Bernoulli **9**(5), 2003, 895–919

A note on parameter differentiation of matrix exponentials, with applications to continuous-time modelling

HENGHSIU TSAI[1] and K.S. CHAN[2]

[1]*Institute of Statistical Science, Academia Sinica, Taipei, Taiwan 115, Republic of China. E-mail: htsai@stat.sinica.edu.tw*
[2]*Department of Statistics & Actuarial Science, University of Iowa, Iowa City, IA 52242, USA. E-mail: kchan@stat.uiowa.edu*

We propose a new analytic formula for evaluating the derivatives of a matrix exponential. In contrast to the diagonalization method, eigenvalues and eigenvectors do not appear explicitly in the derivation, although we show that a necessary and sufficient condition for the validity of the formula is that the matrix has distinct eigenvalues. The new formula expresses the derivatives of a matrix exponential in terms of minors, polynomials, the exponential of the matrix as well as matrix inversion, and hence is algebraically more manageable. For sparse matrices, the formula can be further simplified. Two examples are discussed in some detail. For the companion matrix of a continuous-time autoregressive moving average process, the derivatives of the exponential of the companion matrix can be computed recursively. The second example concerns the exponential of the tridiagonal transition intensity matrix of a finite-state-space continuous-time Markov chain whose instantaneous transitions must be between adjacent states. We present a numerical study to show that the new method may yield numerically more accurate results than the diagonalization method, at the expense of a slight increase in computation.

Keywords: CARMA models; Cayley–Hamilton theorem; finite-state-space continuous-time Markov chain; maximum likelihood estimation; transition intensity matrix

1. Introduction

Various methods of parameter differentiation of a matrix exponential have been studied in statistical mechanics and quantum theory (see Wilcox, 1967), as well as in the mathematics, economics and statistics literature (see Jennrich and Bright 1976; Van Loan 1978; Kalbfleisch and Lawless 1985; Graham 1986; Horn and Johnson 1991; Chen and Zadrozny 2001; and Chan and Munoz-Hernandez 2003). For continuous/discrete state-space modelling (see Jazwinski 1970; Singer 1995), parameter differentiation of a matrix exponential is needed to compute the analytic score function. For continuous-time Markov modelling, an efficient algorithm for the computation of the transition probability matrix and its derivatives with respect to the transition intensity parameters is needed for maximum likelihood estimation. For example, see Kalbfleisch and Lawless (1985) for an approach to

1350–7265

analysing categorical panel data by assuming that the data are obtained from sampling a latent continuous-time finite-state-space Markov process.

We propose in this paper an alternative method for computing the derivatives of a matrix exponential. In contrast to the diagonalization method (see below), eigenvalues and eigenvectors do not appear explicitly in the derivation, although we show that a necessary and sufficient condition for the validity of the formula is that the matrix has distinct eigenvalues. The new formula expresses the derivatives of a matrix exponential in terms of minors, polynomials, the exponential of the matrix as well as matrix inversion, and hence is algebraically more manageable. In particular, we present a numerical study to show that, for nearly non-diagonalizable matrices, the new method may be numerically more accurate than the diagonalization method. When the matrix has repeated eigenvalues, it appears to be hard to extend the results; see the end of Section 2 for a discussion. Fortunately, in most statistical applications that involve matrix exponentials, the distinct eigenvalue assumption is valid. For example, in continuous-time Markov chain modelling, for most models of interest, the transition intensity matrix has distinct eigenvalues for almost all parameter values (see Kalbfleisch and Lawless, 1985).

This paper is organized as follows. In Section 2, we derive the new formula for computing the derivatives of a matrix exponential and a necessary and sufficient condition for the formula to be valid. For sparse matrices, the formula may be further simplified. Two interesting examples are the exponential of the companion matrix arising from a continuous-time autoregressive moving average process and that of the tridiagonal transition intensity matrix arising from a continuous-time Markov chain whose instantaneous transitions must be jumps between adjacent categories. Some simplified formulae for these two examples are given in Section 3. We report in Section 4 a numerical study comparing the new method and the diagonalization method in terms of speed and numerical stability.

2. Main results

Let $A = [a_{ij}]$ be a $p \times p$ matrix whose elements are functions of $\vartheta = (\vartheta_1, \ldots, \vartheta_r)^{\mathrm{T}}$. Recall that the matrix exponential $\mathrm{e}^A = \sum_{j=0}^{\infty} A^j / j!$. There are a number of approaches for computing the partial derivatives of e^{tA} with respect to ϑ_j, where t is a real number. A commonly used method is diagonalization, which assumes that A has distinct eigenvalues $d_1, \ldots, d_p$ so that $A = XDX^{-1}$ and $\mathrm{e}^{tA} = X \operatorname{diag}(\mathrm{e}^{d_1 t}, \ldots, \mathrm{e}^{d_p t}) X^{-1}$, where X is the $p \times p$ matrix whose jth column is a right eigenvector corresponding to d_j and $D = \operatorname{diag}(d_1, \ldots, d_p)$. Furthermore,

$$\frac{\partial \mathrm{e}^{tA}}{\partial \vartheta_u} = X V_u X^{-1}, \qquad u = 1, \ldots, r, \tag{1}$$

where V_u is a $p \times p$ matrix with (i, j)th entry equal to

$$g_{ij}^{(u)} (\mathrm{e}^{d_i t} - \mathrm{e}^{d_j t}) / (d_i - d_j), \qquad i \neq j,$$

$$g_{ii}^{(u)} t \mathrm{e}^{d_i t}, \qquad i = j,$$

and $g_{ij}^{(u)}$ is the (i, j)th entry in $G^{(u)} = X^{-1}(\partial A/\partial \vartheta_u)X$. See Kalbfleisch and Lawless (1985) for the above formula and related discussions; see also Jennrich and Bright (1976) and Chan and Munoz-Hernandez (2003). When A has repeated eigenvalues, an analogous decomposition of A to Jordan canonical form is possible (see Chapter 4 of Cox and Miller, 1965). But, as pointed out by Kalbfleisch and Lawless (1985), this is rarely necessary, since for most models of interest in continuous-time Markov modelling, A has distinct eigenvalues for almost all parameters.

Another approach is based on equation (2.1) of Wilcox (1967) which states that, for $i = 1, \ldots, r$,

$$\frac{\partial \mathrm{e}^{tA}}{\partial \vartheta_i} = \int_0^t \mathrm{e}^{(t-u)A}\left(\frac{\partial A}{\partial \vartheta_i}\right)\mathrm{e}^{uA}\,\mathrm{d}u. \tag{2}$$

One of the main results of this paper is to derive another closed-form solution for $\partial \mathrm{e}^{tA}/\partial \vartheta_i$, based on the Wilcox formula. Let $B_{ij} = \partial A/\partial a_{ij}$, which is a $p \times p$ matrix with a 1 in (i, j)th position and zeros everywhere else. Then, for $1 \leqslant k \leqslant r$,

$$\begin{aligned}
\frac{\partial \mathrm{e}^{tA}}{\partial \vartheta_k} &= \int_0^t \mathrm{e}^{(t-u)A}\,\frac{\partial A}{\partial \vartheta_k}\,\mathrm{e}^{uA}\,\mathrm{d}u \\
&= \int_0^t \mathrm{e}^{(t-u)A}\left(\sum_{i=1}^{p}\sum_{j=1}^{p}\frac{\partial a_{ij}}{\partial \vartheta_k}B_{ij}\right)\mathrm{e}^{uA}\mathrm{d}u \\
&= \sum_{i=1}^{p}\sum_{j=1}^{p}\frac{\partial a_{ij}}{\partial \vartheta_k}\int_0^t \mathrm{e}^{(t-u)A}B_{ij}\,\mathrm{e}^{uA}\,\mathrm{d}u \\
&= \sum_{i=1}^{p}\sum_{j=1}^{p}\frac{\partial a_{ij}}{\partial \vartheta_k}\Sigma_{ij},
\end{aligned} \tag{3}$$

where

$$\Sigma_{ij} = \int_0^t \mathrm{e}^{(t-u)A}B_{ij}\,\mathrm{e}^{uA}\,\mathrm{d}u. \tag{4}$$

A closed-form solution for Σ_{ij} in terms of minors, polynomials, the exponential of the matrix A as well as matrix inversion is given in Theorem 1.

Before stating the main results, we need some notation for an explicit formula for the characteristic polynomial of A. Let $A = [a_{ij}]$ and $B = [b_{ij}]$ be two $p \times p$ matrices. Define $[A, B] = AB - BA$ as the commutator of A and B, and let $|A|$ be the determinant of the matrix A. For vectors $\alpha = [\alpha_1, \ldots, \alpha_q]$ and $\beta = [\beta_1, \ldots, \beta_q]$, where $\alpha_j \in \{1, \ldots, p\}$ and $\beta_j \in \{1, \ldots, p\}$, for $j = 1, \ldots, q(\leqslant p)$, we denote the (sub)matrix that lies in the rows of A indexed by α and the columns indexed by β as $A(\alpha, \beta)$. For example,

$$A([1, 3], [2, 1]) = \begin{bmatrix} a_{12} & a_{11} \\ a_{32} & a_{31} \end{bmatrix}.$$

If $\beta = \alpha$, the submatrix $A(\alpha, \alpha)$ is called a principal submatrix of A and is denoted by $A(\alpha)$; see Horn and Johnson (1985, p. 17). Let $R_0^p = 1$ and, for $1 \leqslant k \leqslant p$, let

$$R_k^p = \sum_{1 \leqslant l_1 < \ldots < l_k \leqslant p} |A([l_1, \ldots, l_k])|. \tag{5}$$

Note that $q(\lambda) = \sum_{k=0}^{p} (-1)^{p-k} R_{p-k}^p \lambda^k$ is the characteristic polynomial of A, and $q(A) = 0$ by the Cayley–Hamilton theorem. Theorem 1 below essentially results from differentiating (5) with respect to a_{ij}. Let $q'(\lambda)$ be the (first) derivative of q with respect to λ. Then $q'(A) = \sum_{k=0}^{p-1} (-1)^{p-k-1} (k+1) R_{p-k-1}^p A^k$, which is independent of t. This fact may result in simpler inferential procedures, as will be illustrated in an example below. The derivative of the matrix exponential is trivial when $p = 1$. For $p \geqslant 2$, we have the following results:

Theorem 1. *For $p \geqslant 2$, and assuming that $q'(A) = \sum_{k=0}^{p-1} (-1)^{p-k-1} (k+1) R_{p-k-1}^p A^k$ is invertible,*

$$\Sigma_{ij} = \left\{ \sum_{k=0}^{p-1} (-1)^{p-k-1} (k+1) R_{p-k-1}^p A^k \right\}^{-1} \left[\left\{ \sum_{k=0}^{p-1} (-1)^{p-k+1} \left(\frac{\partial R_{p-k}^p}{\partial a_{ij}} \right) A^k \right\} t\mathrm{e}^{tA} \right.$$
$$\left. - \sum_{u=0}^{p-2} \sum_{k=u+2}^{p} (-1)^{p-k} (k-u-1) R_{p-k}^p A^{k-u-2} [B_{ij}, \mathrm{e}^{tA}] A^u \right].$$

Theorem 2 gives an explicit representation of the partial derivatives of the R_k^p with respect to the a_{ij}, while Theorem 3 gives a necessary and sufficient condition for $q'(A)$ to be invertible.

Theorem 2. *(a) For $1 \leqslant i \neq j \leqslant p$,*

$$\frac{\partial R_1^p}{\partial a_{ij}} = 0.$$

(b) For $1 \leqslant i \neq j \leqslant p$,

$$\frac{\partial R_2^p}{\partial a_{ij}} = -|A([j], [i])| = -a_{ji}.$$

(c) For $3 \leqslant k \leqslant p$ and $1 \leqslant i \neq j \leqslant p$,

$$\frac{\partial R_k^p}{\partial a_{ij}} = - \sum_{\substack{1 \leqslant l_1 < \ldots < l_{k-2} \leqslant p \\ i \notin \{l_1, \ldots, l_{k-2}\} \\ j \notin \{l_1, \ldots, l_{k-2}\}}} |A([j, l_1, \ldots, l_{k-2}], [i, l_1, \ldots, l_{k-2}])|.$$

(d) For $1 \leqslant i \leqslant p$,

$$\frac{\partial R_1^p}{\partial a_{ii}} = 1.$$

(e) For $2 \leqslant k \leqslant p$ and $1 \leqslant i \leqslant p$,

$$\frac{\partial R_k^p}{\partial a_{ii}} = \sum_{\substack{1 \leqslant l_1 < \ldots < l_{k-1} \leqslant p \\ i \notin \{l_1, \ldots, l_{k-1}\}}} |A([l_1, \ldots, l_{k-1}])|.$$

Theorem 3. *For $p \geqslant 2$, $q'(A) = \sum_{k=0}^{p-1} (-1)^{p-k-1}(k+1)R_{p-k-1}^p A^k$ is invertible if and only if the matrix A has p distinct eigenvalues.*

If A has repeated eigenvalues, Theorem 3 implies that $q'(A)$ is singular, so that Theorem 1 is inapplicable. Now, Theorem 1 may be generalized by considering the equation $m(A) = 0$, where $m(\lambda)$ is the minimal polynomial of A. Indeed, if A is diagonalizable, its minimal polynomial equals $m(\lambda) = \prod(\lambda - \lambda_j)$, where the product is over distinct eigenvalues, in which case, even though the eigenvalues are not distinct, they do not repeat in the minimal polynomial so that $m'(A)$ is invertible. This suggests that the preceding results may be extended to the more general case where A is diagonalizable, or equivalently, its minimal polynomial is of the form $m(\lambda) = \prod(\lambda - \lambda_j)$, where all the λ_j are distinct. However, the coefficients of the minimal polynomial may not admit a simple form. Moreover, Theorem 2 and related results do not appear to be easily generalizable in this more general situation.

3. Applications

3.1. Continuous-time autoregressive moving average processes

For continuous/discrete state-space modelling (see Jazwinski, 1970; Singer, 1995), parameter differentiation of a matrix exponential is needed in computing the analytic score function; indeed, it is also required in other methods of estimation, such as least squares. An example of continuous/discrete state-space modelling is the continuous-time autoregressive moving average (CARMA(p, q)) process, which is defined as the solution of the pth-order differential equation

$$Y_t^{(p)} - \alpha_p Y_t^{(p-1)} - \ldots - \alpha_1 Y_t - \alpha_0 = \sigma[W_t^{(1)} + \beta_1 W_t^{(2)} + \ldots + \beta_q W_t^{(q+1)}], \tag{6}$$

where the superscript $^{(j)}$ denotes j-fold differentiation with respect to t; $\{W_t, t \geqslant 0\}$ is standard Brownian motion, and $\alpha_0, \ldots, \alpha_p, \beta_1, \ldots, \beta_q$ and σ are constants. We assume that $\sigma > 0$, $\alpha_1 \neq 0$ and $\beta_q \neq 0$ and define $\beta_j := 0$ for $j > q$. The derivatives $W_t^{(j)}$, $j > 0$, do not exist in the usual sense; hence, we interpret (6) as being equivalent to the *observation* and *state* equations (the Ys below are observations sampled at possibly unequally spaced epochs $t_1 < t_2 < \ldots < t_n$ from an underlying continuous-time process $\{X_t\}$ and are also referred to as the state vectors: for further discussion, see Arnold 1974; Brockwell 1993; Brockwell and Stramer, 1995):

$$Y_{t_i} = \beta^{\mathrm{T}} X_{t_i}, \qquad i = 1, 2, \ldots, n,$$

$$\mathrm{d}X_t = (AX_t + \alpha_0 \delta_p)\mathrm{d}t + \sigma \delta_p \,\mathrm{d}W_t,$$

where

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ \alpha_1 & \alpha_2 & \alpha_3 & \cdots & \alpha_p \end{bmatrix}, \qquad X_t = \begin{bmatrix} X_t^{(0)} \\ X_t^{(1)} \\ \vdots \\ X_t^{(p-2)} \\ X_t^{(p-1)} \end{bmatrix}, \qquad \delta_p = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \qquad \beta = \begin{bmatrix} 1 \\ \beta_1 \\ \vdots \\ \beta_{p-2} \\ \beta_{p-1} \end{bmatrix}.$$

The solution of the preceding stochastic differential equation is

$$X_t = \mathrm{e}^{At} X_0 + \alpha_0 \int_0^t \mathrm{e}^{A(t-u)} \delta_p \,\mathrm{d}u + \sigma \int_0^t \mathrm{e}^{A(t-u)} \delta_p \,\mathrm{d}W_u.$$

Thus, statistical inference for this model generally requires differentiation of the matrix exponential of At; see below. Note that the companion matrix A is a function of the parameters $\alpha_1, \ldots, \alpha_p$, and, due to the simplicity of the matrix, parameter differentiation of the corresponding matrix exponential can be easily computed by the recursive procedure

$$\frac{\partial \mathrm{e}^{tA}}{\partial \alpha_i} = \left(\frac{\partial \mathrm{e}^{tA}}{\partial \alpha_{i-1}} \right) A, \qquad 2 \leqslant i \leqslant p;$$

see Theorem 4(c) and the appendix for a proof. The partial derivative of e^{tA} with respect to α_1 is given by parts (a) and (b) of the same theorem:

Theorem 4. *(a) For* $p = 1$, $\partial \mathrm{e}^{tA} / \partial \alpha_1 = t\mathrm{e}^{tA}$.

(b) For $p \geqslant 2$,

$$\frac{\partial \mathrm{e}^{tA}}{\partial \alpha_1} = K_{p,0}^{-1} \left\{ t\mathrm{e}^{tA} - \sum_{r=1}^{p-1} K_{p,r} [B_{pr}, \mathrm{e}^{tA}] \right\}, \tag{7}$$

where

$$K_{p,r} = \begin{cases} (p-r)A^{p-r-1} - \displaystyle\sum_{k=r+2}^{p} (k-r-1)\alpha_k A^{k-r-2}, & 0 \leqslant r \leqslant p-2, \\ I & r = p-1, \end{cases}$$

and $[B_{pr}, \mathrm{e}^{tA}] = B_{pr}\mathrm{e}^{tA} - \mathrm{e}^{tA} B_{pr}$ *is the commutator of* B_{pr} *and* e^{tA}.

(c) For $2 \leqslant i \leqslant p$,

$$\frac{\partial \mathrm{e}^{tA}}{\partial \alpha_i} = \left(\frac{\partial \mathrm{e}^{tA}}{\partial \alpha_{i-1}}\right) A. \tag{8}$$

By way of clarification, the expressions for the matrix $\partial \mathrm{e}^{tA}/\partial \alpha_1$, for $p = 1, \ldots, 4$, are as follows: for $p = 1$,

$$\frac{\partial \mathrm{e}^{tA}}{\partial \alpha_1} = t\mathrm{e}^{tA};$$

for $p = 2$,

$$\frac{\partial \mathrm{e}^{tA}}{\partial \alpha_1} = (2A - \alpha_2 I)^{-1}(t\mathrm{e}^{tA} - [B_{21}, \mathrm{e}^{tA}]), \tag{9}$$

for $p = 3$,

$$\frac{\partial \mathrm{e}^{tA}}{\partial \alpha_1} = (3A^2 - 2\alpha_3 A - \alpha_2 I)^{-1}\{t\mathrm{e}^{tA} - (2A - \alpha_3 I)[B_{31}, \mathrm{e}^{tA}] - [B_{32}, \mathrm{e}^{tA}]\}$$

and for $p = 4$,

$$\begin{aligned}\frac{\partial \mathrm{e}^{tA}}{\partial \alpha_1} = {} & (4A^3 - 3\alpha_4 A^2 - 2\alpha_3 A - \alpha_2 I)^{-1}\{t\mathrm{e}^{tA} - (3A^2 - 2\alpha_4 A - \alpha_3 I)[B_{41}, \mathrm{e}^{tA}] \\ & - (2A - \alpha_4 I)[B_{42}, \mathrm{e}^{tA}] - [B_{43}, \mathrm{e}^{tA}]\}.\end{aligned}$$

We now present an example illustrating the use of the new formulae. Suppose that we observed the states X_t from a CAR(p) model over (possibly) unequally spaced epochs (say, t_i), and we wish to compute the conditional least-squares estimators of the parameters. First, note that the sum of squared predictive residuals is

$$g(\alpha_0, \ldots, \alpha_p) = \sum_{i=1}^{N}\{x_{t_i} - \mu - \mathrm{e}^{\Delta_i A}(x_{t_{i-1}} - \mu)\}^{\mathrm{T}}\{x_{t_i} - \mu - \mathrm{e}^{\Delta_i A}(x_{t_{i-1}} - \mu)\},$$

where $\Delta_i = t_i - t_{i-1}$ and $\mu = (-\alpha_0/\alpha_1, 0, \ldots, 0)^{\mathrm{T}}$. For simplicity assume that $\alpha_0 = 0$, so that $\mu = 0$. Therefore, for $1 \leqslant j \leqslant p$,

$$\frac{\partial g(\alpha_0, \ldots, \alpha_p)}{\partial \alpha_j}$$

$$= -2\sum_{i=1}^{N}\left(\frac{\partial e^{\Delta_i A}}{\partial \alpha_j}x_{t_{i-1}}\right)^{\mathrm{T}}(x_{t_i} - e^{\Delta_i A}x_{t_{i-1}})$$

$$= -2\sum_{i=1}^{N}x_{t_{i-1}}^{\mathrm{T}}(A^{\mathrm{T}})^{j-1}\left(\Delta_i e^{\Delta_i A} - \sum_{r=1}^{p-1}K_{p,r}[B_{pr}, e^{\Delta_i A}]\right)^{\mathrm{T}}(K_{p,0}^{-1})^{\mathrm{T}}(x_{t_i} - e^{\Delta_i A}x_{t_{i-1}})$$

$$= -2\sum_{i=1}^{N}\operatorname{tr}\left\{(x_{t_i} - e^{\Delta_i A}x_{t_{i-1}})x_{t_{i-1}}^{\mathrm{T}}(A^{\mathrm{T}})^{j-1}\left(\Delta_i e^{\Delta_i A} - \sum_{r=1}^{p-1}K_{p,r}[B_{pr}, e^{\Delta_i A}]\right)^{\mathrm{T}}(K_{p,0}^{-1})^{\mathrm{T}}\right\}$$

$$= -2\left[\sum_{i=1}^{N}\operatorname{vec}\left\{\left(\Delta_i e^{\Delta_i A} - \sum_{r=1}^{p-1}K_{p,r}[B_{pr}, e^{\Delta_i A}]\right)A^{j-1}x_{t_{i-1}}(x_{t_i} - e^{\Delta_i A}x_{t_{i-1}})^{\mathrm{T}}\right\}\right]^{\mathrm{T}}\operatorname{vec}[(K_{p,0}^{-1})^{\mathrm{T}}].$$

By replacing $K_{p,0}^{-1}$ by the adjoint of $K_{p,0}$ in the preceding expression, high numerical accuracy can be attained even when some of the eigenvalues are nearly identical.

3.2. Tridiagonal intensity matrix in continuous-time Markov processes

Kalbfleisch and Lawless (1985) proposed methods for the analysis of panel data under a continuous-time Markov chain $\{X_t\}$ with a finite state space that has, say, p states. Let $Q = (q_{i,j})$ be a $p \times p$ transition intensity matrix that is constant over any interval between two consecutive integer time points; the intensities $q_{i,j}$ are the rates of transition of the process $\{X_t\}$ from state i to state j over an infinitesimal time period and hence $\sum_j q_{i,j} = 0$ for all i. Furthermore, the transition probability matrix of the process from t to $t+1$ is e^Q. In some applications, the states of a panel of independent individuals following such a transition mechanism are observable over a set of integer time points. The formulae derived below then facilitate a likelihood analysis of such data. For some applications, the matrix Q is a sparse matrix in the sense that only a few elements of Q are non-zero. See Kalbfleisch and Lawless (1985) for examples. Chan and Munoz-Hernandez (2003) adopted the continuous-time Markov processes to model longitudinal data consisting of transitional frequencies classified according to an ordered categorical response variable. The ordering of the categories implies that the continuous-time Markov chain can only jump between adjacent categories over an infinitesimal period, resulting in a tridiagonal transition intensity matrix. For the tridiagonal transition intensity matrix, the coefficients R_k^p and their partial derivatives with respect to the q_{ij} can be further simplified as in Theorem 5 below. Henceforth, assume $p \geqslant 2$, and write the tridiagonal transition intensity matrix as

$$Q = \begin{bmatrix} -q_1 & q_1 & 0 & \cdots & 0 & 0 & 0 \\ q_2 & -q_2 - q_3 & q_3 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & q_{2p-4} & -q_{2p-4} - q_{2p-3} & q_{2p-3} \\ 0 & 0 & 0 & \cdots & 0 & q_{2p-2} & -q_{2p-2} \end{bmatrix}.$$

By equation (3) and the tridiagonality of the matrix, we only need to compute Σ_{ij} for $|i-j| \leqslant 1$, and so only the $\partial R_k^p/\partial q_{ij}$ for $|i-j| \leqslant 1$ are needed. Theorem 5 gives a closed form of the R_k^p and the required derivatives.

Theorem 5. *Assume that* $p \geqslant 2$.

(a) $R_p^p = 0$, *for* $p \geqslant 2$.
(b) *For* $1 \leqslant k \leqslant p-1$,

$$R_k^p = \sum_{i_1=1}^{2p-2k} \sum_{i_2=i_1+2}^{2p-2k+2} \cdots \sum_{i_k=i_{k-1}+2}^{2p-2} (-1)^k q_{i_1} \cdots q_{i_k}.$$

(c) *For* $1 \leqslant i \leqslant p$,

$$\frac{\partial R_1^p}{\partial q_{i,i}} = 1.$$

(d) *For* $1 \leqslant i \leqslant p$ *and* $2 \leqslant k \leqslant p$,

$$\frac{\partial R_k^p}{\partial q_{ii}} = \sum_{i_1=1}^{2p-2k+2} \sum_{i_2=i_1+2}^{2p-2k+4} \cdots \sum_{i_k=i_{k-1}+2}^{2p-2} (-1)^{k-1} q_{i_1}^i \cdots q_{i_{k-1}}^i,$$

where, for $1 \leqslant k \leqslant 2p-2$,

$$q_k^i = \begin{cases} q_k, & \text{if } k \notin \{2i-2, 2i-1\}, \\ 0, & \text{if } k \in \{2i-2, 2i-1\}. \end{cases}$$

(e)

$$\frac{\partial R_1^p}{\partial q_{i,i+1}} = \frac{\partial R_1^p}{\partial q_{i+1,i}} = 0.$$

(f)

$$\frac{\partial R_2^p}{\partial q_{i,i+1}} = -q_{i+1,i} = -q_{2i},$$

and

$$\frac{\partial R_2^p}{\partial q_{i+1,i}} = -q_{i,i+1} = -q_{2i-1}.$$

(g) *For* $1 \leqslant i \leqslant p-1$ *and* $3 \leqslant k \leqslant p$,

$$\frac{\partial R_k^p}{\partial q_{i,i+1}} = -q_{2i} \sum_{i_1=1}^{2p-2k+4} \sum_{i_2=i_1+2}^{2p-2k+6} \cdots \sum_{i_k=i_{k-1}+2}^{2p-2} (-1)^{k-2} \tilde{q}_{i_1}^{\,i} \cdots \tilde{q}_{i_{k-2}}^{\,i},$$

and

$$\frac{\partial R_k^p}{\partial q_{i+1,i}} = -q_{2i-1} \sum_{i_1=1}^{2p-2k+4} \sum_{i_2=i_1+2}^{2p-2k+6} \cdots \sum_{i_k=i_{k-1}+2}^{2p-2} (-1)^{k-2} \tilde{q}^{\,i}_{i_1} \cdots \tilde{q}^{\,i}_{i_{k-2}},$$

where, for $1 \leqslant k \leqslant 2p-2$,

$$\tilde{q}^{\,i}_k = \begin{cases} q_k, & \text{if } k \notin \{2i-2, 2i-1, 2i, 2i+1\}, \\ 0 & \text{if } k \in \{2i-2, 2i-1, 2i, 2i+1\}. \end{cases}$$

Example. By way of clarification, we give the parameter differentiation of the matrix exponential for an example from Chan and Munoz-Hernandez (2003) concerning an analysis of a longitudinal study on election opinions of a group of potential voters. Consider a continuous-time Markov chain that has three (ordered) states, where the model specifies that the transition rates between any two adjacent states to be a linear function of the initial period. Specifically, the transition intensity matrix over the interval $[t, t+1)$ is

$$Q_t = \begin{bmatrix} -q_1 & q_1 & 0 \\ q_2 & -q_2-q_3 & q_3 \\ 0 & q_4 & -q_4 \end{bmatrix} = \begin{bmatrix} -\mathrm{e}^{\theta_5 t+\theta_1} & \mathrm{e}^{\theta_5 t+\theta_1} & 0 \\ \mathrm{e}^{\theta_6 t+\theta_3} & -\mathrm{e}^{\theta_6 t+\theta_3} - \mathrm{e}^{\theta_5 t+\theta_2} & \mathrm{e}^{\theta_5 t+\theta_2} \\ 0 & \mathrm{e}^{\theta_6 t+\theta_4} & -\mathrm{e}^{\theta_6 t+\theta_4} \end{bmatrix}.$$

This model can be used to infer whether the transition intensities are constant against the possibility that they have linear time trends. The transition probability matrix of the Markov chain from t to $t+1$ is $P_t = \mathrm{e}^{Q_t}$. Denote the (u, v)th entry of P_t as $p_{u,v,t}$. Suppose we observed the states of n independent such Markov processes at epochs $t = 0, 1, 2, \ldots, T$ and there were $n_{u,v,t}$ transitions from state u at time t to state v at time $t+1$. Then, conditional on the initial states at $t=0$, the log-likelihood equals $\sum_{t=1}^{T}\sum_{u=1}^{3}\sum_{v=1}^{3} n_{u,v,t}\log(p_{u,v,t})$. Maximum likelihood inference then requires the differentiation of the matrix exponential e^{Q_t}. It follows from Theorems 1 and 5 that

$$\Sigma_{31} = W^{-1}[q_1 q_3 t\mathrm{e}^{tQ} - \{2Q + (q_1+q_2+q_3+q_4)I\}[B_{31}, \mathrm{e}^{tQ}] - [B_{31}, e^{tQ}]Q],$$

where

$$\begin{aligned} W &= 3R_0^3 Q^2 - 2R_1^3 Q + R_2^3 I \\ &= 3Q^2 + 2(q_1+q_2+q_3+q_4)Q + (q_1q_3 + q_1q_4 + q_2q_4)I. \end{aligned}$$

The other Σ_{ij} can be computed from Σ_{31} as follows. First, note that $q_3 \neq 0$ and

$$\begin{aligned} Q\Sigma_{31} &= \int_0^t \mathrm{e}^{(t-u)Q} QB_{31}\mathrm{e}^{uQ}\mathrm{d}u \\ &= \int_0^t \mathrm{e}^{(t-u)Q}(q_3B_{21} - q_4B_{31})\mathrm{e}^{uQ}\mathrm{d}u \\ &= q_3\Sigma_{21} - q_4\Sigma_{31}, \end{aligned}$$

so

$$\Sigma_{21} = \frac{1}{q_3}(Q\Sigma_{31} + q_4\Sigma_{31}).$$

Similarly, we have

$$\Sigma_{32} = \frac{1}{q_1}(\Sigma_{31}Q + q_1\Sigma_{31}),$$

$$\Sigma_{11} = \frac{1}{q_1}(Q\Sigma_{21} + (q_2 + q_3)\Sigma_{21}),$$

$$\Sigma_{22} = \frac{1}{q_1}(\Sigma_{21}Q + q_1\Sigma_{21}),$$

$$\Sigma_{23} = \frac{1}{q_3}(\Sigma_{22}Q + (q_2 + q_3)\Sigma_{22}),$$

$$\Sigma_{33} = t\mathrm{e}^{tQ} - \Sigma_{11} - \Sigma_{22},$$

and

$$\frac{\partial \mathrm{e}^{tQ}}{\partial \theta_1} = -q_1(\Sigma_{11} - \Sigma_{12}),$$

$$\frac{\partial \mathrm{e}^{tQ}}{\partial \theta_2} = -q_3(\Sigma_{22} - \Sigma_{23}),$$

$$\frac{\partial \mathrm{e}^{tQ}}{\partial \theta_3} = q_2(\Sigma_{21} - \Sigma_{22}),$$

$$\frac{\partial \mathrm{e}^{tQ}}{\partial \theta_4} = q_4(\Sigma_{32} - \Sigma_{33}),$$

$$\begin{aligned}\frac{\partial \mathrm{e}^{tQ}}{\partial \theta_5} &= -tq_1\Sigma_{11} + tq_1\Sigma_{12} - tq_3\Sigma_{22} + tq_3\Sigma_{23} \\ &= t\left(\frac{\partial \mathrm{e}^{tQ}}{\partial \theta_1} + \frac{\partial \mathrm{e}^{tQ}}{\partial \theta_2}\right),\end{aligned}$$

$$\begin{aligned}\frac{\partial \mathrm{e}^{tQ}}{\partial \theta_6} &= tq_2\Sigma_{21} - tq_2\Sigma_{22} + tq_4\Sigma_{32} - tq_4\Sigma_{33} \\ &= t\left(\frac{\partial \mathrm{e}^{tQ}}{\partial \theta_3} + \frac{\partial \mathrm{e}^{tQ}}{\partial \theta_4}\right).\end{aligned}$$

We note that the preceding two equalities between the partial derivatives may not be easily realized from the diagonalization method.

4. Numerical study

In this section, we report a numerical study for comparing the speed and numerical stability of our method with the diagonalization method. Now, for nearly non-diagonalizable matrices, the determination of the eigenvectors may be numerically unstable and yet the inverse operation needed in the new method could be well conditioned. We illustrate this observation with an example in the form of the companion matrix of a CARMA(2,0) process (see Section 3.1 for details), i.e.

$$A = \begin{bmatrix} 0 & 1 \\ \alpha_1 & \alpha_2 \end{bmatrix}.$$

Write $\partial \mathrm{e}^A / \partial \alpha_1 = [\partial_{ij}]$. Let the eigenvalues of A be β_1 and β_2; they are negative, and $h = \beta_1 - \beta_2 > 0$. It is readily seen that $\alpha_1 = -\beta_1\beta_2$ and $\alpha_2 = \beta_1 + \beta_2$. It follows from (9) and routine calculations that

$$\begin{aligned}
\partial_{11} &= h^{-3}[\mathrm{e}^{\beta_1}\{2\beta_2 + h(1-\beta_2)\} + \mathrm{e}^{\beta_2}\{-2\beta_1 + h(1-\beta_1)\}], \\
\partial_{12} &= h^{-3}[2(\mathrm{e}^{\beta_2} - \mathrm{e}^{\beta_1}) + h(\mathrm{e}^{\beta_1} + \mathrm{e}^{\beta_2})], \\
\partial_{21} &= h^{-3}[\mathrm{e}^{\beta_1}\{\alpha_1 h + h^2 - 2\alpha_1\} + \mathrm{e}^{\beta_2}\{\alpha_1 h - h^2 + 2\alpha_1\}], \\
\partial_{22} &= h^{-3}[\mathrm{e}^{\beta_1}\{-2\beta_1 + h(1+\beta_1)\} + \mathrm{e}^{\beta_2}\{2\beta_2 + h(1+\beta_2)\}]. \qquad (10)
\end{aligned}$$

Furthermore,

$$\lim_{h\to 0} \frac{\partial \mathrm{e}^A}{\partial \alpha_1} = \frac{1}{6}\begin{bmatrix} (3-\beta_2)\mathrm{e}^{\beta_2} & \mathrm{e}^{\beta_2} \\ (6-\beta_2^2)\mathrm{e}^{\beta_2} & (3+\beta_2)\mathrm{e}^{\beta_2} \end{bmatrix}. \qquad (11)$$

Here, we compute $\partial \mathrm{e}^A/\partial \alpha_1$ by: (a) the new method, namely formula (9); (b) the diagonalization method, namely formula (1); (c) formula (10); and (d) formula (11). In order to compare the stability of methods (a) and (b), we choose $\beta_1 = -1$ and set h to be 0.1^r, for $r = 1, \ldots, 7$. In computing (11), β_2 was replaced by $\beta_1 - h = -1 - h$; this yields more accurate asymptotic values. All computations were done in double precision, using the Compaq Visual Fortran Version 6.1 compiler. For X^{-1} of formula (1) and $(2A - \alpha_2 I)^{-1}$ of formula (9), the DLFCRG subroutine of the IMSL package was called to compute the LU factorization of the matrices and to check for singularity or ill conditioning, and then DLFIRG was called to compute the inverse matrices. The results are listed in Table 1. While all methods are comparable for not too small a difference between the eigenvalues, method (a) clearly outperforms methods (b) and (c) when the difference between the two eigenvalues

Table 1. The values of $\partial e^A/\partial \alpha_1$ computed by methods (a), (b), (c) and (d)

r	(a)		(b)		(c)		(d)	
1	0.236296	0.058338	0.236296	0.058338	0.236296	0.058338	0.227462	0.055479
	0.285912	0.113787	0.285912	0.113787	0.285912	0.113787	0.265742	0.105409
2	0.244336	0.061008	0.244336	0.061008	0.244336	0.061008	0.243420	0.060703
	0.304428	0.121710	0.304428	0.121710	0.304428	0.121710	0.302296	0.120799
3	0.245161	0.061283	0.245161	0.061283	0.245161	0.061283	0.245069	0.061252
	0.306352	0.122535	0.306352	0.122535	0.306351	0.122535	0.306137	0.122443
4	0.245244	0.061310	0.245224	0.061290	0.245137	0.061327	0.245235	0.061307
	0.306545	0.122617	0.306565	0.122637	0.306422	0.122680	0.306523	0.122608
5	0.245247	0.061313	0.257897	0.073958	0.333067	0.116704	0.245251	0.061313
	0.306564	0.122623	0.293919	0.109981	0.333067	0.111022	0.306562	0.122625
6	0.245117	0.061279	31.3456	31.1617	0.000000	−19.4039	0.245253	0.061313
	0.307129	0.122681	−30.7938	−30.9777	110.914	0.000000	0.306566	0.122626
7	0.250000	0.046875	−14916.1	−14916.3	0.000000	15230.6	0.245253	0.061313
	0.281250	0.125000	14916.7	14916.5	0.000000	0.000000	0.306566	0.122626

becomes very small. These results can be understood as follows. The eigenvectors of A can be shown to be proportional to $(1, \beta_1)^{\mathrm{T}}$ and $(1, \beta_2)^{\mathrm{T}}$ and are nearly linearly dependent for nearly zero h, rendering the diagonalization method numerically unstable. On the other hand, the matrix $2A - \alpha_2 I$ appearing in (9) has a condition number equal to 1 and hence is well conditioned; the condition number of a matrix is defined as the product of the spectral norm of the matrix and that of the inverse of the matrix (see Moler and Van Loan 1978). Consequently, the inverse operation in (9) is not particularly sensitive to machine rounding error. We also note that the matrix exponential needed in (9) is computed by the Padé approximation (Ward 1977) that does not make use of diagonalization.

In order to assess the speed of methods (a) and (b), we compute $\partial \mathrm{e}^A/\partial\alpha_1$ and $\partial \mathrm{e}^A/\partial\alpha_2$ by each method 100 000 times, which took 37 and 34 seconds respectively for methods (a) and (b). For this example, the diagonalization method was about 10% faster than the new method. It would be interesting to further identify the conditions under which the new method is numerically more (less) stable than the diagonalization method.

APPENDIX

Proof of Theorem 1. By the Cayley–Hamilton theorem (see Horn and Johnson, 1985, p. 86),

$$\sum_{k=0}^{p}(-1)^{p-k} R_{p-k}^{p} A^k = 0. \tag{12}$$

Taking the partial derivative with respect to a_{ij} on both sides, we have

$$\sum_{k=0}^{p}(-1)^{p-k}\left(\frac{\partial R_{p-k}^{p}}{\partial a_{ij}}\right)A^k + \sum_{k=1}^{p}(-1)^{p-k} R_{p-k}^{p} \sum_{r=0}^{k-1} A^{k-r-1}\left(\frac{\partial A}{\partial a_{ij}}\right)A^r = 0, \tag{13}$$

which implies

$$\sum_{k=1}^{p}(-1)^{p-k} R_{p-k}^{p} \sum_{r=0}^{k-1} A^{k-r-1} B_{ij} A^r = \sum_{k=0}^{p-1}(-1)^{p-k+1}\left(\frac{\partial R_{p-k}^{p}}{\partial a_{ij}}\right)A^k. \tag{14}$$

Premultiplying both sides of equation (14) by $\mathrm{e}^{(t-u)A}$ and postmultiplying by e^{uA} and then integrating with respect to u from 0 to t to get

$$
\begin{aligned}
&\int_0^t \mathrm{e}^{(t-u)A}\left\{\sum_{k=1}^{p}(-1)^{p-k}R_{p-k}^{p}\sum_{r=0}^{k-1}A^{k-r-1}B_{ij}A^{r}\right\}\mathrm{e}^{uA}\,\mathrm{d}u\\
&=\int_0^t \mathrm{e}^{(t-u)A}\left\{\sum_{k=0}^{p-1}(-1)^{p-k+1}\left(\frac{\partial R_{p-k}^{p}}{\partial a_{ij}}\right)A^{k}\right\}\mathrm{e}^{uA}\,\mathrm{d}u. \qquad (15)
\end{aligned}
$$

Using the fact that (see Hale, 1969 p. 95),

$$A\mathrm{e}^{tB}=\mathrm{e}^{tB}A \text{ if and only if } AB=BA,$$

we have

$$\sum_{k=1}^{p}(-1)^{p-k}R_{p-k}^{p}\sum_{r=0}^{k-1}A^{k-r-1}\Sigma_{ij}A^{r}=\left\{\sum_{k=0}^{p-1}(-1)^{p-k+1}\left(\frac{\partial R_{p-k}^{p}}{\partial a_{ij}}\right)A^{k}\right\}t\mathrm{e}^{tA}. \qquad (16)$$

Applying integration by parts to equation (4), we obtain

$$\Sigma_{ij}A=A\Sigma_{ij}+[B_{ij},\,\mathrm{e}^{tA}];$$

postmultiplying both sides by A then gives

$$
\begin{aligned}
\Sigma_{ij}A^2&=A\Sigma_{ij}A+[B_{ij},\,\mathrm{e}^{tA}]A\\
&=A(A\Sigma_{ij}+[B_{ij},\,\mathrm{e}^{tA}])+[B_{ij},\,\mathrm{e}^{tA}]A\\
&=A^2\Sigma_{ij}+A[B_{ij},\,\mathrm{e}^{tA}]+[B_{ij},\,\mathrm{e}^{tA}]A.
\end{aligned}
$$

Applying the same technique recursively, we obtain

$$\Sigma_{ij} A^r = A^r \Sigma_{ij} + \sum_{u=0}^{r-1} A^{r-u-1}[B_{ij}, e^{tA}]A^u, \qquad r \geqslant 1. \tag{17}$$

It follows from equation (17) that the left-hand side of (16) can be rewritten as

$$\begin{aligned}
&(-1)^{p-1} R^p_{p-1} \Sigma_{ij} \\
&\quad + \sum_{k=2}^{p} (-1)^{p-k} R^p_{p-k} \left\{ \mathrm{A}^{k-1}\Sigma_{ij} + \sum_{r=1}^{k-1} A^{k-r-1} \left(A^r \Sigma_{ij} + \sum_{u=0}^{r-1} A^{r-u-1}[B_{ij}, e^{tA}]A^u \right) \right\} \\
&= (-1)^{p-1} R^p_{p-1} \Sigma_{ij} + \sum_{k=2}^{p} (-1)^{p-k} R^p_{p-k} A^{k-1} \Sigma_{ij} \\
&\quad + \sum_{k=2}^{p} (-1)^{p-k} R^p_{p-k} \sum_{r=1}^{k-1} A^{k-1} \Sigma_{ij} + \sum_{k=2}^{p} (-1)^{p-k} R^p_{p-k} \sum_{r=1}^{k-1} \sum_{u=0}^{r-1} A^{k-u-2}[B_{ij}, e^{tA}]A^u \\
&= \left\{ \sum_{k=1}^{p} (-1)^{p-k} k R^p_{p-k} A^{k-1} \right\} \Sigma_{ij} + \sum_{k=2}^{p} (-1)^{p-k} R^p_{p-k} \sum_{u=0}^{k-2} \sum_{r=u+1}^{k-1} A^{k-u-2}[B_{ij}, e^{tA}]A^u \\
&= \left\{ \sum_{r=0}^{p-1} (-1)^{p-r-1} (r+1) R^p_{p-r-1} A^r \right\} \Sigma_{ij} \\
&\quad + \sum_{k=2}^{p} (-1)^{p-k} R^p_{p-k} \sum_{u=0}^{k-2} (k-u-1) A^{k-u-2}[B_{ij}, e^{tA}]A^u \\
&= \left\{ \sum_{r=0}^{p-1} (-1)^{p-r-1} (r+1) R^p_{p-r-1} A^r \right\} \Sigma_{ij} \\
&\quad + \sum_{u=0}^{p-2} \sum_{k=u+2}^{p} (-1)^{p-k} (k-u-1) R^p_{p-k} A^{k-u-2}[B_{ij}, e^{tA}]A^u.
\end{aligned}$$

This proves the result. □

Proof of Theorem 2. The proofs of (a), (b) and (d) are trivial.

(c) First, note that, for $3 \leqslant k \leqslant p$,

$$R_k^p = \sum_{1 \leqslant l_1 < \ldots < l_k \leqslant p} |A([l_1, \ldots, l_k])|$$

$$= \sum_{\substack{1 \leqslant l_1 < \ldots < l_k \leqslant p \\ i,j \in \{l_1, \ldots, l_k\}}} |A([l_1, \ldots, l_k])| + \sum_{\substack{1 \leqslant l_1 < \ldots < l_k \leqslant p \\ i \notin \{l_1, \ldots, l_k\} \\ \text{or } j \notin \{l_1, \ldots, l_k\}}} |A([l_1, \ldots, l_k])|$$

$$= - \sum_{\substack{1 \leqslant l_1 < \ldots < l_{k-2} \leqslant p \\ i \notin \{l_1, \ldots, l_{k-2}\} \\ j \notin \{l_1, \ldots, l_{k-2}\}}} |A([i, j, l_1, \ldots, l_{k-2}], [j, i, l_1, \ldots, l_{k-2}])|$$

$$+ \sum_{\substack{1 \leqslant l_1 < \ldots < l_k \leqslant p \\ i \notin \{l_1, \ldots, l_k\} \\ \text{or } j \notin \{l_1, \ldots, l_k\}}} |A([l_1, \ldots, l_k])|,$$

which implies that, for $1 \leqslant i \neq j \leqslant p$,

$$\frac{\partial R_k^p}{\partial a_{ij}} = - \sum_{\substack{1 \leqslant l_1 < \ldots < l_{k-2} \leqslant p \\ i \notin \{l_1, \ldots, l_{k-2}\} \\ j \notin \{l_1, \ldots, l_{k-2}\}}} |A([j, l_1, \ldots, l_{k-2}], [i, l_1, \ldots, l_{k-2}])|.$$

(e) For $2 \leqslant k \leqslant p$,

$$R_k^p = \sum_{1 \leqslant l_1 < \cdots < l_k \leqslant p} |A([l_1, \ldots, l_k])|$$

$$= \sum_{\substack{1 \leqslant l_1 < \cdots < l_k \leqslant p \\ i \in \{l_1, \ldots, l_k\}}} |A([l_1, \ldots, l_k])| + \sum_{\substack{1 \leqslant l_1 < \cdots < l_k \leqslant p \\ i \notin \{l_1, \ldots, l_k\}}} |A([l_1, \ldots, l_k])|,$$

which implies that, for $1 \leqslant i \leqslant p$,

$$\frac{\partial R_k^p}{\partial a_{ii}} = \sum_{\substack{1 \leqslant l_1 < \cdots < l_{k-1} \leqslant p \\ i \notin \{l_1, \ldots, l_{k-1}\}}} |A([l_1, \ldots, l_{k-1}])|$$

$\square$

Proof of Theorem 3. The characteristic polynomial can be written as

$$q(\lambda) = |\lambda I - A| = \prod_{i=1}^{p}(\lambda - \lambda_{\mathrm{i}}),$$

where the λ_i are the eigenvalues of A. Now, the derivative of the characteristic polynomial is $q'(\lambda) = \sum_{i=1}^{p}\prod_{j\neq i}(\lambda - \lambda_j)$. Hence, $q'(A) = \sum_{i=1}^{p}\prod_{j\neq i}(A - \lambda_j I)$. If v_k is an eigenvector of A corresponding to the eigenvalue λ_k, then

$$q'(A)v_k = \sum_{i=1}^{p}\prod_{j\neq i}(\lambda_k - \lambda_j)v_k.$$

In other words, the eigenvalues of $q'(A)$ are $\sum_{i=1}^{p}\prod_{j\neq i}(\lambda_k - \lambda_j) = \prod_{j\neq k}(\lambda_k - \lambda_j)$, $k = 1, 2, \ldots, p$, which are non-zero if and only if all the eigenvalues of A are distinct. □

Proof of Theorem 4. (a) The proof of (a) is trivial.

(b) The proof for $p = 2$ is trivial. For $p \geqslant 3$, first note that

$$R^p_{p-k-1} = \begin{cases} (-1)^{p-k}\alpha_{k+2}, & \text{for } -1 \leqslant k \leqslant p-2 \\ 1, & \text{for } k = p-1. \end{cases} \tag{18}$$

$$\begin{aligned} \frac{\partial R^p_{p-k}}{\partial a_{p1}} &= \begin{cases} \dfrac{\partial}{\partial \alpha_1}\{(-1)^{p-k+1}\alpha_{k+1}\}, & \text{for } 0 \leqslant k \leqslant p-1, \\ 0, & \text{for } k = p, \end{cases} \\ &= \begin{cases} (-1)^{p+1}, & \text{for } k = 0, \\ 0, & \text{for } 1 \leqslant k \leqslant p, \end{cases} \end{aligned} \tag{19}$$

and

$$[B_{p1}, \mathrm{e}^{tA}]A^u = [B_{p1}A^u, \mathrm{e}^{tA}] = [B_{p(u+1)}, \mathrm{e}^{tA}], \qquad \text{for } 0 \leqslant u \leqslant p-1. \tag{20}$$

Thus,

$$\begin{aligned}
\frac{\partial \mathrm{e}^{tA}}{\partial \alpha_1} &= \int_0^t \mathrm{e}^{(t-u)A}\left(\frac{\partial A}{\partial \alpha_1}\right)\mathrm{e}^{uA}\,\mathrm{d}u \qquad \text{(by equation (2))}\\
&= \int_0^t \mathrm{e}^{(t-u)A} B_{p1}\mathrm{e}^{uA}\,\mathrm{d}u\\
&= \Sigma_{p1} \qquad \text{(by equation (4))}\\
&= \left\{pA^{p-1} + \sum_{k=0}^{p-2}(-1)^{p-k-1}(k+1)(-1)^{p-k}\alpha_{k+2}A^k\right\}^{-1}\Bigg[(-1)^{p+1}(-1)^{p+1}t\mathrm{e}^{tA}\\
&\quad - \sum_{u=0}^{p-3}\sum_{k=u+2}^{p-1}(-1)^{p-k}(-1)^{p-k+1}\alpha_{k+1}(k-u-1)A^{k-u-2}[B_{p1}, \mathrm{e}^{tA}]A^u\\
&\quad - \sum_{u=0}^{p-2}(-1)^{p-p}(p-u-1)A^{p-u-2}[B_{p1}, \mathrm{e}^{tA}]A^u\Bigg] \text{ (by Theorem 1, equations (18) and (19))}\\
&= \left\{pA^{p-1} - \sum_{k=0}^{p-2}(k+1)\alpha_{k+2}A^k\right\}^{-1}\\
&\quad \times \Bigg\{t\mathrm{e}^{tA} + \sum_{u=0}^{p-3}\sum_{k=u+2}^{p-1}(k-u-1)\alpha_{k+1}A^{k-u-2}[B_{p(u+1)}, \mathrm{e}^{tA}]\\
&\quad - \sum_{u=0}^{p-2}(p-u-1)A^{p-u-2}[B_{p(u+1)}, \mathrm{e}^{tA}]\Bigg\} \qquad \text{(by equation (20))}\\
&= K_{p,0}^{-1}\left\{t\mathrm{e}^{tA} + \sum_{r=1}^{p-2}\sum_{k=r+1}^{p-1}(k-r)\alpha_{k+1}A^{k-r-1}[B_{pr}, \mathrm{e}^{tA}] - \sum_{r=1}^{p-1}(p-r)A^{p-r-1}[B_{pr}, \mathrm{e}^{tA}]\right\}\\
&\hspace{20em} \text{(by letting } r = u+1)\\
&= K_{p,0}^{-1}\left\{t\mathrm{e}^{tA} + \sum_{r=1}^{p-2}\sum_{v=r+2}^{p}(v-r-1)\alpha_v A^{v-r-2}[B_{pr}, \mathrm{e}^{tA}] - \sum_{r=1}^{p-1}(p-r)A^{p-r-1}[B_{pr}, \mathrm{e}^{tA}]\right\}\\
&\hspace{20em} \text{(by letting } v = k+1)\\
&= K_{p,0}^{-1}\Bigg[t\mathrm{e}^{tA} - \sum_{r=1}^{p-2}\left\{(p-r)A^{p-r-1} - \sum_{k=r+2}^{p}(k-r-1)\alpha_k A^{k-r-2}\right\}[\mathrm{B}_{pr}, \mathrm{e}^{tA}]\\
&\quad - [B_{p(p-1)}, \mathrm{e}^{tA}]\Bigg] = K_{p,0}^{-1}\left\{t\mathrm{e}^{tA} - \sum_{r=1}^{p-1}K_{p,r}[B_{pr}, \mathrm{e}^{tA}]\right\}.
\end{aligned}$$

This proves the result.

(c) For $2 \leqslant i \leqslant p$,

$$\begin{aligned}\frac{\partial \mathrm{e}^{tA}}{\partial \alpha_i} &= \int_0^t \mathrm{e}^{(t-u)A} \delta_p \delta_i' \mathrm{e}^{uA}\, \mathrm{d}u \\ &= \int_0^t \mathrm{e}^{(t-u)A} \delta_p \delta_{i-1}' A \mathrm{e}^{uA}\, \mathrm{d}u \qquad (\text{because } \delta_{i-1}' A = \delta_i') \\ &= \left(\frac{\partial \mathrm{e}^{tA}}{\partial \alpha_{i-1}} \right) A \qquad (\text{because } A\mathrm{e}^{uA} = \mathrm{e}^{uA} A).\end{aligned}$$

Proof of Theorem 5. The proofs of (c), (e) and (f) are trivial.

We prove (a) and (b) together. Let $q_0 = q_{2p-1} = 0$, then the transition intensity matrix can be rewritten as

$$Q_p = [q_{i,j}] = \begin{bmatrix} -q_0 - q_1 & q_1 & 0 & \cdots & 0 & 0 & 0 \\ q_2 & -q_2 - q_3 & q_3 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & q_{2p-4} & -q_{2p-4} - q_{2p-3} & q_{2p-3} \\ 0 & 0 & 0 & \cdots & 0 & q_{2p-2} & -q_{2p-2} - q_{2p-1} \end{bmatrix},$$

and

$$q_{i,j} = \begin{cases} -q_{2i-2} - q_{2i-1}, & \text{if } 1 \leqslant i = j \leqslant p, \\ q_{2i-1}, & \text{if } 1 \leqslant j = i+1 \leqslant p, \\ q_{2i-2}, & \text{if } 1 \leqslant j = i-1 \leqslant p-1, \\ 0, & \text{otherwise.} \end{cases}$$

For $1 \leqslant i \leqslant p$, let

$$Q_p^i = \begin{bmatrix} -q_1^i & q_1^i & 0 & \cdots & 0 & 0 & 0 \\ q_2^i & -q_2^i - q_3^i & q_3^i & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & q_{2p-4}^i & -q_{2p-4}^i - q_{2p-3}^i & q_{2p-3}^i \\ 0 & 0 & 0 & \cdots & 0 & q_{2p-2}^i & -q_{2p-2}^i \end{bmatrix},$$

where, for $1 \leqslant k \leqslant 2p-2$,

$$q_k^i = \begin{cases} q_k, & \text{if } k \notin \{2i-2, 2i-1\}, \\ 0, & \text{if } k \in \{2i-2, 2i-1\}. \end{cases}$$

For $1 \leqslant i \leqslant p$, let $R_{0,i}^p = 1$. For $p \geqslant 2$ and $1 \leqslant i, k \leqslant p$, define

$$R_{k,i}^{p} = \sum_{i_1=1}^{2p-2k} \sum_{i_2=i_1+2}^{2p-2k+2} \cdots \sum_{i_k=i_{k-1}+2}^{2p-2} (-1)^k q_{i_1}^{i} \cdots q_{i_k}^{i}.$$

We will prove later that, for $p \geqslant 2$, and $1 \leqslant k \leqslant p$,

$$R_k^p = \sum_{i_1=0}^{2p-2k+1} \sum_{i_2=i_1+2}^{2p-2k+3} \cdots \sum_{i_k=i_{k-1}+2}^{2p-1} (-1)^k q_{i_1} \cdots q_{i_k}. \tag{21}$$

But because $q_0 = q_{2p-1} = 0$, we have, by equation (21),

$$R_p^p = \sum_{i_1=0}^{1} \sum_{i_2=i_1+2}^{3} \cdots \sum_{i_p=i_{p-1}+2}^{2p-1} (-1)^p q_{i_1} \cdots q_{i_p} = 0,$$

and, for $1 \leqslant k \leqslant p-1$,

$$R_k^p = \sum_{i_1=1}^{2p-2k} \sum_{i_2=i_1+2}^{2p-2k+2} \cdots \sum_{i_k=i_{k-1}+2}^{2p-2} (-1)^k q_{i_1} \cdots q_{i_k}.$$

This proves parts (a) and (b) of the theorem.

We prove equation (21) by mathematical induction. First, it is easily seen that (21) holds for $p=2$ and $1 \leqslant k \leqslant 2$. Now, suppose (21) holds for R_k^r, where $2 \leqslant r \leqslant p-1$ and $1 \leqslant k \leqslant r$; we wish to show that, for $2 \leqslant k \leqslant p$, R_k^p is given by the right-hand side of (21) (the proof for $k=1$ is trivial).

Note that, for the tridiagonal matrix Q_p, we have

$$|\lambda I - Q_p| = (\lambda - q_{p,p})|\lambda I - Q_{p-1}| - q_{p,p-1} q_{p-1,p} |\lambda I - Q_{p-2}|,$$

which implies that

$$\sum_{k=0}^{p} (-1)^{p-k} R_{p-k}^{p} \lambda^k = (\lambda + q_{2p-2} + q_{2p-1}) \sum_{k=0}^{p-1} (-1)^{p-k-1} R_{p-k-1}^{p-1} \lambda^k$$
$$- q_{2p-2} q_{2p-3} \sum_{k=0}^{p-2} (-1)^{p-k-2} R_{p-k-2}^{p-2} \lambda^k.$$

Comparing the coefficients of λ^k on both sides, we have, for $k = 1, \ldots, p-2$,

$$(-1)^{p-k} R_{p-k}^{p} = (-1)^{p-k} R_{p-k}^{p-1} + (-1)^{p-k-1} (q_{2p-2} + q_{2p-1}) R_{p-k-1}^{p-1}$$
$$- (-1)^{p-k-2} q_{2p-2} q_{2p-3} R_{p-k-2}^{p-2}.$$

Equivalently, we have, for $k = 2, \ldots, p-1$,

$$(-1)^k R_k^p = (-1)^k R_k^{p-1} + (-1)^{k-1} (q_{2p-2} + q_{2p-1}) R_{k-1}^{p-1}$$
$$- (-1)^{k-2} q_{2p-2} q_{2p-3} R_{k-2}^{p-2}. \tag{22}$$

But, for $k = 3, \ldots, p-1$,

$$R_k^{p-1} = \sum_{i_1=0}^{2p-2k-1} \sum_{i_2=i_1+2}^{2p-2k+1} \cdots \sum_{i_k=i_{k-1}+2}^{2p-3} (-1)^k q_{i_1} \cdots q_{i_k},$$

$$R_{k-1}^{p-1} = \sum_{i_1=0}^{2p-2k+1} \sum_{i_2=i_1+2}^{2p-2k+3} \cdots \sum_{i_{k-1}=i_{k-2}+2}^{2p-3} (-1)^{k-1} q_{i_1} \cdots q_{i_{k-1}},$$

$$R_{k-2}^{p-2} = \sum_{i_1=0}^{2p-2k+1} \sum_{i_2=i_1+2}^{2p-2k+3} \cdots \sum_{i_{k-2}=i_{k-3}+2}^{2p-5} (-1)^{k-2} q_{i_1} \cdots q_{i_{k-2}},$$

and so, for $k = 3, \ldots, p-1$, equation (22) becomes

$$\begin{aligned}
(-1)^k R_k^p = {} & \sum_{i_1=0}^{2p-2k-1} \sum_{i_2=i_1+2}^{2p-2k+1} \cdots \sum_{i_{k-1}=i_{k-2}+2}^{2p-5} \sum_{i_k=i_{k-1}+2}^{2p-3} q_{i_1} \cdots q_{i_k} \\
& + \sum_{i_1=0}^{2p-2k} \sum_{i_2=i_1+2}^{2p-2k+2} \cdots \sum_{i_{k-2}=i_{k-3}+2}^{2p-6} \sum_{i_{k-1}=i_{k-2}+2}^{2p-4} \sum_{i_k=2p-2}^{2p-1} q_{i_1} \cdots q_{i_k} \\
& + \sum_{i_1=0}^{2p-2k+1} \sum_{i_2=i_1+2}^{2p-2k+3} \cdots \sum_{i_{k-2}=i_{k-3}+2}^{2p-5} \sum_{i_{k-1}=2p-3}^{2p-3} \sum_{i_k=2p-2}^{2p-2} q_{i_1} \cdots q_{i_k} \\
& + \sum_{i_1=0}^{2p-2k+1} \sum_{i_2=i_1+2}^{2p-2k+3} \cdots \sum_{i_{k-2}=i_{k-3}+2}^{2p-5} \sum_{i_{k-1}=2p-3}^{2p-3} \sum_{i_k=2p-1}^{2p-1} q_{i_1} \cdots q_{i_k} \\
& - \sum_{i_1=0}^{2p-2k+1} \sum_{i_2=i_1+2}^{2p-2k+3} \cdots \sum_{i_{k-2}=i_{k-3}+2}^{2p-5} \sum_{i_{k-1}=2p-3}^{2p-3} \sum_{i_k=2p-2}^{2p-2} q_{i_1} \cdots q_{i_k} \\
= {} & \sum_{i_1=0}^{2p-2k-1} \sum_{i_2=i_1+2}^{2p-2k+1} \cdots \sum_{i_{k-1}=i_{k-2}+2}^{2p-5} \sum_{i_k=i_{k-1}+2}^{2p-3} q_{i_1} \cdots q_{i_k} \\
& + \sum_{i_1=0}^{2p-2k} \sum_{i_2=i_1+2}^{2p-2k+2} \cdots \sum_{i_{k-2}=i_{k-3}+2}^{2p-6} \sum_{i_{k-1}=i_{k-2}+2}^{2p-4} \sum_{i_k=2p-2}^{2p-1} q_{i_1} \cdots q_{i_k} \\
& + \sum_{i_1=0}^{2p-2k+1} \sum_{i_2=i_1+2}^{2p-2k+3} \cdots \sum_{i_{k-2}=i_{k-3}+2}^{2p-5} \sum_{i_{k-1}=2p-3}^{2p-3} \sum_{i_k=2p-1}^{2p-1} q_{i_1} \cdots q_{i_k} \\
= {} & \sum_{i_1=0}^{2p-2k+1} \sum_{i_2=i_1+2}^{2p-2k+3} \cdots \sum_{i_k=i_{k-1}+2}^{2p-1} q_{i_1} \cdots q_{i_k}.
\end{aligned}$$

The proofs for $k = 2$ and p are similar to that of $3 \leqslant k \leqslant p-1$. This proves equation (21).

(d) First note that, for $1 \leqslant l_1 < \ldots < l_{k-1} \leqslant p$ and $i \notin \{l_1, \ldots, l_{k-1}\}$,

$|Q_p([l_1, \ldots, l_{k-1}])| = |Q_p^i([l_1, \ldots, l_{k-1}])|$. For $1 \leqslant l_1 < \ldots < l_{k-1} \leqslant p$ and $i \in \{l_1, \ldots, l_{k-1}\}$, $|Q_p^i([l_1, \ldots, l_{k-1}])| = 0$, because the matrix $Q_p^i([l_1, \ldots, l_{k-1}])$ contains a zero row vector. Thus, by Theorem 2(e), we have that, for $2 \leqslant k \leqslant p$ and $1 \leqslant i \leqslant p$,

$$\begin{aligned}
\frac{\partial R_k^p}{\partial q_{i,i}} &= \sum_{\substack{1 \leqslant l_1 < \ldots < l_{k-1} \leqslant p \\ i \notin \{l_1, \ldots, l_{k-1}\}}} |Q_p([l_1, \ldots, l_{k-1}])| \\
&= \sum_{\substack{1 \leqslant l_1 < \ldots < l_{k-1} \leqslant p \\ i \notin \{l_1, \ldots, l_{k-1}\}}} |Q_p^i([l_1, \ldots, l_{k-1}])| + \sum_{\substack{1 \leqslant l_1 < \ldots < l_{k-1} \leqslant p \\ i \in \{l_1, \ldots, l_{k-1}\}}} |Q_p^i([l_1, \ldots, l_{k-1}])| \\
&= \sum_{1 \leqslant l_1 < \ldots < l_{k-1} \leqslant p} |Q_p^i([l_1, \ldots, l_{k-1}])| \\
&= R_{k-1,i}^p,
\end{aligned}$$

where the last equality follows from equation (5), Theorem 5(b) and the definition of $R_{k,i}^p$.

(g) For $1 \leqslant i \leqslant p$, let

$$\tilde{Q}_p^i = \begin{bmatrix} -\tilde{q}_1^{\,i} & \tilde{q}_1^{\,i} & 0 & \cdots & 0 & 0 & 0 \\ \tilde{q}_2^{\,i} & -\tilde{q}_2^{\,i} - \tilde{q}_3^{\,i} & \tilde{q}_3^{\,i} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \tilde{q}_{2p-4}^{\,i} & -\tilde{q}_{2p-4}^{\,i} - \tilde{q}_{2p-3}^{\,i} & \tilde{q}_{2p-3}^{\,i} \\ 0 & 0 & 0 & \cdots & 0 & \tilde{q}_{2p-2}^{\,i} & -\tilde{q}_{2p-2}^{\,i} \end{bmatrix},$$

and $\tilde{R}_{0,i}^p = 1$. For $p \geqslant 2$ and $1 \leqslant i, k \leqslant p$, define

$$\tilde{R}_{k,i}^p = \sum_{i_1=1}^{2p-2k} \sum_{i_2=i_1+2}^{2p-2k+2} \cdots \sum_{i_k=i_{k-1}+2}^{2p-2} (-1)^k \tilde{q}_{i_1}^{\,i} \cdots \tilde{q}_{i_k}^{\,i}.$$

By Theorem 2(c) we have that, for $3 \leqslant k \leqslant p$ and $1 \leqslant i \leqslant p-1$,

$$\begin{aligned}
\frac{\partial R_k^p}{\partial q_{i,i+1}} &= -\sum_{\substack{1\leq l_1<\ldots<l_{k-2}\leq p\\ i\notin\{l_1,\ldots,l_{k-2}\}\\ \text{and } i+1\notin\{l_1,\ldots,l_{k-2}\}}} |Q_p([i+1, l_1, \ldots, l_{k-2}], [i, l_1, \ldots, l_{k-2}])| \\
&= -q_{i+1,i}\sum_{\substack{1\leq l_1<\ldots<l_{k-2}\leq p\\ i\notin\{l_1,\ldots,l_{k-2}\}\\ \text{and } i+1\notin\{l_1,\ldots,l_{k-2}\}}} |Q_p([l_1, \ldots, l_{k-2}], [l_1, \ldots, l_{k-2}])| \\
&= -q_{i+1,i}\sum_{\substack{1\leq l_1<\ldots<l_{k-2}\leq p\\ i\notin\{l_1,\ldots,l_{k-2}\}\\ \text{and } i+1\notin\{l_1,\ldots,l_{k-2}\}}} |\tilde{Q}_p^i([l_1, \ldots, l_{k-2}], [l_1, \ldots, l_{k-2}])| \\
&\quad - q_{i+1,i}\sum_{\substack{1\leq l_1<\ldots<l_{k-2}\leq p\\ i\notin\{l_1,\ldots,l_{k-2}\}\\ \text{and } i+1\in\{l_1,\ldots,l_{k-2}\}}} |\tilde{Q}_p^i([l_1, \ldots, l_{k-2}], [l_1, \ldots, l_{k-2}])| \\
&= -q_{i+1,i}\sum_{1\leq l_1<\ldots<l_{k-2}\leq p} |\tilde{Q}_p^i([l_1, \ldots, l_{k-2}], [l_1, \ldots, l_{k-2}])| \\
&= -q_{2i}\tilde{R}_{k-2,i}^p.
\end{aligned}$$

The proof for $\partial R_k^p/\partial q_{i+1,i}$ is similar.

Acknowledgement

We are grateful to the associate editor and a referee for helpful comments, and thank Wei-Hsiung Chao for helpful discussions and pointing out some typos in a previous version. We thank the National Science Council of the Republic of China (NSC 90-2118-M-001-039), and the RGC (Hong Kong) for partial support.

References

Arnold, L. (1974) *Stochastic Differential Equations*. New York: Wiley.

Brockwell, P.J. (1993) Threshold ARMA processes in continuous time. In H. Tong (ed.), *Dimension Estimation and Models*, pp. 170–190. River Edge, NJ: World Scientific.

Brockwell, P.J. and Stramer, O. (1995) On the approximation of continuous time threshold ARMA processes. *Ann. Inst. Statist. Math.*, **47**(1), 1–20.

Chan, K.S. and Munoz-Hernandez, B. (2003) A generalized linear model for repeated ordered categorical response data. *Statist. Sinica*, **13**, 207–226.

Chen, B. and Zadrozny, P.A. (2001) Analytic derivatives of the matrix exponential for estimation of linear continuous-time models. *J. Econom. Dynam. Control*, **25**, 1867–1879.

Cox, D.R. and Miller, H.D. (1965) *The Theory of Stochastic Processes*. London: Methuen.

Graham, A. (1981) *Kronecker Products and Matrix Calculus: with Applications*. Chichester: Ellis Horwood.

Hale, J. (1969) *Ordinary Differential Equations*. New York: Wiley-Interscience.

Horn, R.A. and Johnson, C.R. (1985) *Matrix Analysis*. Cambridge: Cambridge University Press.

Horn, R.A. and Johnson, C.R. (1991) *Topics in Matrix Analysis*. Cambridge: Cambridge University Press.

Jazwinski, A.H. (1970) *Stochastic Processes and Filtering Theory.* New York: Academic Press.

Jennrich, R.I. and Bright, P.B. (1976) Fitting systems of linear differential equations using computer generated exact derivatives. *Technometrics*, **18**, 385–392.

Kalbfleisch, J.D. and Lawless, J.F. (1985) The analysis of panel data under a Markov assumption. *J. Amer. Statist. Assoc.*, **80**, 863–871.

Moler, C.B. and Van Loan, C.F. (1978) Nineteen dubious ways to compute the exponential of a matrix. *SIAM Rev.*, **20**, 801–836.

Singer, H. (1995) Analytical score function for irregularly sampled continuous time stochastic processes with control variables and missing values. *Econometric Theory*, **11**, 721–735.

Van Loan, C.F. (1978) Computing integrals involving the matrix exponential. *IEEE Trans. Automat. Control*, **23**, 395–404.

Ward, R.C. (1977) Numerical computation of the matrix exponential with accuracy estimate. *SIAM J. Numer. Anal.*, **14**, 600–610.

Wilcox, R.M. (1967) Exponential operators and parameter differentiation in quantum physics. *J. Math. Phys.*, **8**, 962–982.

Received January 2002 and revised February 2003