A MODIFIED PRONY ALGORITHM FOR EXPONENTIAL FUNCTION FITTING

M. R. OSBORNE* AND G. K. SMYTH†

Abstract. A modification of the classical technique of Prony for fitting sums of exponential functions to data is considered. The method maximizes the likelihood for the problem (unlike the usual implementation of Prony's method, which is not even consistent for transient signals), proves to be remarkably effective in practice, and is supported by an asymptotic stability result. Novel features include a discussion of the problem parametrization and its implications for consistency. The asymptotic convergence proofs are made possible by an expression for the algorithm in terms of circulant divided difference operators.

Key words. Prony's method; Pisarenko's method; differential equations; difference equations; nonlinear least squares; inverse iteration; asymptotic stability; Levenberg algorithm; circulant matrices.

AMS subject classifications. 62J02 65D10 65U05

1. Introduction. Prony's method is a technique for extracting sinusoid or exponential signals from time series data, by solving a set of linear equations for the coefficients of the recurrence equation that the signals satisfy [24] [22] [17]. It is closely related to Pisarenko's method, which finds the smallest eigenvalue of an estimated covariance matrix [32]. Unfortunately, Prony's method is well known to perform poorly when the signal is imbedded in noise; Kahn et al [14] show that it is actually inconsistent. The Pisarenko form of the method is consistent but inefficient for estimating sinusoid signals and inconsistent for estimating damped sinusoids or exponential signals.

A modified Prony algorithm that is equivalent to maximum likelihood estimation for Gaussian noise was originated by Osborne [28]. It was generalized in [39] [30] to estimate any function which satisfies a difference equation with coefficients linear and homogeneous in the parameters. Osborne and Smyth [30] considered in detail the special case of rational function fitting, and proved that the algorithm is asymptotically stable in that case. This paper considers the application to fitting sums of exponential functions.

The modified Prony algorithm for exponential fitting will estimate, for fixed p, any function μ that solves a constant coefficient differential equation

$$\sum_{k=1}^{p+1} \xi_k \mathcal{D}^{k-1}\mu = 0 \tag{1}$$

where $\mathcal{D}$ is the differential operator. Perturbed observations, $y_i = \mu(t_i) + \epsilon_i$, are made at equi-spaced times t_i, $i = 1, \ldots, n$, where the ϵ_i are independent with mean zero and variance σ^2. The solutions to (1) include complex exponentials, damped and undamped sinusoids and real exponentials, depending on the roots of the polynomial with the ξ_k as coefficients. The modified Prony algorithm has the great practical advantage that it will estimate any of these functions according to which best fits the available observations.

* Statistics Research Section, School of Mathematics, Australian National University, GPO Box 4, Canberra, ACT 2601, Australia.

† Department of Mathematics, University of Queensland, Q 4072, Australia.

Although the algorithm estimates all functions in the same way, the practical considerations and asymptotic arguments differ depending on whether the signals are periodic or transient, real or complex. This paper therefore focuses on the specific problem of fitting a sum of real exponential functions

$$\mu(t) = \sum_{j=1}^{p} \alpha_j e^{-\beta_j t} \tag{2}$$

to real data. The α_j and β_j will be assumed real, the β_j distinct and generally non-negative. This paper is mainly concerned with proving the asymptotic stability of the algorithm, but several practical issues are also addressed. The algorithm has been applied elsewhere to real sinusoidal signals [21] [14] and to exponentials with imaginary exponents in complex noise [18].

Real exponential fitting is one of the most important, difficult and frequently occuring problems of applied data analysis. Applications include radioactive decay [38], compartment models [2, Chapter 5] [37, Chapter 8], and atmospheric transfer functions [46]. Estimation of the α_j and β_j is well known to be numerically difficult [19, p. 276] [43] [37, Section 3.4]. General purpose algorithms often have great difficulty in converging to a minimum of the sums of squares. This can be caused by difficulty in choosing initial values, ill-conditioning when two or more β_j are close, and other less important difficulties associated with the fact that the ordering of the β_j is arbitrary. The modified Prony algorithm solves the problem of ordering the β_j and is relatively insensitive to starting values. It also solves the ill-conditioning problem as far as convergence of the algorithm is concerned, but may return a pair of damped sinusoids in place of two exponentials which are coalescing.

In some applications the restriction to positive coefficients α_j is natural. A convex cone characterization is then possible, and special algorithms have been proposed in [6] [46] [15] [10] [35]. We prefer to treat the general problem with freely varying coefficients since this is appropriate for most compartment models. A common attempt to reduce the difficulty of the general problem has been to treat it as a separable regression, i.e., to estimate the coefficients by linear least squares conditional on the rate constants β_j as in [44] [20] [1] [12] [16] [40]. Another approach has been suggested by Ross [34, Section 3.1.4] who suggests that the coefficients of the differential equation (1) comprise a more "stable" parametrization of the problem than do the parameters of (2). Both of these strategies are part of the modified Prony algorithm.

The modified Prony algorithm uses the fact that the $\mu(t_i)$ satisfy an exact difference equation when the t_i are equally spaced. The algorithm directly estimates the coefficients, γ_k say, of this difference equation. In Section 3 it is shown that the residual sum of squares after estimating the α_j can be written in terms of the γ_k. The derivative with respect to $\gamma = (\gamma_1, \ldots, \gamma_{p+1})^T$ can then be written as $2B(\gamma)\gamma$, where B is a symmetric matrix function of γ. The modified Prony algorithm finds the eigenvector of $B(\gamma)\gamma = \lambda\gamma$ corresponding to $\lambda = 0$ by the fixed point iteration in which γ^{k+1} is the eigenvector of $B(\gamma^k)$ with eigenvalue nearest zero. The eigenvalue λ is the Lagrange multiplier for the scale of γ in the homogeneous difference equation. Inverse iteration proves very suitable for the actual computations.

Jennrich [13] shows that, under general conditions, least squares estimators are asymptotically normal and unbiased with covariance matrix of $O(n^{-1})$. Under the same conditions the Gauss-Newton algorithm is asymptotically stable at the least squares estimates. For the results to apply here, it is necessary that the empirical distribution function of the t_i should have a limit as $n \to \infty$. Since the t_i are equally spaced, it is sufficient that they lie in an interval independent of n. Without loss of generality we take this interval to be the unit interval and assume that $t_i = i/n$, $i = 1, \ldots, n$.

Difference equation parametrizations for the exponential functions are discussed in the next section. The modified Prony algorithm is given in Section 3. It is compared with Prony's method and another algorithm of Prony type, and the equivalence of the various parametrizations is discussed. The algorithm is shown to be asymptotically stable, and $\sigma^2 B(\hat{\gamma})^+$ is shown to estimate the asymptotic covariance matrix of the least squares estimator $\hat{\gamma}$ in Section 4. Section 5 shows how the algorithm can accommodate linear constraints on the γ_j. Such constraints may serve for example to include a constant term in $\mu(t)$ or to constrain it to be a sum of undamped sinusoids. A small simulation study is included in Section 6 to illustrate the asymptotic results and to compare the modified Prony algorithm with the Levenberg algorithm.

The asymptotic convergence proofs involve lengthy technical arguments and are relegated to the appendix. The proofs are made possible by an expression for the algorithm in terms of circulant divided difference operators. Circulant methods have often been applied to differential and difference equations, for example in [45] [26] [11] [7]. The theory of circulant matrices was put on a firm basis with the work of Davis [8]. The methods are used here somewhat differently, to compute certain matrix multiplications analytically.

2. Difference and Recurrence Equations. Suppose that $\mu(t)$ satisfies the constant coefficient differential equation (1), and that the polynomial

$$p_\xi(z) = \sum_{k=1}^{p+1} \xi_k z^{k-1} \tag{3}$$

has distinct roots $-\beta_j$ with multiplicities m_j for $j = 1, \ldots, s$. Then (1) may be rewritten as

$$\prod_{j=1}^{s} (\mathcal{D} + \beta_j I)^{m_j} \mu(t) = 0.$$

The general solution for μ may be expressed as

$$\mu(t) = \sum_{j=1}^{s} \sum_{k=1}^{m_j} \alpha_{jk} t^{k-1} e^{-\beta_j t}$$

writing α_{jk} for the coefficients of the fundamental solutions. The roots β_j may in general include complex pairs. If so, then the real part of μ will contain linear combinations of damped trigonometric functions $e^{-\mathrm{Re}\beta_j t} \sin(\mathrm{Im}\beta_j t)$ and $e^{-\mathrm{Re}\beta_j t} \cos(\mathrm{Im}\beta_j t)$.

Now consider discrete approximations to the differential equation. Let Π be the forward shift operator defined by $\Pi\mu(t) = \mu(t + \frac{1}{n})$, and let Δ be the divided difference operator $\Delta = n(\Pi - I)$. It is easy to verify that the operator $(\Delta + \zeta_j I)^m$ with $\zeta_j = n(1 - e^{-\beta_j/n})$ annihilates the term $t^{m-1}e^{-\beta_j t}$. Therefore μ also satisfies the difference equation

$$\prod_{j=1}^{s} (\Delta + \zeta_j I)^{m_j} \mu(t) = 0,$$

which can be written as

$$\sum_{k=1}^{p+1} \gamma_k \Delta^{k-1} \mu(t) = 0 \tag{4}$$

for some suitable choice of γ_k. The γ_k will be called the *difference form* Prony parameters. The ζ_j and γ_k represent discrete approximations to the β_j and ξ_k respectively, in the sense that $\zeta_j \to \beta_j$ and $\gamma_k \to \xi_k$ as $n \to \infty$.

For some purposes a simpler discrete approximation is that in terms of the forward shift operators. The function $t^{m-1}e^{-\beta_j t}$ is also annihilated by the operator $(\Pi - \rho_j I)^m$ with $\rho_j = e^{-\beta_j/n}$. Therefore μ also satisfies the recurrence equation

$$\prod_{j=1}^{s} (\Pi - \rho_j I)^{m_j} \mu(t) = 0$$

which can be written as

$$\sum_{k=1}^{p+1} \delta_k \Pi^{k-1} \mu(t) = 0 \tag{5}$$

for some δ_k. We call the δ_k the *recurrence form* Prony parameters. Since $\rho_j \to 1$ as $n \to \infty$, the δ_k must converge to some multiple of the binomial coefficients $(-1)^{p-k+1}\binom{p}{k-1}$, a limit which is independent of the β_j.

The relationship between the difference and recurrence parameters can be exhibited by equating

$$\sum_{k=1}^{p+1} \gamma_k \Delta^{k-1} = \sum_{k=1}^{p+1} c_k \Pi^{k-1},$$

which when solved for the c_k gives

$$c_j = \sum_{k=j}^{p+1} (-1)^{k-j} \binom{k-1}{j-1} n^{k-1} \gamma_k.$$

That is, $\mathbf{c} = U\boldsymbol{\gamma}$ where U is the nonsingular matrix

$$U = \begin{pmatrix} 1 & -1 & 1 & \cdots & & (-1)^p \\ & 1 & -2 & & & \\ & & 1 & & & \\ & & & \ddots & & \vdots \\ & & & & 1 & -\binom{p}{1} \\ & & & & & 1 \end{pmatrix} \begin{pmatrix} 1 & & & \\ & n & & \\ & & \ddots & \\ & & & n^p \end{pmatrix} \tag{6}$$

and $\mathbf{c} = (c_1, \ldots, c_{p+1})^T$. Obviously $\mathbf{c}$ and $\boldsymbol{\delta}$ are re-scaled versions of one another. The notational convention will be used that $\mathbf{c}$ represents the above function of $\boldsymbol{\gamma}$ while $\boldsymbol{\delta}$ is a function of the rate constants β_j with elements scaled to be $O(1)$.

For the reasons given in the introduction, we will henceforth assume that the roots of $p_\xi(\cdot)$ are distinct and real, so that the general solution for $\mu(t)$ collapses to the sum of real exponential functions (2). The coefficients of the differential equation

ξ_k can then be expressed as the elementary symmetric functions of the β_j. If the ξ_k are scaled so that $\xi_{p+1} = 1$, then the ξ_k are given by

$$\xi_k = \sum_{j=1}^{\binom{p}{p-k+1}} \prod_{\ell \in J_{k,j}} \beta_\ell$$

where $J_{k,j}$ for $j = 1, \ldots, \binom{p}{p-k+1}$ are the possible sets of size $p - k + 1$ drawn from $\{1, \ldots, p\}$. Write this as $\boldsymbol{\xi} = \mathcal{S}(\boldsymbol{\beta})$, after gathering the ξ_k and the β_j into respective vectors. Similarly, in an obvious notation, $\gamma = \mathcal{S}(\boldsymbol{\zeta})$, if γ is scaled so that $\gamma_{p+1} = 1$, and $\boldsymbol{\delta} = \mathcal{S}(-\boldsymbol{\rho})$, if $\boldsymbol{\delta}$ is scaled so that $\delta_{p+1} = 1$.

3. A Modified Prony Algorithm.

3.1. Nonlinear Eigenproblem. Let $\mu_i = \mu(t_i)$, $i = 1, \ldots, n$, let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$ and let X_δ be the $n \times (n-p)$ matrix

$$X_\delta = \begin{pmatrix} \delta_1 & & \\ \vdots & \ddots & \\ & & \delta_1 \\ \delta_{p+1} & & \\ & \ddots & \vdots \\ & & \delta_{p+1} \end{pmatrix}$$

where the δ_k are the recurrence parameters. Then $\boldsymbol{\mu}$ satisfies

$$X_\delta^T \boldsymbol{\mu} = 0$$

which is the matrix version of the recurrence equation (5). Alternatively, we can substitute c_k for δ_k in X_δ and write the resulting matrix X_γ as a function of the difference parameters γ using $\mathbf{c} = U\gamma$. Then

$$X_\gamma^T \boldsymbol{\mu} = 0$$

is the matrix version of the difference equation (4).

We now treat the exponential fitting problem as a separable regression, and use the above matrix equations to give an expression for the reduced sum of squares. Let A be the $n \times p$ matrix function of $\boldsymbol{\beta}$ with elements $A_{ij} = e^{-\beta_j t_i}$, and write

$$\boldsymbol{\mu} = A(\boldsymbol{\beta})\boldsymbol{\alpha}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^T$. Then A is orthogonal to both X_δ and X_γ, and, if the β_j are distinct, all matrices are of full column rank. Let $\mathbf{y} = (y_1, \ldots, y_n)^T$ be the vector of observations. The sum of squares

$$\phi(\boldsymbol{\alpha}, \boldsymbol{\beta}) = (\mathbf{y} - \boldsymbol{\mu})^T(\mathbf{y} - \boldsymbol{\mu})$$

is minimized with respect to $\boldsymbol{\alpha}$ by

$$\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}) = (A^T A)^{-1} A^T \mathbf{y}.$$

Substituting $\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta})$ into ϕ gives the reduced sum of squares

$$\begin{aligned} \psi(\boldsymbol{\beta}) &= \phi(\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}), \boldsymbol{\beta}) \\ &= \mathbf{y}^T(I - P_A)\mathbf{y} \end{aligned}$$

where P_A is the orthogonal projection onto the column space of A. The function ψ is the variable projection functional defined by Golub and Pereyra [12]. We can reparametrize to the Prony parameters by writing

$$\psi = \mathbf{y}^T P_X \mathbf{y} \tag{7}$$

where P_X is the orthogonal projection onto the common column space of X_δ and X_γ. Then ψ is a function of either δ or γ.

We now solve the least squares problem with respect to the Prony parameters. The derivative of ψ with respect to γ can be written

$$\dot{\psi}_\gamma = 2B_\gamma(\gamma)\gamma$$

where B_γ is the symmetric $(p+1)\times(p+1)$ matrix function of γ with elements

$$B_{\gamma ij} = \mathbf{y}^T X_{\gamma i}(X_\gamma^T X_\gamma)^{-1}X_{\gamma j}^T\mathbf{y} - \mathbf{y}^T X_\gamma(X_\gamma^T X_\gamma)^{-1}X_{\gamma i}^T X_{\gamma j}(X_\gamma^T X_\gamma)^{-1}X_\gamma^T\mathbf{y} \tag{8}$$

and where $X_{\gamma j} = \partial X_\gamma/\partial\gamma_j$. Each $X_{\gamma j}$ is a constant matrix representing the $j-1$th order divided difference operator [30]. Since the scale of γ is disposable, we will adjoin the condition $\gamma^T\gamma = 1$ so that the components of γ are $O(1)$. A necessary condition for a minimum of (7) subject to the constraint is

$$(B_\gamma(\gamma) - \lambda I)\gamma = 0 \tag{9}$$

where λ is the Lagrange multiplier for the constraint. This condition corresponds to a special case of the problem considered by Mittleman and Weber [23] and described by them as a nonlinear eigenvalue problem. This is not the usual form of nonlinear eigenvalue problem in which the nonlinearity is in the eigenvalue λ only, and it appears to have been little studied otherwise. Our special case possesses one feature of the ordinary eigenvalue problem not enjoyed by the general form considered in [23]. Solutions to (9) are independent of change of scale, and as a further consequence the corresponding eigenvalues satisfy $\lambda = 0$. This follows because $\psi(\gamma)$ is homogeneous of degree zero in γ. In [30] it is shown that this implies $\gamma^T B_\gamma\gamma = 0$ at all points at which $\dot{\psi}(\gamma)$ is well defined, and this implies the result.

The modified Prony algorithm solves (9) using a succession of linear problems converging to $\lambda = 0$. Given an estimate γ^k of the solution $\hat{\gamma}$, solve

$$\begin{aligned}[B_\gamma(\gamma^k) - \lambda^{k+1}I]\gamma^{k+1} &= 0\\ {\gamma^{k+1}}^T\gamma^{k+1} &= 1\end{aligned}$$

with λ^{k+1} the nearest to zero of such solutions. Convergence is accepted when λ^{k+1} is small compared to $\|B_\gamma\|$. Inverse iteration has proved very satisfactory for solving the linear eigenproblems. A detailed algorithm is given in [30].

An exactly analogous version of the algorithm can be developed in terms of the recurrence parameters. The derivative of ψ with respect to $\boldsymbol{\delta}$ is

$$\dot{\psi}_\delta = 2B_\delta(\boldsymbol{\delta})\boldsymbol{\delta}$$

where B_δ is as for B_γ with X_δ replacing X_γ and with the shift operator $X_{\delta j} = \partial X_\delta/\partial\delta_j$ replacing $X_{\gamma j}$. Up to a scale factor, B_δ is $U^T B_\gamma U$ where U is given by (6). The difference and recurrence versions of the algorithm are distinct algorithms, but share the same stationary values. The recurrence version was the original algorithm developed in [28]. In this paper most emphasis will be given to the difference version because of its suitability for asymptotic arguments.

Some care is needed in considering the reparametrization from $\boldsymbol{\beta}$ to the Prony parameters. The Prony parametrizations are more general, since the difference and recurrence equations may yield more general solutions, possibly including repeated roots and damped trigonometric functions, than the sum of exponential functions given by (2). Since $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ may take values for which there is no corresponding sum of exponentials, solving the least squares problem with respect to the Prony parameters as above is not necessarily equivalent to solving with respect to $\boldsymbol{\beta}$. Theorem 3.1 which is proved in [28] and [39] shows that the Prony parametrization does in fact solve the exponential fitting problem, in the sense that if $\boldsymbol{\beta}$ minimizes the sum of squares then the corresponding elementary symmetric functions give Prony parameters which satisfy (9).

THEOREM 3.1. *Let $\gamma(\boldsymbol{\beta}) = \|\mathbf{s}\|^{-1}\mathbf{s}$, with $\mathbf{s} = \mathcal{S}(\boldsymbol{\zeta})$ and $\boldsymbol{\zeta} = n(1 - e^{-\boldsymbol{\beta}/n})$. If $\boldsymbol{\beta}$ solves $\dot{\psi}_{\beta} = 0$, and the β_j are distinct, then $\gamma(\boldsymbol{\beta})$ solves $\dot{\psi}_{\gamma} = 0$.*

3.2. Other Algorithms. Prony's classical method for exponential fitting consisted of solving the linear system

$$X_\delta^T \mathbf{y} = 0$$

with respect to $\boldsymbol{\delta}$ to interpolate p exponentials through $2p$ points. The direct generalization to the overdetermined case, which consists of minimizing the sum of squares

$$\mathbf{y}^T X_\delta \, X_\delta^T \mathbf{y} \tag{10}$$

subject to $\delta_{p+1} = 1$ is now called Prony's Method. Minimizing (10) subject to $\boldsymbol{\delta}^T \boldsymbol{\delta} = 1$ is equivalent to finding the smallest eigenvalue of the $(p+1) \times (p+1)$ matrix M with components

$$M_{ij} = \mathbf{y}^T X_{\delta i} X_{\delta j}^T \mathbf{y},$$

and this is often called Pisarenko's Method or the Covariance Method. Applications and references are given in [22]. Because of their simplicity, the methods of Prony and Pisarenko have enjoyed considerable popularity over the last three decades, and the techniques have been adapted to other problems, for example [4] [42] [33] [9].

Comparing with (7) it can be seen that (10) ignores the factor $(X_\delta^T X_\delta)^{-1}$ in the objective function. While Prony's and Pisarenko's methods are consistent as $\sigma^2 \downarrow 0$, Kahn et al [14] show that neither algorithm is consistent as $n \uparrow \infty$ for estimating exponentials or damped sinusoids. The methods are useful only for low noise levels regardless of how many observations are available. For estimating pure sinusoids, Kahn et al show that Pisarenko's method is consistent but not efficient while Prony's remains inconsistent; see also [36] [41] [21].

Another attempt is that of Osborne [27] and Bresler and Macovski [5]. They identify the correct objective function (7), and propose an eigenvalue iteration for minimizing it. However they apply reweighting to the objective function rather than to a modification of the necessary conditions (9), and this has the effect of ignoring the second term in the expression (8) for B_δ. Their algorithm does not minimize ψ, but does give consistent estimators of transient signals if a particular choice of scale is made; see [14].

3.3. Calculation of B. A general scheme for the calculation of the matrix B_γ is given in [30]. Some simplification occurs for exponential fitting. The $X_{\gamma j}^T \mathbf{y}$ are the divided differences of the y_i of successive orders. Similarly for the $X_{\gamma j}\mathbf{w}$ where $\mathbf{w} = (X_\gamma^T X_\gamma)^{-1} X_\gamma^T \mathbf{y}$. The matrix X_γ is Toeplitz as well as banded, so only $p+1$ elements need to be stored; these are calculated from $\mathbf{c} = U\gamma$ with U defined by (6). The matrix $X_\gamma^T X_\gamma$ is banded, as is its Choleski decomposition, with p sub-diagonal bands.

The calculation of B_δ is even simpler. The elements of X_δ require no calculation given $\boldsymbol{\delta}$. The $X_{\delta j}^T \mathbf{y}$ are simply windowed shifts of $\mathbf{y}$, and

$$\mathbf{y}^T X_\delta (X_\delta^T X_\delta)^{-1} X_{\delta i}^T X_{\delta j}^T (X_\delta^T X_\delta)^{-1} X_\delta^T \mathbf{y} = \sum_{k=1}^{n-p-|i-j|} v_k v_{k+|i-j|}$$

where the v_k are the components of $\mathbf{v} = (X_\delta^T X_\delta)^{-1} X_\delta^T \mathbf{y}$. See also [28]. The simplicity of the recurrence form tempts one to calculate the difference form Prony matrix from it by $B_\gamma = U^{-T} B_\delta U^{-1}$. This turns out to be equivalent to

$$B_{\gamma ij} = n^{-2p} (\Delta^{i-1T} B_\delta \Delta^{j-1})_{ij}$$

where Δ is the divided difference operator. Unfortunately the elements of B_δ are large and nearly equal, so this calculation involves considerable subtractive cancellation and is not recommended.

3.4. Recovery of the rate constants. Having estimated γ or $\boldsymbol{\delta}$, we can obtain $\boldsymbol{\mu}$ directly from equation (5) of [30]. Usually though it will be necessary to recover the rate constants β_j for the purpose of interpretation. Given the recurrence form parameters we solve $p_\delta(z) = 0$ to obtain roots $\rho_j = e^{-\beta_j/n}$. For large n this is an ill-conditioned problem because the ρ_j cluster near 1. Another aspect of the same problem is that asymptotically the leading significant figures of the δ_k contain no information about the β_j.

This problem does not arise in the difference formulation. Given the difference parameters we solve $p_\gamma(z) = 0$ to obtain roots $\zeta_j = n(1 - e^{-\beta_j/n})$, and the final step

$$\begin{aligned} \beta_j &= -n \log(1 - \zeta_j/n) \\ &= n \sum_{j=1}^{\infty} j^{-1} (\zeta_j/n)^j \end{aligned}$$

will cause problems only in the unlikely event that ζ_j is large and negative. Unfortunately a non-negligible amount of subtractive cancellation does occur in another part of the difference form calculations, namely when forming the $X_{\gamma j}^T \mathbf{y}$ in the calculation of B_γ.

4. Asymptotic Stability. The key result for stability of the modified Prony is the convergence of the matrix $\frac{1}{n} B_\gamma$ to a positive semi-definite limit. The expectation of B_γ is σ^2 times the Fisher information matrix for γ given $\boldsymbol{\alpha}$, namely $E[\sigma^2 \ddot{\psi}_\gamma/2]$. This is shown in Section 7 of [30] to be $\dot{\boldsymbol{\mu}}_\gamma^T P_X \dot{\boldsymbol{\mu}}_\gamma$, where $\dot{\boldsymbol{\mu}}$ is the gradient matrix of $\boldsymbol{\mu}$ with respect to γ.

THEOREM 4.1.

$$\frac{1}{n} B_\gamma(\hat{\gamma}) \xrightarrow{a.s.} \mathcal{J}_0$$

as $n \to \infty$ where

$$\mathcal{J}_0 = \lim_{n\to\infty} \frac{1}{n} \dot{\boldsymbol{\mu}}_\gamma^T P_X \dot{\boldsymbol{\mu}}_\gamma(\gamma_0)$$

is σ^2 times the limiting Fisher information per observation. Also $\mathcal{J}_0$ is positive semi-definite with null space spanned by γ.

Among the consequences of this theorem are that the Moore-Penrose inverse of $\frac{1}{n}B_\gamma(\hat{\gamma})$ estimates the asymptotic covariance matrix of $\sqrt{n}\hat{\gamma}/\sigma$, and that the zero eigenvalue $B_\gamma(\hat{\gamma})$ is asymptotically isolated with multiplicity one.

It is show in [30] that the derivative at $\hat{\gamma}$ of the iteration function defined by the modified Prony algorithm is

$$G_n(\gamma) = B_\gamma(\gamma)^+ \dot{B}_\gamma(\gamma)\gamma.$$

The algorithm is linearly convergent with limiting contraction factor given by the spectral radius of $G_n(\hat{\gamma})$ [31]. Theorem 4.2 combines Theorem 4.1 with the result that $\frac{1}{n}\dot{B}_\gamma(\hat{\gamma})\hat{\gamma} \to 0$.

THEOREM 4.2.

$$G_n(\hat{\gamma}) \xrightarrow{a.s.} 0$$

as $n \to \infty$.

A corollary is that the algorithm is asymptotically stable at $\hat{\gamma}$. The spectral radius of $G_n(\hat{\gamma})$ can in fact be shown to be $O(1/\sqrt{n})$ in probability.

Theorem 4.2 applies also to the recurrence version of the algorithm, as can be seen from Theorem 4.1 of [30]. There is however no corresponding recurrence version of Theorem 4.1. The recurrence matrix B_δ has in fact a very interesting eigenstructure dominated by powers of n. Let H be the $p \times p$ matrix with elements

$$H_{ij} = (-1)^{i-j}\binom{j-1}{i-1}$$

for $i \le j$ and 0 otherwise. Then

$$U = H \begin{pmatrix} 1 & & & \\ & n & & \\ & & \ddots & \\ & & & n^p \end{pmatrix}$$

so that

$$B_\delta = H^{-T} \begin{pmatrix} n^p & & & \\ & \ddots & & \\ & & n & \\ & & & 1 \end{pmatrix} B_\gamma \begin{pmatrix} n^p & & & \\ & \ddots & & \\ & & n & \\ & & & 1 \end{pmatrix} H^{-1}.$$

Let f_k be the polynomial of degree $k-1$ which satisfies $f_k(i) = 0$, for $i = 1, \dots, k-1$, and $f_k(k) = 1$. Let $\mathbf{h}_k = (f_k(1), \cdots, f_k(p+1))^T$. Then $H^T\mathbf{h}_k = \mathbf{e}_k$ is the kth coordinate vector, so the $\mathbf{h}_k$ are the columns of H^{-T}, and

$$\begin{aligned} B_\delta &= \sum_{i,j=1}^{p+1} n^{p-i+1} n^{p-j+1} H^{-T}\mathbf{e}_i\mathbf{e}_i^T B \mathbf{e}_j \mathbf{e}_j^T H^{-1} \\ &= \sum_{i,j=1}^{p+1} n^{p-i+1} n^{p-j+1} B_{ij}\mathbf{h}_i\mathbf{h}_j^T . \end{aligned}$$

The following argument shows that, while B_δ has a zero eigenvalue when evaluated at $\hat{\boldsymbol{\delta}}$, the other eigenvalues have orders which are increasing odd powers of n.

Firstly, for large n, all proper submatrices of $B_\gamma(\hat{\gamma})$ are nonsingular. This follows because $\hat{\gamma}_1$ and $\hat{\gamma}_{p+1}$ are nonzero (none of the true rate constants β_{0j} may be zero) and $\hat{\gamma}$ spans the null space of $B_\gamma(\hat{\gamma})$. In particular, the diagonal elements of $B_\gamma(\hat{\gamma})$ are nonzero—from Theorem 4.1 they are $O(n)$. Let $\mathbf{x}_1, \ldots, \mathbf{x}_{p+1}$ be the orthonormal sequence obtained from $\mathbf{h}_1, \ldots, \mathbf{h}_{p+1}$ by Gram-Schmidt orthonormalization. This is equivalent to

$$\mathbf{x}_k = (HH^T)^{1/2}\mathbf{h}_k$$

where $(HH^T)^{1/2}$ is the Choleski factor of HH^T. The largest and smallest eigenvalues are given by the extreme values of the Rayleigh quotient:

$$\lambda_1 = \max_{\mathbf{z}^T\mathbf{z}=1} \mathbf{z}^T B_\delta \mathbf{z} \ , \quad \lambda_{p+1} = \min_{\mathbf{z}^T\mathbf{z}=1} \mathbf{z}^T B_\delta \mathbf{z} \ .$$

Asymptotically, these are achieved by $\mathbf{z} = \mathbf{x}_1 = (p+1)^{-1/2}\mathbf{1}$ giving $\lambda_1 = O(n^{2p+1})$, and $\mathbf{z} = \mathbf{x}_{p+1}$ giving $\lambda_{p+1} = O(n)$. Defining the remaining eigenvalues recursively, the kth eigenvalue of B_δ in decreasing order is asymptotically equal to

$$\max_{\substack{\mathbf{z}^T\mathbf{x}_j=0,\ j<k \\ \mathbf{z}^T\mathbf{z}=1}} \mathbf{z}^T B_\delta \mathbf{z}$$

which is asymptotically achieved by $\mathbf{z} = \mathbf{x}_k$, and is $O(n^{2p+1-2(k-1)})$.

5. Including a Linear Constraint. Two methods of handling one or more linear constraints on the β_j are considered. The first is convenient with the recurrence form algorithm. The second is convenient when including a constant term with the difference form algorithm.

Suppose prior information about γ can be expressed as the linear constraint $\mathbf{g}^T\gamma = 0$. For example the constant term model

$$\mu(t) = \alpha_1 + \sum_{j=2}^{p} \alpha_j e^{-\beta_j t} \tag{11}$$

corresponds to $\beta_1 = 0$ and hence to $\mathbf{e}_1^T\gamma = 0$ in the difference formulation or $\mathbf{1}^T\boldsymbol{\delta} = 0$ in the recurrence formulation. The appropriate objective function is

$$F(\gamma, \lambda, \nu) = \psi(\gamma) + \lambda(1 - \gamma^T\gamma) + 2\nu s\gamma^T\mathbf{g}$$

where λ and ν are Lagrange multipliers, and s is a scale factor chosen for numerical conditioning. Differentiating gives

$$\begin{aligned} \dot{F}_\gamma &= 2B_\gamma(\gamma)\gamma - 2\lambda\gamma + 2\nu s\gamma \\ \dot{F}_\lambda &= 1 - \gamma^T\gamma \\ \dot{F}_\nu &= 2s\gamma^T\mathbf{g} \ . \end{aligned}$$

The necessary conditions for a minimum may be summarized as the generalized eigenproblem

$$(A - \lambda P)\mathbf{v} = 0 \ , \qquad \mathbf{v}^T P\mathbf{v} = 1 \tag{12}$$

with

$$A = \begin{pmatrix} B_\gamma & s\mathbf{g} \\ s\mathbf{g}^T & 0 \end{pmatrix}, \quad \mathbf{v} = \begin{pmatrix} \gamma \\ \nu \end{pmatrix} \text{and} \quad P = \begin{pmatrix} I_p & 0 \\ 0 & 0 \end{pmatrix}.$$

Premultiplying the equation $\dot{F}_\gamma = 0$ by γ^T shows that λ must be zero at a solution of (12). The eigenproblem is solved by solving the sequence of linear problems

$$(13) \qquad \left(A(\gamma^k) - \lambda^{k+1} P\right)\mathbf{v}^{k+1} = 0\,, \qquad \mathbf{v}^{k+1T} P \mathbf{v}^{k+1} = 1\,.$$

This modifies the detailed algorithm given in Section 5 of [30]. The inverse iteration sequence, which finds the eigenvalue of $A(\gamma^k)$ closest to zero, now becomes

$l := 1$
$\eta^l := 0$
$\mathbf{v}^l :=$ current estimate of γ
repeat (inverse iteration)
  $\mathbf{w}^{l+1} := (A - \eta^l P)^{-1} P\mathbf{v}^l$
  $\mathbf{v}^{l+1} := P\mathbf{w}^{l+1} / \|P\mathbf{w}^{l+1}\|_\infty$
  $\mathbf{w}^{l+2} := (A - \eta^l P)^{-1}\mathbf{v}^{l+1}$
  $\eta^{l+2} := \eta^l + \mathbf{w}^{l+2T}\mathbf{v}^{l+1} / \mathbf{w}^{l+2T}\mathbf{w}^{l+2}$
  $\mathbf{v}^{l+2} := \mathbf{w}^{l+2} / \|P\mathbf{w}^{l+2}\|_2$
  $l := l + 2$
until $|\eta^l - \eta^{l-2}| < \epsilon$.

The eigenvalues of (13) are unaffected by s, since

$$\begin{aligned} \det(A - \lambda P) &= \det \begin{pmatrix} B_\gamma - \lambda I & s\mathbf{g} \\ s\mathbf{g}^T & 0 \end{pmatrix} \\ &= s^2 \det \begin{pmatrix} B_\gamma - \lambda I & \mathbf{g} \\ \mathbf{g}^T & 0 \end{pmatrix}, \end{aligned}$$

so we can take s to have a scale comparable to the elements of B_γ without affecting the rate of convergence of the iteration. The determinant is a polynomial in λ of order only q, so the constraint has reduced the dimension of the eigenproblem. This technique, with δ and B_δ replacing γ and B_γ throughout, is used in Section 6 to fit models of the form (11) using the recurrence form algorithm.

An alternative approach to the constraint is to explicitly deflate the dimension of B_γ. Let W be a $(p+1) \times p$ matrix of full rank satisfying $W^T\mathbf{g} = 0$. We can set $W^T W = I$. Then $W\boldsymbol{\tau} = \gamma$ and (12) is equivalent to

$$(W^T B_\gamma W - \lambda I)\boldsymbol{\tau} = 0\,, \quad \boldsymbol{\tau}^T\boldsymbol{\tau} = 1.$$

If $\mathbf{g} = \mathbf{e}_1$ then W can be chosen as $W = [0\ I_p]^T$ so that $W^T B_\gamma W$ is simply the trailing $p \times p$ matrix of B_γ. This technique has been used to fit models of the form (11) using the difference form algorithm.

6. A Numerical Experiment. Osborne [28] gave an example of the modified Prony algorithm on a real data set, showing excellent convergence behaviour. This section compares the modified Prony algorithm with a good general purpose nonlinear least squares procedure, namely the Levenberg modification of the Gauss-Newton algorithm, on a simulated problem. The modified Prony algorithm was implemented in its recurrence form with the augmentation of Section 5. The Levenberg algorithm was implemented essentially as described by [29], the Levenberg parameter having expansion factor 2, contraction factor 10 and initial value of 1. The tolerance parameter which determines the precision of the estimates required—roughly, the relative change in the root sum of squares—was set to 10^{-7}. Although the Prony and Levenberg convergence criteria are not strictly comparable, the Prony tolerance parameter was adjusted to 10^{-15} so that the two algorithms returned estimates on average of the same precision.

Data was simulated using the mean function $\mu(t) = .5+2e^{-4t}-1.5e^{-7t}$. Data sets were constructed as described in [30] to have standard deviations $\sigma = .03, .01, .003, .001$ and sample sizes $n = 32, 64, 128, 256, 512$. Ten replicates were generated for each of the four distributions: the normal, student's t on 3 d.f. (infinite third moments), log-normal (skew) and Pareto's distribution with $k = 1$ and $\alpha = 3$ (skew and infinite third moments). Uniform deviates were generated from the NAG subroutine G05CAF (Numerical Algorithms Group, 1983) with seed 1984, the first 200 values being discarded for seed independence.

To remove subjectivity, the true parameter values themselves were used as starting values. These were quite far from the least squares estimates for small n and large σ, less so for large n and small σ, as can be seen from Table 3.

The modified Prony convergence results were almost identical for the four distributions. Apparently it is little affected by skewness or by the third and higher moments of the error distribution (although the actual least squares estimates returned are affected). The Levenberg algorithm was adversely affected by non-normality for $n \leq 64$ but was unaffected for $n \geq 128$. Only the results for the normal distribution are reported in detail.

TABLE 1

Median and maximum iteration counts, and number of failures, for exponential fitting. Results for the Prony algorithm are above those for the Levenberg algorithm.

$n\backslash\sigma$	0.030			0.010			0.003			0.001		
32	6	11	6	4	6	5	3	4	1	3	3	0
	40	40	6	33	40	5	26	40	4	16	40	1
64	4	8	5	3	4	5	2	3	1	2	2	0
	32.5	40	5	31.5	40	5	20	40	2	13	22	1
128	3	3	2	2	3	2	2	2	0	1.5	2	0
	16.5	40	2	10	40	2	8	34	0	6	18	0
256	2	3	4	2	2	3	1	1	0	1	1	0
	30	30	4	20	40	4	14	32	1	10	12	1
512	1	1	4	1	1	1	1	1	0	1	1	0
	36.5	40	5	19.5	40	3	13	22	0	7.5	12	0

As Table 1 shows, the Prony algorithm required dramatically fewer iterations than the Levenberg algorithm to estimate the exponential model from the normal data. Furthermore, individual Prony iterations used less machine time on average than those of the Levenberg algorithm, for which many adjustments of the Levenberg parameter were required. The Levenberg algorithm was limited to 40 iterations, and was regarded as failing if it did not converge before this. Prony obliged by always converging, but did so sometimes to complex roots. These were regarded as failures of Prony for the purposes of the current study. However in all such cases the Prony algorithm found a sum of damped sinusoids which fitted the data more accurately than did any sum of exponentials, and in practice this would often be a valid solution. The

Levenberg algorithm failed whenever Prony did. For both programs failure occurred when the estimates of β_2 and β_3 were relatively close together.

TABLE 2
Mean of $\hat{\sigma}$ over 10 replicates. Given are the leading significant figures, those for Prony above those for Levenberg.

$n\backslash\sigma$	0.030	0.010	0.003	0.001
32	2885	97251	292062	9737559
	2942	98661	293552	9737579
64	2889	96409	289308	9644329
	2914	96879	298484	9644324
128	2945	98177	294538	9817982
	2950	98259	294538	9818041
256	2937	97896	293686	9789516
	2940	97925	293686	9789513
512	2981	99362	298085	9936191
	2983	99376	298085	9936189

Table 2 gives estimated standard deviations averaged over the 10 replications. Reflecting as it does the minimized sums of squares, it gives some idea of comparative precision achieved by the two algorithms. However the sums of squares are not strictly comparable when complex roots occur — in those cases, Prony always achieves a lower sum of squares by including implicitly trigonometric terms in the mean function.

TABLE 3
Means and standard deviations of estimates of β_2 and β_3. True values are 4 and 7 respectively.

$n\backslash\sigma$	0.030	0.010	0.003	0.001
32	4.089(1.4)	4.127(.78)	4.138(.40)	4.065(.18)
	17.08(28.)	7.420(2.0)	6.872(.82)	6.901(.36)
64	3.937(1.0)	4.083(.60)	4.101(.31)	4.030(.11)
	8.629(3.7)	7.169(1.4)	6.876(.63)	6.952(.23)
128	3.930(.66)	4.007(.39)	4.029(.20)	4.005(.06)
	7.680(1.9)	7.132(.85)	6.977(.39)	6.995(.12)
256	4.022(.83)	4.071(.47)	4.024(.18)	4.004(.06)
	7.721(2.1)	7.072(1.0)	6.992(.36)	7.001(.12)
512	4.216(.65)	4.139(.37)	4.043(.13)	4.012(.04)
	6.974(1.7)	6.830(.78)	6.930(.27)	6.979(.09)

Table 3 gives means and standard deviations of the smaller and larger estimated rate constants respectively.

7. Concluding Remarks. In the simulations and in other experiments, the modified Prony algorithm has been found not only to converge rapidly but to be remarkably tolerant of poor starting values. A complete explanation of this behaviour has not yet been made, but the reparametrization from the rate constants to the Prony parameters is undoubtedly an important part. Only average performance was observed from the modified Prony algorithm in fitting rational functions for which no reparametrization is involved [30].

The eigenstructure of B_δ, described in Section 4, raises a potential problem for the convergence criterion of the modified Prony algorithm, but one that seems mitigated in practice. With three exponential terms in the simulations the largest eigenvalue of $n^{-1}B_\delta$ is $O(n^6)$. This suggests that the very rapid convergence of the algorithm for large sample sizes is an artifact of numerical ill-conditioning in B_δ. The algorithm, however, actually returns excellent estimates, even for $n = 512$, and a fully satisfactory explanation of this phenomenon is yet to be made. While the eigenvalues of the difference form Prony matrix B_γ are all of the same order, limited experiments suggest that the recurrence and difference form algorithms are very similar in their practical behaviour.

Appendix. Proofs of the Stability Theorems. Proofs of the stability Theorems 4.1 and 4.2 require that X_γ and $X_{\gamma i}$ be related through matrices that have explicit eigen-factorizations. This is achieved by augmenting X_γ to the $n \times n$ circulant matrix

$$C = \begin{pmatrix} c_1 & & & c_{p+1} & \cdots & c_2 \\ \vdots & \ddots & & & \ddots & \vdots \\ & & c_1 & & & c_{p+1} \\ c_{p+1} & & & c_1 & & \\ & \ddots & \vdots & \vdots & \ddots & \\ & & c_{p+1} & c_p & \cdots & c_1 \end{pmatrix}.$$

Then $X_\gamma = CP^T$, where P is the $(n-p) \times n$ matrix $(I \;\; 0)$ which picks out the leading $n - p$ columns of C. In fact,

$$C^T = \sum_{k=1}^{q+1} \gamma_k \Delta^{k-1} = \sum_{k=1}^{q+1} c_k \Pi^{k-1}$$

where Π is the circulant forward shift matrix $\operatorname{circ}(0, 1, 0, \ldots, 0)$ and Δ is the circulant difference matrix $\Pi = n(\Pi - I)$. Write $C_k = \frac{\partial C}{\partial \gamma_k} = \Delta^{k-1}$ and $D_k = C_k C^{-1}$ for $k = 1, \ldots, q+1$. Note that the D_k are smoothing operators, since C^{-1} is the solution operator for a difference equation. Then we have the key identity

$$X_{\gamma i}^T = PC_i^T = PC^T C^{-T} C_i^T = X_\gamma^T D_i^T,$$

which leads to the following expansion for B_γ:

LEMMA 7.1. *The components of* $\frac{1}{n}B_\gamma$ *can be expanded as*

$$\begin{aligned} \frac{1}{n}B_{\gamma ij} = & \frac{1}{n}(\boldsymbol{\mu}_0 + \boldsymbol{\epsilon})^T D_i (I - P_A) D_j^T (\boldsymbol{\mu}_0 + \boldsymbol{\epsilon}) \\ & - \frac{1}{n}(\boldsymbol{\mu}_0 + \boldsymbol{\epsilon})^T (I - P_A) D_i^T D_j (I - P_A)(\boldsymbol{\mu}_0 + \boldsymbol{\epsilon}) \end{aligned}$$

where $\boldsymbol{\mu}_0 + \boldsymbol{\epsilon} = \mathbf{y}$. The importance of this lemma lies in its representation for B_γ in terms of projections and smoothing operators.

The remainder of the proof of Theorem 4.1 consists of using the law of large numbers to show that the terms involving $\boldsymbol{\epsilon}$ in the above expansion are asymptotically negligible. A similar application of the law of large numbers was given in [30]. For example, the components of $\frac{1}{n}A^T\boldsymbol{\epsilon} \xrightarrow{a.s.} 0$ because each column $\mathbf{a}_j$ of A is smooth in the sense that it can be defined as the values taken by a continuous function, namely

$e^{-\beta_j t}$, at the time points $t = t_1, \ldots, t_n$. The convergence moreover is uniform in $\boldsymbol{\beta}$ because $e^{-\beta_j t}$ is jointly continuous in β_j and t. The lemmas below show that terms like $\frac{1}{n}A^T D_u^T \epsilon$ and $\frac{1}{n}A^T D_u^T D_v \epsilon$ tend to zero also because $D_u \mathbf{a}_j$ and $D_v^T D_u \mathbf{a}_j$ are smooth in the same sense as above.

The lemmas are proved directly by construction using the properties of circulant matrices. A circulant matrix of the form of C^T has complex eigenvalues

$$\lambda_i = p_c(\omega^{i-1}) = \sum_{k=1}^{q+1} c_k \omega^{(i-1)(k-1)}$$

where $\omega = \exp\{\frac{2\pi}{n}\sqrt{-1}\}$ is the nth fundamental root of unity. Also

$$C^T = \Omega^* \Lambda \Omega$$

where Ω is the $n \times n$ Fourier matrix defined by $\Omega^*_{ij} = \omega^{(i-1)(j-1)}$, which is both unitary and circulant, and where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. See [3] or [8]. For any vector $\mathbf{z}$, $\Omega\mathbf{z}$ is the discrete Fourier transform and $\Omega^*\mathbf{z}$ is the inverse discrete Fourier transform. Also the polynomial $p_c(\cdot)$ is the transfer function of C^T. Since D_u and D_v are also circulant, the vectors $D_u \mathbf{a}_j$ and the $D_v^T D_u \mathbf{a}_j$ can be constructed by explicitly evaluating the discrete Fourier transform of $\mathbf{a}_j$, multiplying by the appropriate transfer function, and inverting back to the time domain.

LEMMA 7.2. *The sequence*

$$f_k = \rho^{k-1} \qquad k = 1, \ldots, n\,,$$

where ρ is any constant, has discrete Fourier transform

$$F_k = n^{-1/2}(1-\rho^n)(\omega^{k-1}-\rho)^{-1}\omega^{k-1} \qquad k = 1, \ldots, n$$

where ω is the fundamental n*th root of unity.*

Proof. Follows from summing a geometric series in $\omega^{-(k-1)}\rho$, and using $\omega^n = 1$. □

LEMMA 7.3. *The sequence*

$$F_k = n^{-1/2}(\omega^{k-1}-\rho)^{-2}\omega^{2(k-1)} \qquad k = 1, \ldots, n\,,$$

where ρ is any constant, has inverse discrete Fourier transform

$$f_k = (1-\rho^n)^{-2} k \rho^{k-1} \qquad k = 1, \ldots, n\,.$$

Proof. Uses geometric series identities and

$$\sum_{j=0}^{n-1} \omega^{mj} = 0$$

for positive integers m. □

The next two lemmas follow from the partial fraction theorem.

LEMMA 7.4. *If $p(z)$ is a polynomial of degree less than r, then*

$$F(z) = \frac{p(z)}{(z-a_1)\cdots(z-a_r)} = \sum_{j=1}^{r} \frac{b_j}{z-a_j}$$

with

$$b_j = Q_j^{-1} p(a_j)\,, \quad Q_j = \prod_{\substack{k=1\\k\neq j}}^{r} (a_j - a_k)\,.$$

LEMMA 7.5. *If $p(z)$ is a polynomial of degree at most r, then*

$$F(z) = \frac{p(z)}{(z-a_1)^2(z-a_2)\cdots(z-a_r)} = \frac{b_1 z}{(z-a_1)^2} + \sum_{j=2}^{r} \frac{b_j}{z-a_j} - \sum_{j=1}^{r} \frac{b_j}{z-a_1}$$

with

$$b_1 = \frac{p(a_1)}{a_1 Q_1}\,, \quad b_j = \frac{p(a_j)}{(a_j - a_1) Q_j}\,, \quad Q_j = \prod_{\substack{k=1\\k\neq j}}^{r} (a_j - a_k)\,.$$

LEMMA 7.6. *For each u, there exist functions f_j, continuous on $[0,1]$, such that*

$$(D_u A)_{ij} = f_j(t_i) + O\left(\frac{1}{n}\right)$$

uniformly for $i = 1, \ldots, n$ and $j = 1, \ldots, p$.

Proof. The operator D_u can be written

$$D_u = \Delta^{T(u-1)} \prod_{j=1}^{p} (\Delta^T + \zeta_j I)^{-1}$$

which has transfer function

$$\lambda(z) = \frac{n^{u-1}}{n^p} \frac{(z^{-1}-1)^{u-1}}{\prod_{j=1}^{p}(z^{-1}-\rho_j)} \qquad z = \omega^0, \ldots, \omega^{n-1}\,.$$

Here $\rho_j = e^{-\beta_j/n}$ and $\zeta_j = n(1-\rho_j)$. Using Lemma 7.2, $D_u \mathbf{a}_1$ has discrete Fourier transform

$$F(z) = \lambda(z) n^{-1/2} \rho_1 z \frac{1-\rho_1^n}{z-\rho_1} \qquad z = \omega^0, \ldots, \omega^{n-1}$$

which, using Lemma 7.4, can be written as

$$n^{-1/2} \frac{n^{u-1}}{n^p} \rho_1 (1-\rho_1^n) \left\{ \sum_{j=1}^{p} \frac{b_j}{z^{-1}-\rho_j} + \frac{c}{1-z^{-1}\rho_1} \right\}$$

with

$$b_j = \frac{(\rho_j - 1)^{u-1}}{(1-\rho_j\rho_1)\prod_{\substack{k=1\\k\neq j}}^{p}(\rho_j - \rho_k)}$$

and

$$c = \frac{(\rho_1^{-1}-1)^{u-1}}{\prod_{j=1}^{p}(\rho_1^{-1}-\rho_j)}\,.$$

Reversing Lemma 7.2, we obtain $(D_u\mathbf{a}_1)_s$ as

$$\begin{aligned}
&\frac{n^{u-1}}{n^p}\rho_1(1-\rho_1^n)\left\{\sum_{j=1}^{p}\frac{b_j(-\rho_j^{-1})}{1-\rho_j^{-n}}\rho_j^{-(s-1)}+\frac{c}{1-\rho_1^n}\rho_1^{s-1}\right\}\\
&=\frac{n^{u-1}}{n^p}\left\{-\sum_{j=1}^{p}\frac{b_j\rho_1(1-\rho_1^n)}{1-\rho_j^{-n}}\rho_j^{-s}+c\rho_1^s\right\}.
\end{aligned}$$

Using the fact that $\zeta_j\to\beta_j$, we find that

$$\frac{n^{u-1}}{n^p}b_j=b_{j\infty}+o\left(\frac{1}{n}\right),\quad \frac{n^{u-1}}{n^p}c=c_\infty+o\left(\frac{1}{n}\right)$$

where

$$b_{j\infty}=\frac{(-\beta_j)^{u-1}}{(\beta_1+\beta_j)\prod_{\substack{k=1\\k\neq j}}^{p}(\beta_k-\beta_j)}$$

and

$$c_\infty=\frac{(-\beta_1)^{u-1}}{\prod_{k=1}^{p}(\beta_1+\beta_k)}$$

do not depend on n. So let

$$f_1(t)=c_\infty e^{-\beta t}-\sum_{j=1}^{p}b_{j\infty}\frac{1-e^{-\beta_1}}{1-e^{\beta_j}}e^{\beta_j t}.$$

The other functions $f_2,\ldots,f_p$ are defined similarly. □

LEMMA 7.7. *For each u and v, there exist functions g_j, continuous on $[0,1]$, such that*

$$(D_v^TD_uA)_{ij}=g_j(t_i)+O\left(\frac{1}{n}\right)$$

uniformly for $i=1,\ldots,n$ and $j=1,\ldots,p$.

Proof. The operator

$$D_v^TD_u=\Delta^{v-1}\Delta^{u-1^T}\prod_{j=1}^{p}(\Delta+\zeta_jI)^{-1}(\Delta^T+\zeta_jI)^{-1}$$

has transfer function

$$\lambda(z)=\frac{n^{u+v-2}}{n^{2p}}\frac{(z^{-1}-1)^{u-1}(z-1)^{v-1}}{\prod_{j=1}^{p}(z^{-1}-\rho_j)(z-\rho_j)}\quad z=\omega^0,\ldots,\omega^{n-1}.$$

Therefore $D_v^TD_u\mathbf{a}_1$ has discrete Fourier transform

$$\begin{aligned}
F(z)&=\lambda(z)n^{-1/2}\rho_1(1-\rho_1^n)\frac{z}{z-\rho_1}\\
&=n^{-1/2}\frac{n^{u+v-2}}{n^{2p}}\rho_1(1-\rho_1^n)\frac{zz^{p-u+1}(1-z)^{u-1}(z-1)^{v-1}}{(z-\rho_1)^2\prod_{j=2}^{p}(z-\rho_j)\prod_{j=1}^{p}(1-z\rho_j)},
\end{aligned}$$

which can be written, using Lemma 7.5, as

$$n^{-1/2}\frac{n^{u+v-2}}{n^{2p}}\rho_1(1-\rho_1^n)z\left\{\frac{b_1 z}{(z-\rho_1)^2}+\sum_{j=2}^{p}\frac{b_j}{(z-\rho_j)}+\sum_{j=1}^{p}\frac{c_j}{(1-z\rho_j)}-\frac{\sum_{j=1}^{p}(b_j+c_j)}{(z-\rho_1)}\right\}$$

with

$$b_1 = \frac{\rho_1^{(p-u+1)}(1-\rho_1)^{u-1}(\rho_1-1)^{v-1}}{\rho_1\prod_{k=2}^{p}(\rho_1-\rho_k)\prod_{k=1}^{p}(1-\rho_1\rho_k)}$$
$$b_j = \frac{\rho_j^{(p-u+1)}(1-\rho_j)^{u-1}(\rho_j-1)^{v-1}}{(\rho_j-\rho_1)^2\prod_{\substack{k=2\\k\neq j}}^{p}(\rho_j-\rho_k)\prod_{k=1}^{p}(1-\rho_j\rho_k)}$$

and

$$c_j = \frac{\rho_j^{-(p-u+1)}(1-\rho_j^{-1})^{u-1}(\rho_j^{-1}-1)^{v-1}}{(\rho_j^{-1}-\rho_1)^2\prod_{k=2}^{p}(\rho_1^{-1}-\rho_k)\prod_{\substack{k=1\\k\neq j}}^{p}(1-\rho_j^{-1}\rho_k)}\ .$$

Using Lemmas 7.2 and 7.3 to invert $F(z)$, we obtain $(D_v^T D_u \mathbf{a}_1)_s$ as

$$\frac{n^{u+v-2}}{n^{2p}}\rho_1(1-\rho_1^n)\left\{\frac{b_1}{(1-\rho_1^n)^2}s\rho_1^{s-1}+\sum_{j=2}^{p}\frac{b_j}{1-\rho_j^n}\rho_j^{s-1}\right.$$
$$\left.-\sum_{j=1}^{p}\frac{c_j\rho_j^{-1}}{1-\rho_j^{-n}}\rho_j^{-(s-1)}-\sum_{j=1}^{p}\frac{b_j+c_j}{1-\rho_1^n}\rho_1^{s-1}\right\}\ .$$

Using $\zeta_j \to \beta_j$ we find

$$\frac{n^{u+v-2}}{n^{2p}}nb_1 = b_{1\infty}+o\left(\frac{1}{n}\right)\ ,\quad \frac{n^{u+v-2}}{n^{2p}}b_j = b_{j\infty}+o\left(\frac{1}{n}\right)$$
$$\frac{n^{u+v-2}}{n^{2p}}c_j = c_{j\infty}+o\left(\frac{1}{n}\right)$$

with

$$b_{1\infty} = \frac{(-1)^{v-1}\beta_1^{u+v-2}}{2\beta_1\prod_{k=2}^{p}(\beta_k^2-\beta_1^2)}$$
$$b_{j\infty} = \frac{(-1)^{v-1}\beta_j^{u+v-2}}{2\beta_j(\beta_1-\beta_j)\prod_{\substack{k=1\\k\neq j}}^{p}(\beta_k^2-\beta_j^2)}$$
$$c_{j\infty} = \frac{(-1)^{u-1}\beta_j^{u+v-2}}{2\beta_j(\beta_1+\beta_j)\prod_{\substack{k=1\\k\neq j}}^{p}(\beta_k^2-\beta_j^2)}$$

which do not depend on n. So let

$$g_1(t) = \frac{c_{1\infty}}{1-e^{-\beta_1}}te^{-\beta_1 t}+\sum_{j=2}^{p}b_{j\infty}\frac{1-e^{-\beta_1}}{1-e^{-\beta_j}}e^{-\beta_j t}$$
$$-\sum_{j=1}^{p}c_{j\infty}\frac{1-e^{-\beta_1}}{1-e^{\beta_j}}e^{\beta_j t}-\sum_{j=1}^{p}(b_{j\infty}+c_{j\infty})e^{-\beta_j t}\ .$$

The other functions $g_2, \ldots, g_p$ can be defined in a similar fashion. □

The final lemma differs from Lemma 7.7 in that D_u and D_v are evaluated at the current value of γ while A is evaluated at the true value γ_0. Its proof is similar to that of Lemma 7.7, with the difference that all the poles of the discrete Fourier transform $F(z)$ are simple, and each function $g_{0j}(t)$ includes a term in $e^{-\beta_0 t}$ as well as in $e^{-\beta_k t}$ and $e^{\beta_k t}$, $k = 1, \ldots, p$.

LEMMA 7.8. *Let $A_0 = A(\gamma_0)$. For each u and v, there exists functions g_{0j}, continuous on $[0,1]$, such that*

$$(D_v^T D_u A_0)_{ij} = g_{0j}(t_i) + O\left(\frac{1}{n}\right)$$

uniformly for $i = 1, \ldots, n$ and $j = 1, \ldots, p$.

Proof of Theorem 4.1. Consider the expansion for $\frac{1}{n}B_\gamma$ given in Lemma 7.1. The terms

$$\frac{1}{n}\epsilon^T D_i D_j^T \epsilon - \frac{1}{n}\epsilon^T D_i^T D_j \epsilon$$

cancel out of this expansion — D_i and D_j commute since they are circulants. Repeated application of Lemmas 7.6 to 7.8 and the law of large numbers [30, Theorem 4] shows that all other terms which involve ϵ converge to zero. The first term for example is

$$\frac{1}{n}\boldsymbol{\mu}_0^T D_i P_A D_j^T \epsilon = (\frac{1}{n}\boldsymbol{\mu}_0^T D_i A)(\frac{1}{n}A^T A)^{-1})(\frac{1}{n}A^T D_j^T \epsilon). \tag{14}$$

The middle term $\frac{1}{n}A^T A$ converges to the positive definite matrix with elements

$$\int_0^1 e^{-(\beta_i+\beta_j)t} dt$$

for $i, j = 1, \ldots, p$, and each element of $\frac{1}{n}A^T D_j^T \epsilon$ converges to zero almost surely, by Lemma 7.6 and the law of large numbers. Lemma 7.6 also shows that each element of $\frac{1}{n}\boldsymbol{\mu}_0^T D_i A$ converges to a constant, hence the whole term (14) converges to zero. Moreover the convergence is uniform for γ in a compact set. Similarly, the term

$$\epsilon^T P_A D_i^T D_j P_A \epsilon = (\frac{1}{n}\epsilon^T A)(\frac{1}{n}A^T A)^{-1}(\frac{1}{n}A^T D_i^T D_j A)(\frac{1}{n}A^T A)^{-1}(\frac{1}{n}A^T \epsilon)$$

converges to zero. Lemma 7.6 shows that $\frac{1}{n}A^T D_i^T D_j A$ converges to a constant $p \times p$ matrix, and the law of large numbers shows that $\frac{1}{n}A^T \epsilon$ converges to zero. This term is in fact of smaller order than the first, since it includes two factors which converge to zero.

The other terms involving ϵ are treated in the same way, and require Lemmas 7.7 and 7.8. The remaining terms can be identified with $\mathcal{J}_0$, thus completing the proof. □

Proof of Theorem 4.2. Section 7 of [30] gives an expression for $\dot{B}_\gamma \gamma$ and shows that $E(\dot{B}_\gamma(\gamma_0)\gamma_0) = 0$. The methods used above for Theorem 4.1 can be used to prove that

$$\frac{1}{n}\dot{B}_\gamma(\hat{\gamma})\hat{\gamma} \overset{a.s.}{\to} 0$$

as $n \to \infty$. Now $\hat{\gamma}^T \dot{B}_\gamma(\hat{\gamma})\hat{\gamma} = 0$ so

$$G(\hat{\gamma}) = \left[\frac{1}{n}B_\gamma(\hat{\gamma}) + \hat{\gamma}\hat{\gamma}^T\right]^{-1} \frac{1}{n}\dot{B}_\gamma(\hat{\gamma})\hat{\gamma}.$$

Theorem 4.1 shows that $\frac{1}{n}B_\gamma(\hat{\gamma})+\hat{\gamma}\hat{\gamma}^T \overset{a.s.}{\to} \mathcal{J}_0+\gamma_0\gamma_0^T$ which is positive definite, which completes the theorem. □

REFERENCES

[1] R. H. BARHAM AND W. DRANE, *An algorithm for least estimation of nonlinear parameters when some of the parameters are linear*, Technometrics, 14 (1972), pp. 757–766.

[2] D. M. BATES AND D. WATTS, *Nonlinear Regression Analysis and its Applications*, Wiley, New York, 1988.

[3] R. BELLMAN, *Introduction to Matrix Analysis*, 2nd ed., McGraw-Hill, New York, 1970.

[4] M. BENSON, *Parameter fitting in dynamic models*, Ecol. Mod., 6 (1979), pp. 97–115.

[5] Y. BRESLER AND A. MACOVSKI, *Exact maximum likelihood parameter estimation of superimposed exponential signals in noise*, IEEE Trans. Acoust. Speech Signal Process., 34 (1986), pp. 1081–1089.

[6] D. G. CANTOR AND J. W. EVANS, *On approximation by positive sums of powers*, SIAM J. Appl. Math., 18 (1970), pp. 380–388.

[7] J. C. R. CLAEYSSEN AND L. A. DOS SANTOS LEAL, *Diagonalization and spectral decomposition by factor block circulant matrices*, Linear Algebra Appl., 99 (1988), pp. 41–61.

[8] P. J. DAVIS, *Circulant Matrices*, Wiley, New York, 1979.

[9] H. DERIN, *Estimating components of univariate Gaussian mixtures using Prony's method*, IEEE Trans. Pattern Analysis Machine Intelligence, 9 (1987), pp. 142–148.

[10] J. W. EVANS, W. B. GRAGG AND R. J. LEVEQUE, *On least squares exponential sum approximation with positive coefficients*, Math. Comp., 34 (1980), pp. 203–211.

[11] A. E. GILMOUR, *Circulant matrix methods for the numerical solution of partial differential equations by* FFT *convolutions*, Appl. Math. Modelling, 12 (1988), pp. 44–50.

[12] G. H. GOLUB AND V. PEREYRA, *The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate*, SIAM J. Numer. Anal., 10 (1973), pp. 413–32.

[13] R. I. JENNRICH, *Asymptotic properties of non-linear least squares estimators*, Ann. Math. Statist., 40 (1969), pp. 633–643.

[14] M. KAHN, M. S. MACKISACK, M. R. OSBORNE AND G. K. SMYTH, *On the consistency of Prony's method and related algorithms*, J. Comput. Graph. Statist., 1 (1992), pp. 329–349.

[15] D. W. KAMMLER, *Least squares approximation of completely monotonic functions by sums of exponentials*, SIAM J. Num. Anal., 16 (1979), pp. 801–818.

[16] L. KAUFMANN, *A variable projection method for solving separable nonlinear least squares problems*, BIT, 15 (1975), pp. 49–57.

[17] S. M. KAY, *Modern Spectral Estimation: Theory and Application*, Prentice-Hall, Englewood Cliffs, 1988.

[18] D. KUNDU, *Estimating the parameters of undamped exponential signals*, Technometrics, 35 (1993), pp. 215–218.

[19] C. LANZCOS, *Applied Analysis*, Prentice-Hall, Englewood Cliffs, 1956.

[20] W. H. LAWTON AND E. A. SYLVESTRE, *Elimination of linear parameters in nonlinear regression*, Technometrics, 13 (1971), pp. 461–467.

[21] M. S. MACKISACK, M. R. OSBORNE AND G. K. SMYTH, *A modified Prony algorithm for estimating sinusoidal frequencies*, J. Statist. Comp. Simul., to appear.

[22] S. L. MARPLE, *Digital Spectral Analysis*, Prentice-Hall, Englewood Cliffs, 1987.

[23] H. D. MITTLEMAN AND H. WEBER, *Multi-grid solution of bifurcation problems*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 49–60.

[24] R. J. MULHOLLAND, J. R. CRUZ AND J. HILL, *State-variable canonical forms for Prony's method*, Int. J. Systems Sci., 17 (1986), pp. 55–64.

[25] NUMERICAL ALGORITHMS GROUP, NAG FORTRAN *Library Manual Mark* 10, Numerical Algorithms Group, Oxford, 1983.

[26] R. D. NUSSBAUM, *Circulant matrices and differential-delay equations*, J. Differential Equations, 60 (1985), pp. 201–217.

[27] M. R. OSBORNE, *A class of nonlinear regression problems*, in Data Representation, R. S. Anderssen and M. R. Osborne, eds., University of Queensland Press, St Lucia, 1970, pp. 94–101.

[28] ———, *Some special nonlinear least squares problems*, SIAM J. Numer. Anal., 12 (1975), pp. 571–592.

[29] ———, *Nonlinear least squares — the Levenberg algorithm revisited*, J. Austral. Math. Soc. B, 19 (1976), pp. 343–357.

[30] M. R. Osborne and G. K. Smyth, *A modified Prony algorithm for fitting functions defined by difference equations*, SIAM J. Sci. Stat. Comput., 12 (1991), pp. 362–382.

[31] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[32] H. Ouibrahim, *Prony, Pisarenko, and the matrix pencil: a unified presentation*, IEEE Trans. Acoust. Speech Signal Process., 37 (1989), pp. 133–134.

[33] E. Pelikan, *Spectral analysis of* ARMA *processes by Prony's method*, Kybernetika, 20 (1984), pp. 322–328.

[34] G. J. S. Ross, *Nonlinear Estimation*, Springer-Verlag, New York, 1990.

[35] A. Ruhe, *Fitting empirical data by positive sums of exponentials*, SIAM J. Sci. Statist. Comp., 1 (1980), pp. 481–498.

[36] H. Sakai, *Statistical analysis of Pisarenko's method for sinusoidal frequency estimation*, IEEE Trans. Acoust. Speech Signal Process., 32 (1984), pp. 95–101.

[37] G. A. F. Seber and C. J. Wild, *Nonlinear Regression*, Wiley, New York, 1989.

[38] M. R. Smith, S. Cohn-Sfetcu and H. A. Buckmaster, *Decomposition of multicomponent exponential decays by spectral analytic techniques*, Technometrics, 18 (1976), pp. 467–482.

[39] G. K. Smyth, *Coupled and separable iterations in nonlinear estimation*, Ph.D. thesis, Department of Statistics, The Australian National University, Canberra, 1985.

[40] ———, *Partitioned algorithms for maximum likelihood and other nonlinear estimation*, Centre for Statistics Research Report 3, University of Queensland, St Lucia, 1993.

[41] P. Stoica and A. Nehorai, *Study of the statistical performance of the Pisarenko harmonic decomposition method*, Comm. Radar Signal Process., 153 (1988), pp. 161–168.

[42] J. M. Varah, *A spline least squares method for numerical parameter estimation in differential equations*, SIAM J. Sci. Stat. Comput., 3 (1982), pp. 28–46.

[43] ———, *On fitting exponentials by nonlinear least squares*, SIAM J. Sci. Stat. Comput., 6 (1985), pp. 30–44.

[44] D. Walling, *Non-linear least squares curve fitting when some parameters are linear*, Texas J. Science, 20 (1968), pp. 119–24.

[45] A. C. Wilde, *Differential equations involving circulant matrices*, Rocky Mountain J. Math., 13 (1983), pp. 1–13.

[46] W. J. Wiscombe and J. W. Evans, *Exponential-sum fitting of radiative transmission functions*, J. Comp. Physics, 24 (1977), pp. 416–444.