Open access • Journal Article • DOI:10.1137/0912020

# A Modified Prony Algorithm for Fitting Functions Defined by Difference Equations



Michael R. Osborne, Gordon K. Smyth

**Published on:** 01 Mar 1991 - Siam Journal on Scientific and Statistical Computing (Society for Industrial and Applied Mathematics)

**Topics:** Hessian equation, Prony's method, Nonlinear eigenproblem, Hessian matrix and Explained sum of squares

# A Modified Prony Algorithm for Fitting Functions Defined by Difference Equations

M. R. Osborne *  G. K. Smyth †

This technical report dated 29 June 1987 was published in

SIAM J. on Sci. and Statist. Comp., Vol 12 (1991), pages 362–382.

**Abstract**

This paper reformulates, generalizes and investigates the stability of the modified Prony algorithm introduced by Osborne (1975), with special reference to rational and exponential fitting. The algorithm, originally for exponential functions, is generalized to the least squares fitting of any function which satisfies a linear homogeneous difference equation. Using the difference equation formulation, the problem is expressed as a separable regression, and hence as a nonlinear eigenproblem in terms of the coefficients of the difference equation. The eigenproblem involves finding the null space of a matrix of data differences $B$, and is solved using a variant of inverse iteration. Stability of the algorithm is shown to depend on the fact that $B$ closely approximates the Hessian of the sum of squares. The expectations of $B$ and the Hessian are evaluated. In the case of rational fitting, the relative difference between $B$ and the Hessian is shown to converge to zero almost surely. Some details of the implementation of the algorithm are given. A simulation study compares the modified Prony algorithm with the Levenberg algorithm on a rational fitting problem, and supports the theoretical results.

KEYWORDS: Prony's method; differential equations; nonlinear least squares; rational fitting; nonlinear eigenproblem; inverse iteration; asymptotic stability; Gauss-Newton algorithm.

AMS(MOS) subject classifications: 62J02 65D10

Short title: A Modified Prony Algorithm

*Statistics Section, School of Mathematics, Australian National University, Canberra 2601, Australia.

†Department of Mathematics, University of Queensland, St Lucia 4067, Australia.

# 1 Introduction

This paper reformulates, generalizes and investigates the stability of the modified Prony algorithm introduced by Osborne (1975). The algorithm, originally for exponential functions, is generalized to provide least squares fitting of any function which satisfies a linear homogeneous difference equation (e.g. to fit compartment models). The main interest is in functions for which the difference equation is an exact discrete analogue of the limiting differential equation, and especially in the specific examples of rational and exponential function fitting.

In a classic paper, Prony (1795) proposed a method for interpolating a sum of exponential functions through a series of data values at equally spaced points. He solved a system of linear equations for the coefficients of a difference equation satisfied by the exponential functions, and obtained the exponential functions from the roots of the polynomial with those coefficients. Prony's method was extended to the least squares context by, for example, Householder (1950), and is actively used in systems engineering and time series (for example, Marple (1987), Mulholland et al (1986)). But a formulation which returned the actual least squares estimators was not produced until Osborne (1975), who proposed a modified Prony algorithm in which a nonlinear eigenproblem is solved for the coefficients of the difference equation. Osborne (1975) showed, by way of a simulation study, that the algorithm converged with surprising speed, but didn't provide a theoretical justification. He also indicated an extension to rational function fitting, without verifying the numerical properties.

This paper is directed towards proving that the algorithm is asymptotically stable, ie that it converges at a rapid rate from good starting values for sufficiently large data sets. (This is analogous to the result proved by Jennrich (1969) for the Gauss-Newton algorithm.) Although the general form of the proof is shown, the proof is completed here only for rational functions. The rational fitting case involves a minimum of special complication, and least obscures the general case. The corresponding result for exponential functions was proved by Smyth (1985), and will be given in a separate paper to be devoted to the special features of the exponential fitting case.

The problem considered is that of fitting a given continuous function $\mu$ to the data pairs $(t_i, y_i), i = 1, \ldots, n$ by least squares. Let $\mu_i = \mu(t_i)$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$. The errors $y_i - \mu_i$ are assumed to be independent and to have equal and finite variances. We show that if the $\mu_i$ satisfy a difference equation, linear and homogeneous in a parameter vector $\boldsymbol{\gamma}$, then the least squares problem may be formulated as a nonlinear eigenproblem, which can be solved using a variant of inverse iteration. Section 2 shows that the

difference equation can be expressed in matrix terms as

$$X(\boldsymbol{\gamma})^T \boldsymbol{\mu} = 0$$

where $X$ is a banded linear matrix function of $\boldsymbol{\gamma}$. Section 3 shows that the least squares problem, concentrated in terms of the nonlinear parameters, may be solved by minimizing

$$\psi(\boldsymbol{\gamma}) = \mathbf{y}^T P_X \mathbf{y}$$

where $P_X$ is the projection onto the space $\mathcal{R}(X)$ spanned by the columns of $X$. Furthermore the gradient of $\psi$ is given by

$$\dot{\psi}(\boldsymbol{\gamma}) = 2B(\boldsymbol{\gamma})\boldsymbol{\gamma}$$

where $B$ is a symmetric matrix function of $\boldsymbol{\gamma}$. Section 4 relates the results to rational and exponential fitting. Section 5 describes in detail the modified Prony algorithm, and briefly discusses the effect of the constraint used to standardize $\boldsymbol{\gamma}$. Section 6 shows that a sufficient condition for convergence is that $B$ should closely approximate $\ddot{\psi}/2$, and Section 7 evaluates the expectations of these quantities. Sections 8 to 10 discuss the asymptotic eigenstructure of $X$ and $B$, and show that for rational fitting $B^+(\ddot{\psi}/2 - B) \to 0$ with probability one, thus establishing the stability of the algorithm.

Rational fitting is usually accomplished using the Gauss-Newton algorithm or its Levenberg or Marquardt modifications (Ratkowsky, 1985; Bates and Watts, 1988), so Section 11 compares the modified Prony algorithm with the Levenberg algorithm on a rational fitting problem using a small simulation study.

# 2 The Difference Equation

We assume that $\mu(t)$ satisfies a difference equation of the form

$$\sum_{k=1}^{p+1} d_k(t;\boldsymbol{\gamma}) \Delta^{k-1} \mu(t) = 0 \tag{1}$$

and that the coefficients $d_k$ are continuous in $t$ and linear and homogeneous in some unknown $\boldsymbol{\gamma} \in \mathbb{R}^{q+1}$. We call $\boldsymbol{\gamma}$ the *Prony parameters*. For simplicity, we will assume that the times $t_i$ at which $\mu$ is sampled are equally spaced over a fixed interval, although the results of this paper require only that the empirical distribution of the $t_i$ converges to a non-trivial probability distribution. Without loss of generality we take that interval to be $[0, 1]$ so that $t_i = 1/n$, and define $\Delta$ to be the forward difference operator

$$\Delta f(t) = n\left[f(t + 1/n) - f(t)\right].$$

Generally $\boldsymbol{\gamma}$ and $\mathbf{d} = (d_1, \ldots, d_{p+1})^T$ will depend on $n$, but we will usually not make this explicit in our notation. Since the scales of $\boldsymbol{\gamma}$ and $\mathbf{d}$ are disposable, we will adjoin the condition $\gamma_{q+1} = 1$ or $\|\boldsymbol{\gamma}\| = 1$ with respect to some inner product, whatever is most convenient, and similarly for $\mathbf{d}$.

By assumption, the $d_k$ may be expanded into

$$d_k(t; \boldsymbol{\gamma}) = \sum_{j=1}^{q+1} \gamma_j d_{kj}(t)$$

where the $d_{kj}$ are known continuous functions, so (1) may be rewritten as

$$\sum_{j=1}^{q+1} \gamma_j \left( \sum_{k=1}^{p+1} d_{kj}(t) \Delta^{k-1} \right) \mu(t) = 0.$$

This may be translated into matrix terms in the following way. Let $\boldsymbol{\Pi}$ be the $n \times n$ circulant forward shift operator

$$\boldsymbol{\Pi} = \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ & & \vdots & & \\ 1 & 0 & 0 & \ldots & 0 \end{pmatrix}$$

and let $\boldsymbol{\Delta}$ be the circulant difference matrix $\boldsymbol{\Delta} = n(\boldsymbol{\Pi} - I)$. Then $\boldsymbol{\mu}$ satisfies

$$P \sum_{j=1}^{q+1} \gamma_j \sum_{k=1}^{p+1} \langle d_{kj}(\mathbf{t}) \rangle \, \boldsymbol{\Delta}^{k-1} \boldsymbol{\mu} = 0$$

where $\langle d_{kj}(\mathbf{t}) \rangle$ denotes the diagonal matrix with diagonal elements $d_{kj}(t_1)$ to $d_{kj}(t_n)$, and $P = (I_{n-p}\ 0)$ selects the leading $(n-p)$ rows. (Throughout this paper, scalar functions are taken to act componentwise on vectors, and $\langle \cdot \rangle$ denotes a diagonal matrix with the components of the vector argument down the diagonal.) Write also

$$C_j^T = \sum_{k=1}^{p+1} \langle d_{kj}(\mathbf{t}) \rangle \, \boldsymbol{\Delta}^{k-1}$$

and

$$C = \sum_{j=1}^{q+1} C_j \gamma_j.$$

Let $X_j = C_j P^T$, the leading $n - p$ columns of $C_j$, and let

$$X = \sum_{j=1}^{q+1} \gamma_j X_j.$$

Then (1) may be re-expressed in matrix terms as

$$PC(\boldsymbol{\gamma})^T\boldsymbol{\mu} = 0$$

or

$$X(\boldsymbol{\gamma})^T\boldsymbol{\mu} = 0. \tag{2}$$

Note that $X$ is a $n \times (n-p)$ matrix with $p+1$ bands, and that it has full column rank if $d_{p+1} \neq 0$. Although the above definition in terms of circulant operators is often convenient, the last $p$ columns of $C$ are arbitrary as far as (2) is concerned; in the next section we will assume that they have been chosen to ensure that $C$ is invertible.

The above derivation depended on $\mu(t)$ satisfying exactly a difference equation. In practice, it will often be the case that $\mu(t)$ came as the general solution of a differential equation, say of

$$\sum_{k=1}^{p+1} b_k(t;\boldsymbol{\xi})\mathcal{D}^{k-1}\mu(t) = 0 \tag{3}$$

where $\mathcal{D}$ is the differential operator and the coefficients $b_k$ are linear and homogeneous in $\boldsymbol{\xi}$. Until Section 8 of this paper we will assume only the difference equation (1). In the remaining sections, which deal with asymptotics, we will assume that $\mu$ satisfies the differential equation (3) also, and that the difference equation converges to it, in the sense that $\boldsymbol{\gamma} \to \boldsymbol{\xi}$ and $d \to b$ as $n \to \infty$.

The most general expression for $\boldsymbol{\mu}$ which allows a characterization of the form (2) is in fact

$$\boldsymbol{\mu} = Z(\boldsymbol{\gamma})^{-T}W\boldsymbol{\alpha} \tag{4}$$

where $Z$ is a linear function of $\boldsymbol{\gamma}$ and $W$ is a constant $n \times p$ matrix. Then

$$X(\boldsymbol{\gamma}) = Z(\boldsymbol{\gamma})H$$

where $H$ is any full rank $n \times (n-p)$ matrix orthogonal to $W$. The matrix $H$ may be chosen to have $p+1$ bands,

$$H = \begin{pmatrix} h_{11} & & & \\ \vdots & \ddots & & \\ & & & h_{n-p,1} \\ h_{1,p+1} & & & \\ & \ddots & & \vdots \\ & & & h_{n-p,p+1} \end{pmatrix},$$

by determining each $(h_{j,1}, \ldots, h_{j,p+1})^T$ to be orthogonal to

$$\begin{pmatrix} w_{j,1} & \ldots & w_{j,p} \\ \vdots & & \vdots \\ w_{j+p,1} & \ldots & w_{j+p,1} \end{pmatrix},$$

which comprises rows $j$ to $j+p$ of $W$.

# 3 A Nonlinear Eigenproblem

We now invert the matrix difference equation (2) to obtain an explicit expression for $\boldsymbol{\mu}$ of the form (4). One way is to note that quite generally

$$C(\boldsymbol{\gamma})^T\boldsymbol{\mu} = \begin{pmatrix} 0 \\ \boldsymbol{\alpha} \end{pmatrix}$$

for some $\boldsymbol{\alpha} \in \mathbb{R}^p$. Therefore, assuming $C$ is invertible,

$$\begin{aligned} \boldsymbol{\mu} &= C(\boldsymbol{\gamma})^{-T}Q^T\boldsymbol{\alpha} \\ &= A(\boldsymbol{\gamma})\boldsymbol{\alpha} \end{aligned} \tag{5}$$

where $A = C^{-T}Q^T$ and $Q$ is the $p \times n$ matrix $(0\ I_p)$. The $p$ columns of $A$ correspond to a particular set of fundamental solutions of the difference equation (1), depending on how the last $p$ columns of $C$ have been chosen. In any case, (5) displays $\boldsymbol{\mu}$ as a separable regression, with $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ as the linear and nonlinear parameters respectively. For any fixed value of $\boldsymbol{\gamma}$ the sums of squares

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

is minimized by

$$\hat{\boldsymbol{\alpha}}(\boldsymbol{\gamma}) = (A^TA)^{-1}A^T\mathbf{y}. \tag{6}$$

Substituting this back into $\phi$ gives the reduced sum of squares

$$\begin{aligned} \psi(\boldsymbol{\gamma}) &= \phi(\hat{\boldsymbol{\alpha}}(\boldsymbol{\gamma}), \boldsymbol{\gamma}) \\ &= \mathbf{y}^T(I - A(A^TA)^{-1}A^T)\mathbf{y} \\ &= \mathbf{y}^T(I - P_A)\mathbf{y} \end{aligned} \tag{7}$$

where $P_A$ is the orthogonal projection onto $\mathcal{R}(A)$. The least squares problem may be solved by minimizing $\psi$ with respect to $\boldsymbol{\gamma}$, and recovering the least squares estimate of $\boldsymbol{\alpha}$ from (6). Note that $X^TA = 0$, so that the columns of $X$ and $A$ span orthogonal spaces. Therefore (7) can be written more conveniently as

$$\psi(\boldsymbol{\gamma}) = \mathbf{y}^TP_X\mathbf{y}. \tag{8}$$

Since $P_X = X(X^TX)^{-1}X^T$, $\psi$ has partial derivatives

$$\begin{aligned} \dot{\psi}_i &= 2\mathbf{y}^TX_i(X^TX)^{-1}X^T\mathbf{y} - 2\mathbf{y}^TX(X^TX)^{-1}X_i^TX(X^TX)^{-1}X^T\mathbf{y} \\ &= 2\sum_{j=1}^{q+1}\boldsymbol{\gamma}_jB_{ij} \end{aligned}$$

with

$$B_{ij} = \mathbf{y}^T X_i (X^T X)^{-1} X_j^T \mathbf{y} - \mathbf{y}^T X (X^T X)^{-1} X_i^T X_j (X^T X)^{-1} X^T \mathbf{y}$$

Therefore, gathering the $B_{ij}$ into a symmetric matrix function of $\boldsymbol{\gamma}$, we have

$$\dot{\psi}(\boldsymbol{\gamma}) = 2B(\boldsymbol{\gamma})\boldsymbol{\gamma}.$$

Assuming that the constraint on $\boldsymbol{\gamma}$ takes the form $\boldsymbol{\gamma}^T\boldsymbol{\gamma} = 1$, then $\psi(\boldsymbol{\gamma})$ is minimized by a solution of the nonlinear system

$$\dot{\psi}(\boldsymbol{\gamma}) = 2\dot{B}(\boldsymbol{\gamma})\boldsymbol{\gamma} = \lambda\boldsymbol{\gamma} \tag{9}$$

where $\lambda$ is the Lagrange multiplier associated with the constraint. The interpretation and solution of this equation as a nonlinear eigenproblem is the basis of the modified Prony algorithm. Because of its fundamental role in the algorithm, we call $B$ the *Prony matrix*.

The derivation and definition of $B$ given above differs from Osborne (1975) in it's definition of $\boldsymbol{\gamma}$ and in that the components of $B$ are displayed as quadratic functions of $\mathbf{y}$. These changes make possible asymptotic arguments and distributional calculations. The various Prony parametrizations $\boldsymbol{\gamma}$ for which the normal equations can be expressed in the form (9) are in fact linearly related to one another up to a scale factor, and the specific choice amongst them does not affect the algorithm's rate of convergence near to a stationary value (Section 6).

# 4 Rational and Exponential Fitting

One class of linear homogeneous differential equations for which there are corresponding difference equations is

$$p_a(\mathcal{D})(q(t;\boldsymbol{\gamma})\mu(t)) = 0$$

where $p_a(\cdot)$ is any polynomial with known coefficients, and $q(t;\boldsymbol{\gamma})$ is any continuous function of time, linear and homogeneous in $\boldsymbol{\gamma}$. The general solution of these equations has the form

$$\mu(t) = \frac{f(t)}{q(t;\boldsymbol{\gamma})}$$

where $f(t)$ is the general solution of $p_a(\mathcal{D})f(t) = 0$.

Rational functions arise in the particular case that $p_a(\mathcal{D}) = \mathcal{D}^p$ and $q(t;\boldsymbol{\gamma}) = p_\gamma(t) = \sum_{j=1}^{q+1} \gamma_j t^{j-1}$. The general solution is then

$$\mu(t) = \frac{p_\alpha(t)}{p_\gamma(t)} \tag{10}$$

with $p_\alpha(t) = \sum_{j=1}^{p} \alpha_j t^{j-1}$. The rational function (10) satisfies the difference equation

$$\Delta^p(p_\gamma(t)\mu(t)) = 0 \tag{11}$$

which is of the form (1) with $d_k(t;\boldsymbol{\gamma}) = \binom{p}{k-1}\Delta^{p-k+1}p_\gamma(t)$. In matrix terms, (11) is

$$P\boldsymbol{\Delta}^p \langle p_\gamma(\mathbf{t})\rangle \boldsymbol{\mu} = 0$$

where $\boldsymbol{\Delta}$ is now the circulant difference matrix. This leads to

$$X^T = P\boldsymbol{\Delta}^p \langle p_\gamma(\mathbf{t})\rangle$$

and hence

$$X_j^T = \frac{dX^T}{d\gamma_j} = P\boldsymbol{\Delta}^p \left\langle \mathbf{t}^{j-1} \right\rangle$$

where $\langle \mathbf{t}^{j-1}\rangle$ denotes the diagonal matrix with elements $t_1^{j-1}$ to $t_n^{j-1}$.

Let $T$ be the $n \times p$ matrix with elements $T_{ij} = t_i^{j-1}$, so that $p_\alpha(\mathbf{t}) = T\boldsymbol{\alpha}$. For rational fitting, the straightforward choice of $C^T$ to be $\boldsymbol{\Delta}^p \langle p_\gamma(\mathbf{t})\rangle$ is not invertible. If we choose the last $p$ rows of $C^T$ to be instead $(T^TT)^{-1}T^T \langle p_\gamma(\mathbf{t})\rangle$, then the general formula (5) for $\boldsymbol{\mu}$ gives $A = \langle p_\gamma(\mathbf{t})\rangle^{-1} T$, and recovers the parametrization given above in (10).

Comparing rational functions with the general form (4), the denominator matrix $Z(\boldsymbol{\gamma})$ in this case is the diagonal matrix $\langle p_\gamma(\mathbf{t})\rangle$. This greatly simplifies the convergence proofs in Sections 9 and 10, and is the reason for focusing on rational functions in this paper. The matrix $H$ is the difference operator $\boldsymbol{\Delta}^p P^T$ which has the polynomials as its null space. Extension of rational fitting to non-equally spaced $t_i$ requires only a suitable choice of $H$. This is relatively straightforward; a direct formula is given in Osborne (1975).

Suppose now that $\mu(t)$ satisfies the constant coefficient differential equation

$$\sum_{k=1}^{p+1} \xi_k \mathcal{D}^{k-1}\mu(t) = 0,$$

and that the polynomial $p_\xi(z) = \sum_{k=1}^{p+1} \xi_k z^{k-1}$ has real, distinct roots $-\beta_j$. Then the general solution for $\mu(t)$ may be expressed as a sum of exponential functions

$$\mu(t) = \sum_{j=1}^{p} \alpha_j e^{-\beta_j t}, \tag{12}$$

and the corresponding difference equation is

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

where the polynomial $\sum_{k=1}^{p+1} \gamma_k z^{k-1}$ has roots $-\zeta_j = n(1-e^{-\beta_j/n})$. (Notice that $\zeta_j \to \beta_j$ as $n \to \infty$.) In matrix terms the difference equation is $PC^T\boldsymbol{\mu} = 0$ with

$$C^T = \sum_{k=1}^{p+1} \gamma_k \boldsymbol{\Delta}^{k-1}.$$

For exponential fitting, the $C_j^T$ are simply $\boldsymbol{\Delta}^{j-1}$, and $C$ is a, generally invertible, circulant matrix.

Since $C$ has $p+1$ bands, it has $p+1$ distinct non-zero elements, $\mathbf{c}^T = (c_1, \ldots, c_{p+1})$ say. In Osborne (1975), the algorithm was parametrized directly in terms of $\mathbf{c}$ rather than $\boldsymbol{\gamma}$. These two parametrizations will be compared in a separate paper on exponential fitting.

For exponential fitting, (5) expresses $\boldsymbol{\mu}$ as a linear combination of periodic solutions of the difference equation, rather than the more obvious definition with $A_{ij} = e^{-\beta_j t_i}$. Factoring $C$ here, using the complex eigendecomposition available for circulant matrices, shows that rational and exponential fitting are, in a sense, dual problems. Let $C = F\Lambda F^*$ where $F$ is the Fourier matrix (Davis, 1979) and $\Lambda$ is the diagonal matrix of complex eigenvalues. Then

$$F^*\boldsymbol{\mu} = \Lambda^{-1} F^* Q^T \boldsymbol{\alpha},$$

which expresses the discrete Fourier transform of $\boldsymbol{\mu}$ as a rational function of the complex frequency $e^{2\pi\sqrt{-1}/n}$. In other words, exponential fitting corresponds to rational fitting in the frequency domain.

# 5 A Modified Prony Algorithm

We now consider the solution of (9). Henceforth we will scale $\boldsymbol{\gamma}$ to have norm one, so the problem becomes that of minimizing the reduced sum of squares $\psi(\boldsymbol{\gamma})$ subject to this constraint. Let

$$F(\boldsymbol{\gamma}, \lambda) = \psi(\boldsymbol{\gamma}) + \lambda(1 - \boldsymbol{\gamma}^T\boldsymbol{\gamma})$$

where $\lambda$ is a Lagrange multiplier. Then

$$\dot{F}_\gamma = 2B(\boldsymbol{\gamma})\boldsymbol{\gamma} - 2\lambda\boldsymbol{\gamma}$$

and

$$\dot{F}_\lambda = 1 - \boldsymbol{\gamma}^T\boldsymbol{\gamma}$$

so the necessary conditions for a stationary point are

$$(B(\boldsymbol{\gamma}) - \lambda I)\boldsymbol{\gamma} = 0 \tag{13}$$

and

$$\boldsymbol{\gamma}^T\boldsymbol{\gamma} = 1.$$

We now show that $\lambda$ must be zero at a solution of (13). It can be seen from (8) that $\psi(\boldsymbol{\gamma})$ does not depend on $\|\boldsymbol{\gamma}\|$. Thus $\dot{\psi}$ must be in a direction orthogonal to $\boldsymbol{\gamma}$, so that

$$\boldsymbol{\gamma}^T \dot{\psi}(\boldsymbol{\gamma}) = 2\boldsymbol{\gamma}^T B(\boldsymbol{\gamma})\boldsymbol{\gamma} = 0. \tag{14}$$

Premultiplying $\dot{F}_\gamma = 0$ by $\boldsymbol{\gamma}^T$, and using (14) gives the result that $\lambda = 0$. This suggests the following iteration. Given a current estimate $\boldsymbol{\gamma}^k$, solve the linear eigenproblem

$$\begin{aligned} (B(\boldsymbol{\gamma}^k) - \lambda^{k+1} I)\boldsymbol{\gamma}^{k+1} &= 0 \\ {\boldsymbol{\gamma}^{k+1}}^T \boldsymbol{\gamma}^{k+1} &= 1 \end{aligned} \tag{15}$$

for $\boldsymbol{\gamma}^{k+1}$ and $\lambda^{k+1}$, with $\lambda^{k+1}$ the nearest to zero of such solutions. Convergence can be accepted when $\lambda^{k+1}$ is small compared with $\|B\|$. This iterative scheme is the *modified Prony algorithm*.

## Constraint Invariance

The constraint $\boldsymbol{\gamma}^T\boldsymbol{\gamma} = 1$, which we use here and in the remainder of this paper, is of course to some extent arbitrary. In particular, the result that the Lagrange multiplier is zero at a solution holds for any constraint $G(\boldsymbol{\gamma}) = 0$ such that $\hat{\boldsymbol{\gamma}}^T \dot{G}(\hat{\boldsymbol{\gamma}}) \neq 0$, for example for any linear constraint $\boldsymbol{\kappa}^T\boldsymbol{\gamma} = 0$ such that $\boldsymbol{\kappa}$ is not orthogonal to $\hat{\boldsymbol{\gamma}}$. Quadratic norm constraints are of special interest, because they correspond to linear reparametrizations of the Prony parameters. To see this, suppose we had chosen $\|\boldsymbol{\gamma}\| = 1$ with respect to some other inner product, say $\boldsymbol{\gamma}^T A\boldsymbol{\gamma} = 1$ with $A$ a positive definite matrix. The objective function would have been

$$F(\boldsymbol{\gamma}, \lambda) = \psi(\boldsymbol{\gamma}) + \lambda(1 - \boldsymbol{\gamma}^T A\boldsymbol{\gamma})$$

leading to the generalized nonlinear eigenproblem

$$(B(\boldsymbol{\gamma}) - \lambda A)\boldsymbol{\gamma} = 0$$

and to the sequence of linear problems

$$(B(\boldsymbol{\gamma}^k) - \lambda^{k+1} A)\boldsymbol{\gamma}^{k+1} = 0.$$

This is equivalent to

$$(A^{-\frac{1}{2}} B(\boldsymbol{\gamma}^k) A^{-\frac{T}{2}} - \lambda^{k+1} I)\boldsymbol{\delta}^{k+1} = 0$$

where $\boldsymbol{\delta} = A^{\frac{T}{2}}\boldsymbol{\gamma}$ and $A^{\frac{1}{2}}$ is any square root satisfying $A^{\frac{1}{2}} A^{\frac{T}{2}} = A$. If $B$ has a zero at $\hat{\boldsymbol{\gamma}}$ then $A^{-\frac{1}{2}} B A^{-\frac{T}{2}}$ has a zero at $\hat{\boldsymbol{\delta}} = A^{\frac{T}{2}}\hat{\boldsymbol{\gamma}}$, so using $A$ as the inner product matrix is equivalent to the linear reparametrization from $\boldsymbol{\gamma}$ to $\boldsymbol{\delta}$. This leads to a different modified Prony sequence, but one which has the same ultimate rate of convergence, as we show in Section 6.

## Implementation

When implemented efficiently, an iteration of the modified Prony algorithm requires computer time roughly similar to that of an iteration of a Gauss-Newton type algorithm. The linear eigenproblem (15) can be solved extremely rapidly by inverse iteration (Wilkinson, 1965). When testing for convergence, only an order of magnitude estimate of $\|B\|$ is required: we have found $\frac{1}{q+1}\sum_{ij}|B_{ij}|$ to be adequate. Our version of the modified Prony algorithm is given below:

$\boldsymbol{\gamma}^0$ := starting value
$\lambda^0 := 0$
$k := 0$
repeat
  $\epsilon := \tau\frac{1}{q+1}\sum_{ij}|B_{ij}|$
  $\mathbf{v}^0 := \boldsymbol{\gamma}^k$
  $\eta^0 := \lambda^k$
  $\ell := 0$
  {find eigenvalue of B closest $\lambda^k$ by inverse iteration}
  repeat
    $\mathbf{w}^{\ell+1} := (B-\eta^\ell I)^{-1}\mathbf{v}^\ell$
    {normalize $\mathbf{w}^{\ell+1}$ for numerical stability}
    $\mathbf{v}^{\ell+1} := \mathbf{w}^{\ell+1}/\|\mathbf{w}^{\ell+1}\|_\infty$
    $\mathbf{w}^{\ell+2} := (B-\eta^\ell I)^{-1}\mathbf{v}^{\ell+1}$
    $\eta^{\ell+2} := \eta^\ell + \mathbf{w}^{\ell+2^T}\mathbf{v}^{\ell+1}/\mathbf{w}^{\ell+2^T}\mathbf{w}^{\ell+2}$
    $\mathbf{v}^{\ell+2} := \mathbf{w}^{\ell+2}/\|\mathbf{w}^{\ell+2}\|_2$
    $\ell := \ell+2$
  until $|\eta^\ell - \eta^{\ell-2}| < \epsilon$
  $\boldsymbol{\gamma}^{k+1} := \mathbf{v}^\ell$
  $\lambda^{k+1} := \eta^\ell$
  $k := k+1$
until $|\lambda^k| < \epsilon$

In the above algorithm, the symmetric system

$$(B-\eta^\ell I)\mathbf{w}^{\ell+1} = \mathbf{v}^\ell \tag{16}$$

may be solved by Gaussian elimination, or by a method specially adapted to symmetric systems such as diagonal pivoting (Bunch and Kaufman, 1977). Inverse iteration is implemented in a similar way to that suggested by Osborne (1978), with the feature that $B-\eta^\ell I$ is factored once to solve two linear systems. The choice of the constant $\tau$ (tolerance) reflects the precision required and the maximum condition number of a matrix that can be stored in finite arithmetic. For example, in double precision on a 36 bit machine, $\tau = 10^{-15}$ gives precise solutions.

Empirical evidence with exponential fitting suggests that the modified Prony algorithm has an impressive ability to converge to some stationary value even from quite poor starting values. For exponential fitting the algorithm may be started from $\boldsymbol{\gamma} = \mathbf{e}_1$ (the first coordinate vector), for which $\mu(t)$ is constant. For rational fitting, in contrast, good starting values are often required and no general recipe can be given.

## Calculation of $B$

The efficient calculation of the Prony matrix $B$, is the key to the efficient implementation of the algorithm. This may be done in $O(nq^2)$ flops, for which the banded structure of $X$ is crucial. The following scheme is suggested for computation.

| | |
|---|---|
| $Y := (X_1^T\mathbf{y} \ \ldots \ X_{q+1}^T\mathbf{y})$ | An $(n-p)\times(q+1)$ matrix which does not depend on $\boldsymbol{\gamma}$. |
| $M := (X^TX)^{\frac{1}{2}}$ | The Choleski factor, a lower triangular $(n-p)^2$ matrix with $q+1$ bands. |
| $Y := M^{-1}Y$ | $(X^TX)^{-\frac{1}{2}}X_j^T\mathbf{y}, j = 1,\ldots,q+1.$ |
| $\mathbf{v} := M^{-T}Y\boldsymbol{\gamma}$ | $(X^TX)^{-1}X^T\mathbf{y}$, an $n-p$ vector. |
| $V := (X_1\mathbf{v} \ \ldots \ X_{q+1}\mathbf{v})$ | $X_j(X^TX)^{-1}X^T\mathbf{y}, j = 1,\ldots,q+1,$ an $n\times(q+1)$ matrix. |
| $B := Y^TY - V^TV$ | |

If desired, the fitted means may be extracted easily by

$$\hat{\mu} = P_X\mathbf{y} = V\boldsymbol{\gamma}\ . \tag{17}$$

For rational fitting the $X_j^T\mathbf{y}$ are $P\boldsymbol{\Delta}^p\left\langle\mathbf{t}^{j-1}\right\rangle\mathbf{y}$, which can be calculated by repeatedly differencing the $\left\langle\mathbf{t}^{j-1}\right\rangle\mathbf{y}$. Also $X^T$ is $P\boldsymbol{\Delta}^p\left\langle p_\gamma(\mathbf{t})\right\rangle$. For rational fitting, it is efficient to obtain $V^TV$ from $\mathbf{w} = \boldsymbol{\Delta}^{pT}P^T\mathbf{v}$ and

$$(V^TV)_{ij} = \mathbf{w}^T\left\langle\mathbf{t}^{i+j-2}\right\rangle\mathbf{w}$$

without first calculating and storing $V$.

For exponential fitting, $X_j^T\mathbf{y} = P\boldsymbol{\Delta}^{j-1}\mathbf{y}$, which can be calculated by recursive differencing. Similarly for $X_j\mathbf{v} = \boldsymbol{\Delta}^{j-1T}P^T\mathbf{v}$. In the exponential fitting case, the matrix $X$ is Toeplitz as well as banded, so only the $p+1$ nonzero components of the first column need to be stored. These components,

$\mathbf{c}^T = (c_1, \ldots, c_{p+1})$ say, are related to $\boldsymbol{\gamma}$ through $\mathbf{c} = U\boldsymbol{\gamma}$ with

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

and again this can be calculated recursively.

# 6 The Convergence Matrix

In this section we derive sufficient conditions for local convergence of the modified Prony algorithm. Let us express the modified Prony iteration formally as

$$\boldsymbol{\gamma}^{k+1} = F(\boldsymbol{\gamma}^k)$$

where $F$ is defined implicitly by (15). The iteration has a fixed point at the least squares estimate. It has a point of attraction there also if it converges to $\hat{\boldsymbol{\gamma}}$ from any point in a surrounding neighbourhood. In that case we say it is *stable* at $\hat{\boldsymbol{\gamma}}$. A sufficient condition for stability is given by Ostrowski's Theorem (Ortega and Rheinboldt, 1970), namely that $F$ is Frechét differentiable and

$$\rho(\dot{F}(\hat{\boldsymbol{\gamma}})) < 1$$

where $\rho(\cdot)$ denotes spectral radius. Following Smyth (1987) and Kass and Smyth (1989), we call

$$\dot{F}(\hat{\boldsymbol{\gamma}}) = \frac{d\boldsymbol{\gamma}^{k+1}}{d\boldsymbol{\gamma}^k}(\hat{\boldsymbol{\gamma}})$$

the *convergence matrix* and $\rho(\dot{F}(\hat{\boldsymbol{\gamma}}))$ the *convergence factor*.

We now obtain the convergence matrix of the modified Prony algorithm. The theorem assumes that the null space of $B(\hat{\boldsymbol{\gamma}})$ has dimension one, so that it is spanned by $\hat{\boldsymbol{\gamma}}$. That this is asymptotically so, is proved in Section 9. The derivative $\dot{B}$ of $B$ with respect to $\boldsymbol{\gamma}$ is a three dimensional tensor, to be thought of as an array of partial derivatives $\frac{\partial B}{\partial \gamma_k}, k = 1, \ldots, q+1$. Therefore $\dot{B}(\boldsymbol{\gamma})\boldsymbol{\gamma}$ denotes the $(q+1) \times (q+1)$ matrix with k$^{th}$ column $\frac{\partial B}{\partial \gamma_k}\boldsymbol{\gamma}$.

**Lemma 1** *For all* $\boldsymbol{\gamma}$, $\boldsymbol{\gamma}^T\dot{\boldsymbol{\gamma}} = 0$ *and* $\boldsymbol{\gamma}^T B(\boldsymbol{\gamma})\boldsymbol{\gamma} = 0$. *At the least squares estimate* $\hat{\boldsymbol{\gamma}}^T\dot{B}(\hat{\boldsymbol{\gamma}})\hat{\boldsymbol{\gamma}} = 0$.

**Proof** The first result follows from differentiating $\boldsymbol{\gamma}^T\boldsymbol{\gamma} = 1$. The second has already been mentioned in Section 5, and follows from a geometric argument. The third follows from it by differentiating, since $B(\hat{\boldsymbol{\gamma}})\hat{\boldsymbol{\gamma}} = 0$. □

**Theorem 1** *The convergence matrix of the modified Prony algorithm is*

$$B(\hat{\boldsymbol{\gamma}})^{+}\dot{B}(\hat{\boldsymbol{\gamma}})\hat{\boldsymbol{\gamma}}$$

*where $B(\hat{\boldsymbol{\gamma}})^{+}$ is the Moore-Penrose inverse of $B(\hat{\boldsymbol{\gamma}})$.*

**Proof** Differentiating (15) with respect to $\boldsymbol{\gamma}^k$, and evaluating at $\hat{\boldsymbol{\gamma}}$ gives

$$(\dot{B}(\hat{\boldsymbol{\gamma}}) - \dot{\lambda}(\hat{\boldsymbol{\gamma}})I)\hat{\boldsymbol{\gamma}} + (B(\hat{\boldsymbol{\gamma}}) - \lambda(\hat{\boldsymbol{\gamma}})I)\dot{F}(\hat{\boldsymbol{\gamma}}) = 0. \tag{18}$$

It was shown in Section 5 that $\lambda(\hat{\boldsymbol{\gamma}}) = 0$. Premultiplying (18) by $\hat{\boldsymbol{\gamma}}^T$ and applying Lemma 1 and $B(\hat{\boldsymbol{\gamma}})\hat{\boldsymbol{\gamma}} = 0$, shows that $\dot{\lambda}(\hat{\boldsymbol{\gamma}}) = 0$ also. So we have

$$\dot{B}(\hat{\boldsymbol{\gamma}})\hat{\boldsymbol{\gamma}} + B(\hat{\boldsymbol{\gamma}})\dot{F}(\hat{\boldsymbol{\gamma}}) = 0.$$

But

$$B\dot{\boldsymbol{\gamma}} = (B + \boldsymbol{\gamma}\boldsymbol{\gamma}^T)\dot{\boldsymbol{\gamma}}$$

since $\boldsymbol{\gamma}^T\dot{\boldsymbol{\gamma}} = 0$, so

$$\begin{aligned}
\dot{F}(\hat{\boldsymbol{\gamma}}) &= -(B(\hat{\boldsymbol{\gamma}}) + \hat{\boldsymbol{\gamma}}\hat{\boldsymbol{\gamma}}^T)^{-1}\dot{B}(\hat{\boldsymbol{\gamma}})\hat{\boldsymbol{\gamma}} \\
&= -(B(\hat{\boldsymbol{\gamma}})^{+} + \hat{\boldsymbol{\gamma}}\hat{\boldsymbol{\gamma}}^T)\dot{B}(\hat{\boldsymbol{\gamma}})\hat{\boldsymbol{\gamma}} \\
&= -B(\hat{\boldsymbol{\gamma}})^{+}\dot{B}(\hat{\boldsymbol{\gamma}})\hat{\boldsymbol{\gamma}}
\end{aligned}$$

applying Lemma 1 again. □

Differentiating (9) gives

$$\ddot{\psi}(\boldsymbol{\gamma}) = 2B(\boldsymbol{\gamma}) + 2\dot{B}(\boldsymbol{\gamma})\boldsymbol{\gamma},$$

which shows that the convergence matrix measures the relative difference between $B$ and $\frac{1}{2}\ddot{\psi}$. This result is analogous to that for the Gauss-Newton algorithm, which converges when the Fisher information matrix $\dot{\mu}^T\dot{\mu}$ is a good approximation to $\frac{1}{2}\ddot{\phi}$.

The following lemma and theorem show that all linearly related Prony parametrizations have similar convergence matrices.

**Lemma 2** *If $B$ is a symmetric matrix with null space spanned by $\boldsymbol{\gamma}$ ($\boldsymbol{\gamma}^T\boldsymbol{\gamma} = 1$), then*

$$B^{+} = \lim_{\epsilon \to 0}(I - \boldsymbol{\gamma}\boldsymbol{\gamma}^T)(B + \epsilon\mathbf{r}\mathbf{r}^T)^{-1}(I - \boldsymbol{\gamma}\boldsymbol{\gamma}^T)$$

*for any $\mathbf{r}$ for which $\mathbf{r}^T\boldsymbol{\gamma} \neq 0$.*

**Proof** Let

$$B = \begin{pmatrix} Z & \boldsymbol{\gamma} \end{pmatrix} \begin{pmatrix} \Lambda & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} Z^T \\ \boldsymbol{\gamma}^T \end{pmatrix}$$

be an eigen-decomposition of $B$, so that

$$\begin{aligned} (B + \epsilon \mathbf{r}\mathbf{r}^T) &= \begin{pmatrix} Z & \boldsymbol{\gamma} \end{pmatrix} \left[ \begin{pmatrix} \Lambda & 0 \\ 0 & 0 \end{pmatrix} + \epsilon \begin{pmatrix} Z^T \\ \boldsymbol{\gamma}^T \end{pmatrix} \mathbf{r}\mathbf{r}^T \begin{pmatrix} Z & \boldsymbol{\gamma} \end{pmatrix} \right] \begin{pmatrix} Z^T \\ \boldsymbol{\gamma}^T \end{pmatrix} \\ &= \begin{pmatrix} Z & \boldsymbol{\gamma} \end{pmatrix} \left[ \begin{pmatrix} \Lambda & 0 \\ 0 & 0 \end{pmatrix} + \epsilon \begin{pmatrix} Z^T\mathbf{r}\mathbf{r}^T Z & Z^T\mathbf{r}\mathbf{r}^T\boldsymbol{\gamma} \\ \boldsymbol{\gamma}^T\mathbf{r}\mathbf{r}^T Z & (\boldsymbol{\gamma}^T\mathbf{r})^2 \end{pmatrix} \right] \begin{pmatrix} Z^T \\ \boldsymbol{\gamma}^T \end{pmatrix} \end{aligned}$$

and

$$(B + \epsilon \mathbf{r}\mathbf{r}^T)^{-1} = \begin{pmatrix} Z & \boldsymbol{\gamma} \end{pmatrix} \begin{bmatrix} (\Lambda + \epsilon Z^T\mathbf{r}\mathbf{r}^T Z + \epsilon \mathbf{r}^T Z Z^T \mathbf{r})^{-1} & \cdot \\ \cdot & \cdot \end{bmatrix} \begin{pmatrix} Z^T \\ \boldsymbol{\gamma}^T \end{pmatrix}$$

since $\boldsymbol{\gamma}^T\mathbf{r} \neq 0$. Hence

$$\begin{aligned} (I - \boldsymbol{\gamma}\boldsymbol{\gamma}^T)(B + \epsilon \mathbf{r}\mathbf{r}^T)^{-1}(I - \boldsymbol{\gamma}\boldsymbol{\gamma}^T) &= \begin{pmatrix} Z & 0 \end{pmatrix} \begin{bmatrix} \Lambda^{-1} + o(\epsilon) & \cdot \\ \cdot & \cdot \end{bmatrix} \begin{pmatrix} Z^T \\ 0 \end{pmatrix} \\ &= Z\left(\Lambda^{-1} + o(\epsilon)\right) Z^T \\ &= B^+ + o(\epsilon). \end{aligned}$$

□

**Theorem 2** *The eigenvalues of $B^+\dot{B}\boldsymbol{\gamma}$ are invariant under linear reparametrization.*

**Proof** Let $\boldsymbol{\delta} = U^{-1}\boldsymbol{\gamma}$ with $U$ a nonsingular known matrix, and let $B_\delta$ and $B_\gamma$ be the Prony matrices for $\boldsymbol{\delta}$ and $\boldsymbol{\gamma}$ respectively. Now

$$\dot{\psi}_\delta = U^T \dot{\psi}_\gamma,$$

so

$$B_\delta \boldsymbol{\delta} = U^T B_\gamma \boldsymbol{\gamma} = U^T B_\gamma U \boldsymbol{\delta}$$

and

$$B_\delta = U^T B_\gamma U.$$

Differentiating again

$$\begin{aligned} \dot{B}_\delta \boldsymbol{\delta} &= U^T(\dot{B}_\gamma U \boldsymbol{\delta})U \\ &= U^T(\dot{B}_\gamma \boldsymbol{\gamma})U. \end{aligned}$$

Hence

$$\begin{aligned} B_\delta^+ \dot{B}_\delta \boldsymbol{\delta} &= \lim_{\epsilon \to 0} (I - \boldsymbol{\delta}(\boldsymbol{\delta}^T\boldsymbol{\delta})^{-1}\boldsymbol{\delta}^T)(B_\delta + \epsilon \boldsymbol{\delta}\boldsymbol{\delta}^T)^{-1}\dot{B}_\delta \boldsymbol{\delta} \\ &= \lim_{\epsilon \to 0} (I - \boldsymbol{\delta}(\boldsymbol{\delta}^T\boldsymbol{\delta})^{-1}\boldsymbol{\delta}^T)(U^T B_\gamma U + \epsilon \boldsymbol{\delta}\boldsymbol{\delta}^T)^{-1}U^T(\dot{B}_\gamma \boldsymbol{\gamma})U \\ &= \lim_{\epsilon \to 0} U^{-1}U(I - \boldsymbol{\delta}(\boldsymbol{\delta}^T\boldsymbol{\delta})^{-1}\boldsymbol{\delta}^T)U^{-1}(B_\gamma + \epsilon U^{-T}\boldsymbol{\delta}\boldsymbol{\delta}^T U^{-1})^{-1}\dot{B}_\gamma \boldsymbol{\gamma} U \\ &= \lim_{\epsilon \to 0} U^{-1}(I - \boldsymbol{\gamma}(\boldsymbol{\gamma}^T\boldsymbol{\gamma})^{-1}\boldsymbol{\gamma}^T)(B_\gamma + \epsilon U^{-T}\boldsymbol{\delta}\boldsymbol{\delta}^T U^{-1})^{-1}\dot{B}_\gamma \boldsymbol{\gamma} U \\ &= U^{-1} B_\gamma^+ \dot{B}_\gamma \boldsymbol{\gamma} U, \end{aligned}$$

since $\boldsymbol{\delta}^T U^{-1}\boldsymbol{\gamma} = \boldsymbol{\delta}^T\boldsymbol{\delta} \neq 0$, and this is similar to

$$B_\gamma^+ \dot{B}_\gamma \boldsymbol{\gamma}.$$

□

# 7 Expectations

Calculating expectations for $B$ and $\dot{B}\boldsymbol{\gamma}$ shows that the convergence factor converges to zero as the variance of the $y_i$, $\sigma^2 \to 0$. (Sections 9 and 10 show that it also converges to zero for fixed $\sigma^2$ and $n \to \infty$.)

**Theorem 3**

$$\mathbb{E}(B(\boldsymbol{\gamma}_0)) = V_0$$

*where*

$$V_{0ij} = \boldsymbol{\mu}^T X_i (X^T X)^{-1} X_j^T \boldsymbol{\mu}$$

*and*

$$\mathbb{E}(\dot{B}(\boldsymbol{\gamma}_0)\boldsymbol{\gamma}_0) = 0.$$

**Proof** All expressions in this proof are to be evaluated at the true value $\boldsymbol{\gamma}_0$. Using $X^T\mathbf{y} = X^T(\mathbf{y} - \boldsymbol{\mu})$, and the standard identities for rearranging terms in matrix traces, we can write

$$\begin{aligned} B_{ij} &= \text{tr } X_i(X^TX)^{-1}X_j^T\mathbf{y}\mathbf{y}^T \\ &\quad -\text{tr } X(X^TX)^{-1}X_i^TX_j(X^TX)^{-1}X^T(\mathbf{y}-\boldsymbol{\mu})(\mathbf{y}-\boldsymbol{\mu})^T \end{aligned}$$

which has expectation

$$\text{tr } X_i(X^TX)^{-1}X_j^T(I\sigma^2 + \boldsymbol{\mu}\boldsymbol{\mu}^T) - \text{tr } X(X^TX)^{-1}X_i^TX_j(X^TX)^{-1}X^TI\sigma^2$$

$$= \boldsymbol{\mu}^T X_i (X^T X)^{-1} X_j^T \boldsymbol{\mu}$$

To do the same for $\dot{B}\boldsymbol{\gamma}$ we need an explicit expression. The simplest way is to differentiate

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

twice with respect to $\boldsymbol{\gamma}$, and to use

$$\dot{B}\boldsymbol{\gamma} = \frac{1}{2}\ddot{\psi} - B$$

This gives

$$\begin{aligned}
(\dot{B}\boldsymbol{\gamma})_{ij} \;=\; & -\mathbf{y}^T X_i (X^T X)^{-1} X_j^T X (X^T X)^{-1} X^T \mathbf{y} - \text{symmetry term} \\
& -\mathbf{y}^T X_i (X^T X)^{-1} X^T X_j (X^T X)^{-1} X^T \mathbf{y} - \text{symmetry term} \\
& +\mathbf{y}^T X (X^T X)^{-1} X_i^T X (X^T X)^{-1} X_j^T X (X^T X)^{-1} X^T \mathbf{y} \\
& +\text{symmetry term} \\
& +\mathbf{y}^T X (X^T X)^{-1} X^T X_i (X^T X)^{-1} X_j^T X (X^T X)^{-1} X^T \mathbf{y} \\
& +\mathbf{y}^T X (X^T X)^{-1} X_i^T X (X^T X)^{-1} X^T X_j (X^T X)^{-1} X^T \mathbf{y}
\end{aligned}$$

where "symmetry term" means as for the last term but with $i$ and $j$ interchanged. The theorem follows by writing $X^T\mathbf{y} = X^T(\mathbf{y} - \boldsymbol{\mu})$, applying the trace identities, and cancelling out like terms. □

**Corollary 1** *$B(\hat{\boldsymbol{\gamma}})^+ \dot{B}(\hat{\boldsymbol{\gamma}})\hat{\boldsymbol{\gamma}} \to 0$ as $\sigma^2 \to 0$.*

**Proof** As in the proof of Theorem 1, we use Lemma 1 to write

$$B(\hat{\boldsymbol{\gamma}})^+ \dot{B}(\hat{\boldsymbol{\gamma}})\hat{\boldsymbol{\gamma}} = (B(\hat{\boldsymbol{\gamma}}) + \hat{\boldsymbol{\gamma}}\hat{\boldsymbol{\gamma}}^T)^{-1} \dot{B}(\hat{\boldsymbol{\gamma}})\hat{\boldsymbol{\gamma}}. \tag{19}$$

Now $B(\boldsymbol{\gamma}_0) \to V_0$, $\dot{B}(\boldsymbol{\gamma}_0)\boldsymbol{\gamma}_0 \to 0$ and $\hat{\boldsymbol{\gamma}} \to \boldsymbol{\gamma}_0$ as $\sigma^2 \to 0$, so (19) has the same limit as

$$(B(\boldsymbol{\gamma}_0) + \boldsymbol{\gamma}_0\boldsymbol{\gamma}_0^T)^{-1} \dot{B}(\boldsymbol{\gamma}_0)\boldsymbol{\gamma}_0,$$

namely

$$(V_0 + \boldsymbol{\gamma}_0\boldsymbol{\gamma}_0^T)^{-1} 0 = 0.$$

□

Note that

$$V_0 = \dot{\boldsymbol{\mu}}_\gamma^T P_X \dot{\boldsymbol{\mu}}_\gamma$$

since $X^T\boldsymbol{\mu} = 0$ implies that

$$X^T \dot{\boldsymbol{\mu}}_\gamma = -\dot{X}_\gamma^T \boldsymbol{\mu}$$

One interpretation of $V_0$, is that it is the Fisher information matrix for $\boldsymbol{\gamma}$, conditional on the linear parameters $\boldsymbol{\alpha}$.

# 8 Asymptotic Assumptions

Most of the remainder of this paper will deal with asymptotic results, so we now need to make specific assumptions about our asymptotic framework. These are basically that $y$ and $\mu$ satisfy regularity conditions, and that $\mu$ satisfies a differential equation like (3) and its discrete analogue (1). The parameters $\boldsymbol{\xi}$ in (3) parametrize $p$ special solutions of the differential equation, and the function $\mu$ can be expressed as a linear combination of these, with coefficients $\alpha_1, \ldots, \alpha_p$ say. Let $\boldsymbol{\theta} = (\boldsymbol{\alpha}^T, \boldsymbol{\xi}^T)^T$. Assume that

(a) $\boldsymbol{\theta} \in \Theta$, where $\Theta$ is a compact subset of $\mathbb{R}^p \times S$ with $S$ the unit sphere in $\mathbb{R}^{q+1}$, which contains the true value $\boldsymbol{\theta}_0$ as an interior point. The errors $\epsilon_{n,i} = y_i - \mu_i(\boldsymbol{\theta}_0)$ form a triangular array of independent deviates, which are identically distributed with mean 0 and finite variance $\sigma^2$. The time points $t_i$ are equally spaced on the unit interval, so that $t_i = i/n$.

(b) The function
$$\int_0^1 (\mu(t;\boldsymbol{\theta}_0) - \mu(t;\boldsymbol{\theta}))^2 dt$$
has a unique minimum in $\Theta$ at $\boldsymbol{\theta}_0$.

(c) The function $\mu$ is twice continuously differentiable with respect to $\boldsymbol{\xi}$, and $p$-times continuously differentiable with respect to $t$. All these derivatives are jointly continuous in $t$ and $\boldsymbol{\theta}$.

(d) The information matrix $\mathcal{I}$, defined by
$$\mathcal{I}_{ij} = \int_0^1 \frac{\partial \mu}{\partial \theta_i}(t;\boldsymbol{\theta}_0)\frac{\partial \mu}{\partial \theta_j}(t;\boldsymbol{\theta}_0)dt$$
is positive semi-definite. Its null space has dimension one, and is spanned by $(0, \boldsymbol{\xi}_0^T)^T$.

(e) The difference equation (1) converges to the differential equation (3) in the sense that the $d_k$ converge to the $b_k$ and $\boldsymbol{\gamma}$ converges to $\boldsymbol{\xi}$.

Conditions (b) to (d) are similar to Jennrich (1969). Condition (b) asserts that $\mu$ is parametrized uniquely at $\boldsymbol{\theta}_0$, while (d) guarantees that $\boldsymbol{\theta}$ is estimable up to a scale factor for $\boldsymbol{\xi}$.

We need the following form of the law of large numbers.

**Theorem 4 (Law of large numbers)** *If $f(t;\boldsymbol{\theta})$ is jointly continuous on $[0,1] \times \Theta$, then*
$$\frac{1}{n}\sum_{i=1}^{n} f(t_i;\boldsymbol{\theta})\epsilon_{n,i} \to 0$$
*uniformly in $\boldsymbol{\theta}$.*

**Proof** This result is not proved in detail because it is of a standard type. For fixed $\boldsymbol{\theta}$, the result follows from Theorem 4.1.3 of Stout (1974) or from Corollary 1 of Stout (1968), and from their generalizations to triangular arrays of errors as described in Stout (1968). The proof that the result holds uniformly in $\boldsymbol{\theta}$ follows closely that of Theorem 4 of Jennrich (1969). □

Under the above conditions, the least squares estimator $\hat{\boldsymbol{\theta}}$ can be shown to be strongly consistent and asymptotically normal. Proofs would be similar to those of Jennrich (1969).

The above conditions generally hold for exponential fitting, and for rational fitting provided that the denominator polynomial $p_\gamma(t)$ is non-zero on $[0, 1]$.

# 9 Eigenstructure of $B$

**Theorem 5** *The least and greatest singular values of $X$ are $O(1)$ and $O(n^p)$ respectively.*

**Proof** The vectors which correspond to small singular values of $X$ are obtained from realizations of smooth functions at the time points $t_i$. Let $f$ be a $p$ times continuously differentiable function on $[0, 1]$ that is not a solution of the differential equation (3). Write

$$\|f\|^2 = \int_0^1 f(t)^2 dt$$

and let

$$\mathbf{z} = \|f(\mathbf{t})\|^{-1} f(\mathbf{t})$$

Now let $i \to \infty$ in such a way that $t_i = i/n \to t$. Then

$$\begin{aligned} (X^T\mathbf{z})_i &= \|f(\mathbf{t})\|^{-1} \sum_{k=1}^{p+1} d_k(t_i; \boldsymbol{\gamma}) \Delta^k f(t_i) \\ &\to \|f\|^{-1} \sum_{k=1}^{p+1} b_k(t; \boldsymbol{\xi}) \mathcal{D}^k f(t) \end{aligned}$$

which is a nonzero constant. On the other hand, the elements of $X$ are $O(n^p)$ through the contribution of $\boldsymbol{\Delta}^p$, so the largest singular value is $O(n^p)$. □

Theorem 5 implies that the condition number of $X$ will be large, and appears to have consequences for the numerical accuracy of $(X^TX)^{1/2}$, $B$ and $\hat{\boldsymbol{\mu}}$. Experience with the exponential and rational fitting examples suggests, though, that this problem is mitigated by the fact that $X$ possesses a compact analytic specification in terms of the Prony parameters.

The condition number of $B$ itself is $O(1)$, as is shown by the following theorem. The form of the proof given here is quite general, although it is completed only for rational functions. Rational functions offer the least complication because the operators $D_k$ (see proof) are already diagonal. For exponential fitting the proof must be prepared by a series of rather long lemmas which diagonalize the $D_k$ using discrete Fourier transform methods; these were proved by Smyth (1985) and, because of their length, will appear separately.

**Theorem 6** *In the rational fitting case,*

$$\frac{1}{n}B(\hat{\boldsymbol{\gamma}}) \stackrel{a.s.}{\to} V_0$$

*where*

$$V_0 = \lim_{n\to\infty} \frac{1}{n}\dot{\boldsymbol{\mu}}_\gamma^T P_X \dot{\boldsymbol{\mu}}_\gamma(\boldsymbol{\gamma}_0)$$

**Proof** We need to establish that $B(\boldsymbol{\gamma})$ converges almost surely, uniformly in a neighbourhood of $\boldsymbol{\gamma}_0$, to a continuous function $V(\boldsymbol{\gamma})$ which is equal to $V_0$ at $\boldsymbol{\gamma}_0$. It will be convenient to write

$$D_k = C_k C^{-1}$$

for $k = 1, \ldots, q+1$. In the rational fitting case, the $D_k$ are the diagonal matrices

$$\left\langle p_\gamma(\mathbf{t})^{-1}\mathbf{t}^{k-1}\right\rangle.$$

Let $P = (I_{n-p}\ 0)$. Substituting

$$X_i^T = PC_i^T = PC^TC^{-T}C_i^T = X^TD_i^T$$

into the expression for $B$ gives

$$B_{ij} = \mathbf{y}^T D_i P_X D_j^T \mathbf{y} - \mathbf{y}^T P_X D_i^T D_j P_X \mathbf{y}.$$

Expanding $P_X$ as $I - P_A$ (as in Section 3) and $\mathbf{y}$ as $\boldsymbol{\mu}_0 + \boldsymbol{\epsilon}$ gives

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

and we consider the terms appearing in this expansion individually. The two terms

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

cancel out, since $D_i$ and $D_j$ are diagonal matrices. Repeated application of Theorem 4 shows that all other terms which involve $\boldsymbol{\epsilon}$ converge to zero. Consider for example the term

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

The k$^{th}$ element of $\frac{1}{n}A^T D_j^T \boldsymbol{\epsilon}$ is

$$\frac{1}{n}\sum_{i=1}^{n} \frac{t_i^{j+k-2}}{p_\gamma(t_i)^2}\epsilon_{n,i}$$

which converges to zero almost surely by Theorem 4. The whole term converges to zero almost surely, since $\frac{1}{n}\boldsymbol{\mu}_0^T D_i A$ converges to a constant vector with elements

$$\int_0^1 \mu_o(t)\frac{t^{i+k-2}}{p_\gamma(t)^2}dt$$

for $k = 1, \ldots, p$, while $\frac{1}{n}A^T A$ converges to a constant positive definite matrix with elements

$$\int_0^1 \frac{t^{k+\ell}}{p_\gamma(t)^2}dt$$

for $k, \ell = 1, \ldots, p$. Moreover the convergence is uniform for $\boldsymbol{\gamma}$ in a compact set. Similarly, the term

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

causes no problems, since $\frac{1}{n}A^T D_i^T D_j A$ converges to a constant $p \times p$ matrix, and Theorem 4 is applied to show that each of the $\frac{1}{n}A^T\boldsymbol{\epsilon}$ converge to zero. This term is in fact of smaller order than the first.

The only terms in the expansion for $\frac{1}{n}B_{ij}$ which remain are

$$\frac{1}{n}\boldsymbol{\mu}_0^T D_i(I - P_A)D_j^T\boldsymbol{\mu}_0 - \frac{1}{n}\boldsymbol{\mu}_0^T(I - P_A)D_i^T D_j(I - P_A)\boldsymbol{\mu}_0,$$

and these converge to a constant, $V_{ij}$ say. Again the convergence is uniform for $\boldsymbol{\gamma}$ in a compact set. The proof is completed by gathering the $V_{ij}$ into a matrix function $V(\boldsymbol{\gamma})$ of $\boldsymbol{\gamma}$, and observing that $V(\boldsymbol{\gamma}_0) = V_0$. □

**Corollary 2** *With probability one, $n^{-1}B(\hat{\boldsymbol{\gamma}})$ has a positive semi-definite limit, the Moore-Penrose inverse of which is the asymptotic covariance matrix of $n^{1/2}\hat{\boldsymbol{\gamma}}$.*

**Corollary 3** *With probability one, the zero eigenvalue of $B(\hat{\boldsymbol{\gamma}})$ is asymptotically isolated with multiplicity one.*

**Proof** $n^{-1}B(\hat{\boldsymbol{\gamma}})$ asymptotically approximates $V_0$, which by the assumptions of Section 8 has one zero eigenvalue with eigenvector $\boldsymbol{\gamma}_0$. □

# 10 Asymptotic Stability

**Theorem 7** *The modified Prony iteration applied to rational fitting is almost surely asymptotically stable.*

**Proof** Theorem 1 gives an expression for the convergence matrix, which can be rewritten using Lemma 1 and Corollary 3 of Theorem 6 as

$$\left(\frac{1}{n}B(\hat{\boldsymbol{\gamma}}) + \hat{\boldsymbol{\gamma}}\hat{\boldsymbol{\gamma}}^T\right)^{-1} \frac{1}{n}\dot{B}(\hat{\boldsymbol{\gamma}})\hat{\boldsymbol{\gamma}}. \tag{20}$$

It is sufficient to prove that (20) converges to zero almost surely. Theorem 6 shows that $\frac{1}{n}B(\hat{\boldsymbol{\gamma}}) + \hat{\boldsymbol{\gamma}}\hat{\boldsymbol{\gamma}}^T$ converges to the positive definite matrix $V_0 + \boldsymbol{\gamma}_0\boldsymbol{\gamma}_0^T$. It can also be proved that

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

although the proof is not given here because it is tedious and follows the same form as that of Theorem 6 (it uses the expansion for $\dot{B}(\boldsymbol{\gamma})\boldsymbol{\gamma}$ given in Section 7). This establishes the proof: the matrix ratio (20) converges to zero because its numerator converges to zero, while its denominator converges to a positive definite matrix. □

The asymptotic stability of the modified Prony algorithm is actually closely related to the convergence of the usual Hessian of the sum of squares $\frac{1}{n}\ddot{\phi}$. Consider the rational function parametrization in terms of $\boldsymbol{\beta}^T = (\gamma_2 \dots \gamma_{q+1})/\gamma_1$. We can write $\boldsymbol{\mu} = A(\boldsymbol{\beta})\boldsymbol{\alpha}$ with

$$A_{ij} = \frac{t_i^{j-1}}{1 + \sum_{k=1}^{q} \beta_k t^k}.$$

The Jacobian of the transformation $\dot{\gamma}_\beta$ is a rectangular matrix with full column rank, and null space spanned by $\boldsymbol{\gamma}$. Therefore, $(\dot{\gamma}_\beta(\boldsymbol{\beta})\ \boldsymbol{\gamma})$ is a square nonsingular matrix. The, now standard, results of Jennrich (1969) can be applied to show that

$$\lim_{n\to\infty} \frac{1}{n}\ddot{\phi}_\theta(\hat{\boldsymbol{\theta}}) \overset{a.s.}{=} \lim_{n\to\infty} \frac{2}{n}\dot{\mu}_\theta(\boldsymbol{\theta}_0)^T\dot{\mu}_\theta(\boldsymbol{\theta}_0)$$

which is positive definite. So

$$\frac{1}{n}\ddot{\psi}_\beta(\hat{\boldsymbol{\beta}}) = \frac{1}{n}(\ddot{\phi}_\beta - \ddot{\phi}_{\beta\alpha}\ddot{\phi}_\alpha^{-1}\ddot{\phi}_{\alpha\beta})(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$$

almost surely has the same limit as

$$\frac{2}{n}\dot{\mu}_\beta(I - P_A)\dot{\mu}_\beta(\boldsymbol{\theta}_0) = \frac{2}{n}\dot{\mu}_\beta P_X\dot{\mu}_\beta(\boldsymbol{\theta}_0)$$

which again is positive definite. Therefore

$$\frac{1}{2n}\ddot{\psi}_\gamma(\hat{\boldsymbol{\gamma}}) = \frac{1}{n}\begin{pmatrix} \dot{\gamma}_\beta(\hat{\boldsymbol{\beta}}) & \hat{\boldsymbol{\gamma}} \end{pmatrix}^{-T} \begin{pmatrix} \ddot{\psi}_\beta(\hat{\boldsymbol{\beta}}) & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \dot{\gamma}_\beta(\hat{\boldsymbol{\beta}}) & \hat{\boldsymbol{\gamma}} \end{pmatrix}^{-1}$$

almost surely has the same limit as

$$\frac{2}{n}\dot{\mu}_\gamma^T P_X \dot{\mu}(\boldsymbol{\gamma}_0) = \frac{1}{n}\begin{pmatrix} \dot{\gamma}_\beta(\boldsymbol{\beta}_0) & \boldsymbol{\gamma}_0 \end{pmatrix}^{-T} \begin{pmatrix} \dot{\mu}_\beta^T P_X \dot{\mu}_\beta(\boldsymbol{\beta}_0) & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \dot{\gamma}_\beta(\boldsymbol{\beta}_0) & \boldsymbol{\gamma}_0 \end{pmatrix}^{-1}$$

which is positive semi-definite, with null space spanned by $\boldsymbol{\gamma}_0$. This limit may be recognized to be $V_0$. Therefore, the fact that

$$\frac{1}{n}\ddot{\psi}(\hat{\boldsymbol{\gamma}}) = \frac{1}{n}(B(\hat{\boldsymbol{\gamma}}) + \dot{B}(\hat{\boldsymbol{\gamma}})\hat{\boldsymbol{\gamma}}) \stackrel{a.s.}{\to} V_0$$

can be seen to follow from the usual convergence results. (This approach was taken by Osborne and Smyth (1986).)

# 11 A Numerical Experiment

The purpose of this section is to compare the modified Prony algorithm with a good general purpose nonlinear squares procedure, namely the Levenberg modification of the Gauss-Newton algorithm (Fletcher, 1980). Osborne (1975) reported some results for exponential fitting (very favorable for the Prony algorithm), so the results here are for rational fitting. The modified Prony algorithm was implemented as described in Section 5. The symmetric linear system appearing in the inverse iteration sequence was solved by diagonal pivoting, as implemented by J. Bunch in LINPACK (Dongarra et al, 1979). The Levenberg algorithm was implemented essentially as described by Osborne (1976), the Levenberg parameter having expansion factor 2, contraction factor 10 and initial value 1.

The convergence criterion used by the Levenberg algorithm was

$$(\text{ssf}^{1/2} - \text{ssr}^{1/2})/(1 + \text{ssf}^{1/2}) < \tau, \tag{21}$$

where *ssf* is the actual sum of squares and *ssr* is the sum of squares based on a linearization of the problem, and the tolerance $\tau$ was set to $10^{-7}$. Although the Prony and Levenberg convergence criteria are not strictly comparable, the Prony tolerance parameter was adjusted to $10^{-10}$ so that the two algorithms returned estimates that were on average of the same precision.

All calculations were performed in double precision in Fortran 77 on a Sperry Univac 1100/82 computer.

## 11.1 Test Problem

Data was simulated using the mean function

$$\mu(t) = (\alpha_1 + \alpha_2 t)/(1 + \beta_1 t + \beta_2 t^2) = (.5 + .5t)/(1 - .5t + .1t^2).$$

Data sets were constructed to have standard deviations $\sigma = .03, .01, .003, .001$ and sample sizes $n = 32, 64, 128, 256, 512$. Random deviates were associated with the means in such as way that each data set comprised every second point of the next larger data set. This was done so that sample sizes could be compared along a common sequence of random deviates.

Ten replicates were generated for each of the four distributions: the normal, student's t on 3 d.f. (infinite third moments), lognormal (skew) and Pareto's distribution with $k = 1$ and $\alpha = 3$ (skew and infinite third moments). Random deviates were generated by inverting the distribution functions, i.e.

$$\epsilon = \sigma F^{-1}(U)$$

with $U$ a pseudo random number generated by the NAG subroutine G05CAF (Numerical Algorithms Group, 1983) with seed equal to 1984. The first 10,000 values generated by G05CAF were ignored (most were used for exponential fitting simulations), while the next 5120 were used for the simulations reported here.

In order to make an objective choice, the true parameter values themselves were used as starting values. These were quite far from the least squares estimates for small $n$ and large $\sigma$, less so for large $n$ and small $\sigma$, as can be seen from Table 3.

## 11.2 Results

As it turned out, the convergence results were similar for all four distributions; for example, Table 1 gives the iteration counts for the modified Prony algorithm for the first replicate. It appears that the convergence rates achieved by the algorithms are little affected by skewness or by the third and higher moments of the error distribution (although the actual least squares estimates returned will be affected). So only the results for the normal distribution are reported below.

Iteration counts for the normal simulations are given in Table 2. As expected from the form of the convergence matrices, and also from the proximity of the starting values to the final estimates, both algorithms converged more rapidly for large $n$ and small $\sigma$. (See Smyth (1987) and Kass and Smyth (1989) for a discussion of the convergence matrix for the Gauss-Newton algorithm.) The pattern for the Levenberg algorithm was obscured somewhat by its careful convergence criterion (21), which caused it to produce relatively

more precise estimates when the sum of squares was small. This is discussed further below. In most cases the Prony algorithm required slightly fewer iterations than the Levenberg, although both algorithms were finished within 6 iterations most of the time even for $n = 32$ and $\sigma = .03$. Computer time was not strictly recorded, but the two algorithms took similar amounts of elapsed time. Table 3 gives the means and standard deviations of $\hat{\beta}_1$ and $\hat{\beta}_2$. For two samples with $n = 32$ and $\sigma = .03$, Prony converged to a stationary value which was not the least squares estimate; except for these, the programs returned effectively identical estimates. The two samples for which Prony failed gave rise to final Prony matrices $B(\hat{\boldsymbol{\gamma}})$ which were indefinite, while all other cases gave positive semi-definite matrices.

Two changes to the Levenberg algorithm, specifically tuning it to this problem, allowed it to converge rather more rapidly, as shown in Table 4. The first was to set the Levenberg parameter to zero, so that the algorithm reduced to the unmodified Gauss-Newton algorithm. The second was to remove the constant from the denominator of the convergence criterion (21). This constant is a standard safety device in general purpose programs, but when $ssf$ is small it causes the program to seek a small absolute rather than relative different between $ssf^{1/2}$ and $ssr^{1/2}$. Without the constant, the number of iterations required decreases consistently with $n$, as expected. Although the precision of estimation accepted by the new criterion is no longer equivalent to that of Prony, the differences between the parameter estimates returned are very small.

Experimentation showed that both Prony and Gauss-Newton algorithms were sensitive to the starting values for small $n$ and large $\sigma$. This is in contrast with results for exponential fitting, for which the modified Prony algorithm is remarkably robust with respect to the starting value.

We conclude that, for this rational fitting problem, the modified Prony algorithm is intermediate in performance between a general purpose Levenberg algorithm and a Gauss-Newton type algorithm specially tuned to the problem.

Simulations for exponential fitting will be given in a separate paper, which will focus on the special problems of exponential fitting. Curiously, the modified Prony algorithm actually appears to perform better for exponential fitting than it does for rational fitting, despite the well known fact that other algorithms find exponential fitting very much more difficult (as did the Levenberg algorithm in our simulations).

# Acknowledgement

The authors are grateful to Professor E.J. Hannan for valuable discussions.

# References


Bates, D.M. and Watts, D.G. (1988). *Nonlinear Regression Analysis and its Applications*, Wiley, New York.

Bunch. J.R. and Kaufmann, L. (1977). 'Some stable methods for calculating inertia and solving symmetric linear systems', *Math. Comp.* 31:162–179.

Davis, P.J. (1979). *Circulant Matrices*, Wiley, New York.

Dongarra, J.J. et al. (1979). *LINPACK Users' Guide,* SIAM, Philadelphia.

Fletcher, R. (1980). *Practical Methods of Optimization Vol. 1: Unconstrained Optimization*, Wiley, Chichester.

Householder, A.S. (1950). 'On Prony's method of fitting exponential decay curves and multiple-hit survival curves', *Oak Ridge National Lab. Report ORNL-455*, Oak Ridge, Tennessee.

Jennrich, R.I. (1969). 'Asymptotic properties of non-linear least squares estimators', *Ann. Math. Statist.* 40:633–643.

Kass, R. and Smyth, G.K. (1989). 'Convergence of the Scoring Algorithm: a Geometric Interpretation'. In preparation.

Marple, S.L. (1987). *Digital Spectral Analysis with Applications*, Prentice-Hall, Englewood Cliffs, New Jersey.

Mulholland, R.J., Cruz, J.R. and Hill, J. (1986). 'State-variable canonical forms for Prony's method', *Int. J. Systems Sci.* 17:55–64.

Numerical Algorithms Group (1983). *NAG FORTRAN library manual Mark 10*, Numerical Algorithms Group, Oxford.

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

Osborne, M.R. (1975). 'Some special nonlinear least squares problems', *SIAM J. Numer. Anal.* 12:571–592.

Osborne, M.R. (1976). 'Nonlinear least squares — the Levenberg algorithm revisited', *J. Austral. Math. Soc. B* 19:343–357.

Osborne, M.R. (1978). 'Inverse iteration, Newton's method, and non-linear eigenvalue problems', Proceedings of Conference in Honour of J.H. Wilkinson, Institute of Mathematics and its Applications, Royal Society, London. Pub. 19:21–53.

Osborne, M.R. and Smyth, G.K. (1986). 'An algorithm for exponential fitting revisited', Contribution to Essays in Time Series and Allied Processes: Papers in honour of E. J. Hannan. J. Gani and M. B. Priestley (Editors). Applied Probability Trust, Sheffield 419–430.

Prony, R. (1795). 'Essai éxperimental et analytique: sur les lois de la dilatabilité de fluides élastiques et sur celles de la force expansive de la vapeur de léau et de la vapeur de l'alkool, à différentes températures', *J. de L'École Polytechnique* Vol. 1, no. 2:24–76.

Ratkowsky, D.A. (1983). *Nonlinear Regression Modelling: A Unified Practical Approach*, Marcel Dekker, New York.

Smyth, G.K. (1985). *Coupled and Separable Iterations in Nonlinear Estimation*, PhD Thesis, Australian National University, Canberra.

Smyth, G.K. (1987). 'Curvature and Convergence.' Proceedings of the Statistical Computing Section. American Statistical Society, Virginia 278–283.

Stout, W.F. (1968). 'Some results on the complete and almost sure convergences of linear combinations of independent random variables and martingale differences', *Ann. Math. Statist.* 5:1549–1562.

Stout, W.F. (1974). *Almost Sure Convergence*, Academic Press, New York.

Wilkinson, J.H. (1965). *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford.

Table 1: Iteration counts for the modified Prony algorithm for one replicate of four distributions.

| $\sigma$ | $n$ | Normal | Student | Log-normal | Pareto |
|---|---|---|---|---|---|
| .030 | 32 | 5 | 5 | 5 | 4 |
| | 64 | 4 | 5 | 4 | 4 |
| | 128 | 4 | 4 | 4 | 4 |
| | 256 | 5 | 4 | 4 | 4 |
| | 512 | 4 | 4 | 4 | 4 |
| .010 | 32 | 4 | 4 | 3 | 3 |
| | 64 | 4 | 4 | 3 | 3 |
| | 128 | 3 | 3 | 3 | 3 |
| | 256 | 3 | 3 | 3 | 3 |
| | 512 | 3 | 3 | 3 | 3 |
| .003 | 32 | 3 | 3 | 3 | 3 |
| | 64 | 3 | 3 | 3 | 3 |
| | 128 | 2 | 2 | 2 | 2 |
| | 256 | 2 | 2 | 2 | 2 |
| | 512 | 3 | 3 | 2 | 2 |
| .001 | 32 | 2 | 2 | 2 | 2 |
| | 64 | 2 | 2 | 2 | 2 |
| | 128 | 2 | 2 | 2 | 2 |
| | 256 | 2 | 2 | 2 | 2 |
| | 512 | 2 | 2 | 2 | 2 |

Table 2: Median and maximum iteration counts for normal simulations. Results for the Prony algorithm are given above those for the Levenberg algorithm.

| $n\backslash\sigma$ | .030 | | .010 | | .003 | | .001 | |
|---|---|---|---|---|---|---|---|---|
| 32 | 5 | 17 | 4 | 6 | 3 | 3 | 2 | 3 |
| | 6 | 10 | 5 | 7 | 5 | 6 | 5 | 5 |
| 64 | 5 | 12 | 4 | 5 | 3 | 3 | 2 | 2 |
| | 6 | 8 | 5.5 | 6 | 5 | 5 | 5 | 5 |
| 128 | 6 | 7 | 4 | 4 | 2.5 | 3 | 2 | 3 |
| | 6 | 7 | 5 | 6 | 5 | 5 | 5 | 5 |
| 256 | 4.5 | 6 | 3 | 4 | 2 | 3 | 2 | 2 |
| | 6 | 6 | 5 | 6 | 5 | 5 | 5 | 5 |
| 512 | 4 | 5 | 3 | 3 | 3 | 3 | 2 | 2 |
| | 6 | 6 | 5 | 6 | 5 | 5 | 5 | 5 |

Table 3: Means and standard deviations, over 10 replicates, of the least squares estimates of $\beta_1$ and $\beta_2$. True values are $\beta_1 = -.5$ and $\beta_2 = .1$.

| $n\backslash\sigma$ | .03 | .01 | .003 | .001 |
|---|---|---|---|---|
| 32 | $-.2910(.24)$ | $-.4474(.055)$ | $-.4853(.015)$ | $-.4952(.005)$ |
| | $-0.238(.14)$ | $.0692(.034)$ | $.0914(.095)$ | $.0972(.003)$ |
| 64 | $-.5042(.21)$ | $-.5025(.066)$ | $-.5099(.020)$ | $-.5003(.007)$ |
| | $.1033(.12)$ | $.1017(.038)$ | $.1006(.011)$ | $.1002(.004)$ |
| 128 | $-.4694(.14)$ | $-.4918(.046)$ | $-.4977(.013)$ | $-.4993(.005)$ |
| | $.0816(.08)$ | $.0950(.027)$ | $.0986(.008)$ | $.9954(.003)$ |
| 256 | $-.4979(.13)$ | $-.5002(.041)$ | $-.5002(.012)$ | $-.5001(.004)$ |
| | $.0966(.08)$ | $.0994(.025)$ | $.0999(.007)$ | $.1000(.002)$ |
| 512 | $-.5180(.16)$ | $-.5080(.051)$ | $-.5026(0.15)$ | $-.5009(.005)$ |
| | $.1097(.09)$ | $.1044(.030)$ | $.1014(.009)$ | $.1005(.003)$ |

Table 4: Median and maximum iteration counts for unmodified Gauss-Newton with a simplified convergence criterion.

| $n\backslash\sigma$ | .030 | | .010 | | .003 | | .001 | |
|---|---|---|---|---|---|---|---|---|
| 32 | 4 | 5 | 3 | 3 | 3 | 3 | 2 | 3 |
| 64 | 3 | 5 | 3 | 4 | 3 | 3 | 2.5 | 3 |
| 128 | 3 | 5 | 3 | 3 | 2.5 | 3 | 2 | 3 |
| 256 | 3 | 3 | 2 | 3 | 2 | 3 | 2 | 2 |
| 512 | 3 | 3 | 2 | 3 | 2 | 3 | 2 | 2 |