On the consistency of Prony's method and related algorithms*

M.H. Kahn† M.S.Mackisack‡ M.R.Osborne§ G.K.Smyth¶

January 1992

Abstract

Modification's of Prony's classical technique for estimating rate constants in exponential fitting problems have many contemporary applications. Here the consistency of Prony's method and of related algorithms based on maximum likelihood is discussed as the number of observations $n \to \infty$ by considering the simplest possible models for fitting sums of exponentials to observed data. Two sampling regimes are relevant, corresponding to transient problems and problems of frequency estimation; and these are associated with rather different kinds of behaviour. The general pattern is that the stronger results are obtained for the frequency estimation problem. However, the algorithms considered are all scaling dependent and consistency is not automatic. *A new feature emerges which is the importance of an appropriate choice of scale in order to ensure consistency of the estimates in certain cases.* The tentative conclusion is that algorithms referred to as ORA (Objective function Reweighting Algorithm) are superior to their exact maximum likelihood counterparts referred to as GRA (Gradient condition Reweighting Algorithm), especially in the frequency estimation problem. This conclusion does not extend to fitting other families of functions such as rational functions.

1 Introduction

The basic problem considered in this paper is the estimation of the parameter vector $\boldsymbol{\beta} \in R^p$ in fitting to observations $y_i,\ i = 1, 2, \ldots, n$ by models of the form

$$y_i = \sum_{k=1}^{p} \alpha_k \phi_k(\beta_k, t_i) + \epsilon_i \tag{1}$$

*Kahn, M., Mackisack, M. S., Osborne, M. R., and Smyth, G. K. (1992). On the consistency of Prony's method and related algorithms. *J. Comput. Graph. Statist.* **1**, 329–349.

†Statistics Research Section,Mathematical Sciences School,Australian National University

‡School of Mathematics, Queensland University of Technology

§Statistics Research Section,Mathematical Sciences School,Australian National University

¶Department of Mathematics, University of Queensland

where the random variables ϵ_i, $i = 1, 2, \ldots, n$ are assumed independent with mean zero and variance σ^2, and where the further assumption of normality is made to simplify subsequent calculations. Here $\phi_k(\beta_k, t)$ is given either by

- **Model 1:** $\phi_k(\beta_k, t) = e^{-\beta_k t},\ k = 1, 2, \ldots, p$, where β_k is assumed to have positive real part, and if its imaginary part is nonzero then its complement also occurs in (1), or
- **Model 2:** $\phi_k(\beta_k, t) = \cos(\beta_k t + \omega_k),\ k = 1, 2, \ldots, p$, where the β_k, and ω_k are real.

It is assumed that the observation points t_i are equispaced (so that $t_i = t_1 + (i-1)h$). However, *it is necessary to distinguish between the sampling strategies appropriate in the two cases.* In the first the condition $Rl(\beta_k) > 0$ ensures that the signal is transient so that measurements made for large t contain no information. Here we can assume that all observations are made in a fixed interval which can be chosen to be $0 \leq t \leq 1$, and in this case we have $t_1 = 0$ and $h = 1/(n-1)$. The importance of this point was first noted by Malinvaud [3]. In contrast, in the second case the signal is persistent so that observations made for arbitrarily large time contain information. Here it is appropriate to choose the scale of t so that $h = 1$. Also note that estimation of the $\alpha_i, i = 1, \ldots, p$ is not considered here for either model (but see [6], for example).

The restriction to equispaced data points is the basis for the common element in the algorithmic approaches to these two different classes of estimation problem because both models then satisfy difference equations with constant coefficients. In the first problem, where $\phi_k(\beta_k, t) = e^{-\beta_k t}$, there exist real coefficients $c_i(\boldsymbol{\beta})$, $i = 1, 2, \ldots, p+1$ such that

$$\sum_{j=1}^{p+1} c_j \phi_k(\beta_k, t_{i+j-1}) = 0,\ i = 1, 2 \ldots, n-p,\ k = 1, 2 \ldots, p. \tag{2}$$

Also, given $\boldsymbol{c}$, $\boldsymbol{\beta}$ is given by

$$\beta_i = -\frac{1}{h} \log(\lambda_i) \tag{3}$$

where $\lambda_i,\ i = 1, 2, \ldots, p$ are the roots of the equation

$$\sum_{i=1}^{p+1} c_i \lambda^{i-1} = 0. \tag{4}$$

The idea of reparametrizing the problem in terms of the recurrence parameters c_i of the difference equation (2) goes back to Prony [9]. This is not the only possibility of interest, and one important alternative [7], [8] proves to be the difference parameters, d_i, defined by

$$\sum_{j=1}^{p+1} d_j h^{-(p-j+1)} \Delta^{(p-j+1)} \phi_k(\beta_k, t_i) = 0,\ i = 1, 2, \ldots, n-p,\ k = 1, 2, \ldots, p \tag{5}$$

where Δ is the usual forward difference operator.

The frequency estimation problem, when $\phi_k(\beta_k, t) = \cos(\beta_k t + \omega_k)$, is slightly different in that the assumed model really is a linear combination of $2p$ complex

exponential terms, and the condition that the roots be pure imaginary means that (4) must factorize in the form

$$\eta_0 \prod_{m=1}^{p} (\lambda^2 - (2 - \eta_m^2)\lambda + 1), \tag{6}$$

and this implies that $c_k = c_{2p+2-k}$, $k = 1, 2, \ldots, 2p+1$, and that (6) has roots of the form $\lambda_k = e^{\pm i\beta_k}$ where $2 - \eta_k^2 = 2\cos(\beta_k)$. The corresponding difference form is

$$\eta_0 \left(\prod_{j=1}^{p} (\delta^2 + \eta_j^2) E \right) \phi_k(\beta_k, t_i) = 0,\ i = 1, 2, \ldots, n-2p,\ k = 1, 2, \ldots, p \tag{7}$$

where δ^2 is the second central difference operator and E is the forward shift operator.

Prony's method in its modern guise, see [2], is a naive application of least squares to estimate $\boldsymbol{c}$. Let

$$r_k = \sum_{i=1}^{p+1} c_i y_{k+i-1},\ k = 1, 2, \ldots, n-p. \tag{8}$$

Introducing matrices $X_c^T : R^n \to R^{n-p}$, and $Y_c : R^{p+1} \to R^{n-p}$ where

$$(X_c^T)_{ij} = c_{j-i+1},\ j = 1, \ldots, i+p, = 0,\ \text{otherwise} \tag{9}$$

$$(Y_c)_{ij} = y_{i+j-1},\ j = 1, 2, \ldots, p+1,\ i = 1, 2, \ldots, n-p, \tag{10}$$

then (8) can be written

$$\boldsymbol{r} = X_c^T \boldsymbol{y} = Y_c \boldsymbol{c}. \tag{11}$$

and the least squares problem is to minimize $\boldsymbol{r}^T\boldsymbol{r}$. However, as $\boldsymbol{r}$ is homogeneous in the components of $\boldsymbol{c}$, it is necessary to impose a scale - say

$$\psi(\boldsymbol{c}) = \frac{1}{2}. \tag{12}$$

Typical choices of scale include $\psi(\boldsymbol{c}) = \frac{1}{2}\boldsymbol{c}^T\boldsymbol{c}$, and $\psi(\boldsymbol{c}) = \boldsymbol{s}^T\boldsymbol{c}$ for some $\boldsymbol{s}$ prescribed. Prony's original choice was $\boldsymbol{s} = (1, 0, \ldots, 0)^T$. With this notation the estimation problem becomes

$$\min_{\boldsymbol{c}, \psi(\boldsymbol{c})=1/2} \boldsymbol{r}^T\boldsymbol{r}, \tag{13}$$

and the corresponding necessary conditions give the estimating equation

$$Y_c^T Y_c \boldsymbol{c} = \lambda \nabla\psi(\boldsymbol{c})^T \tag{14}$$

where λ is the Lagrange multiplier associated with the constraint (13).

Remark 1.1 There are equivalent definitions for X_d, and Y_d in the difference form for the basic recurrence. Note that X considered as an operator on $\boldsymbol{y} \to \boldsymbol{r}$ is independent of the parametrization except for the choice of scale because it is determined by the condition (2) that $0 = X^T\mathcal{E}\{\boldsymbol{y}\}$. For the difference formulation Y_d has the components

$$(Y_d)_{ij} = h^{-p+j-1}\Delta^{p-j+1} y_i,\ i = 1, 2, \ldots, n-p\ , j = 1, 2, \ldots, p+1. \tag{15}$$

In section 2 of this paper we summarise results that show that the conventional Prony's method does not give consistent estimates of the true parameter values $\boldsymbol{\beta}^*$

for model 1 for any choice of scale and that, for model 2, consistent estimates of $\boldsymbol{\beta}^*$ can be obtained for one form only of the constraint $\psi(\boldsymbol{c})$. In section 3 the problem with Prony's method is diagnosed as an uncritical use of least squares. Modifying the sum of squares turns the problem into a maximum likelihood one but also makes it nonlinear. Two approaches to its solution are considered, the Gradient condition Reweighting Algorithm (GRA) and the Objective function Reweighting Algorithm (ORA). In section 4 it is shown that, for the difference formulation of the modified Prony's method, (5), and for a simple model with one exponential, the ORA algorithm is consistent provided the coefficients of the difference equation satisfy the constraint $\psi(\boldsymbol{d}) = \frac{1}{2}d_1^2 = 1$. Section (5) considers the recurrence form parametrization of the modified Prony's method for estimating the exponential parameter in the same simple model and gives the form of the scaling of $\boldsymbol{c}$ that provides a consistent parameter estimate for the ORA algorithm. The final section of the paper outlines a study of the performance of both the GRA and ORA in numerical experiments with a simple model of the second type. Here the scaling does not prove to be as important.

2 Consistency of Prony's Method

We ask how well does $\widehat{\boldsymbol{\beta}}$ computed from $\widehat{\boldsymbol{c}}$, the solution to (14), approximate the true parameter values $\boldsymbol{\beta}^*$ for the particular model observed when the amount of data is large. The standard tool for answering this consistency question is the law of large numbers. We need here the form used in Osborne and Smyth [7] (it can be found, for example, in Stout [13] where precise conditions are given). To cope with the sampling of transient signals it is appropriate to introduce triangular arrays of random variables corresponding to realisations of the basic experiment for an increasing sequence of values of n. Let $\epsilon_{nj},\ j = 1, 2, \ldots, n,\ n = 1, 2, \ldots$ form such an array of iid random variables with $\mathcal{E}\{\,\epsilon_{nj}\} = \nu$, and $\mathcal{V}\{\,\epsilon_{nj}\} = \sigma^2$, and let uniformly bounded constants γ_{nj} be given. Two cases occur:

1. If $\lim_{n\to\infty} \gamma_{nj} \to \gamma(\frac{j}{n})$ as $n \to \infty$, where $\gamma(t)$ is continuous on $[0, 1]$, then with probability one
$$\frac{1}{n}\sum_{j=1}^{n} \gamma_{nj}\epsilon_{nj} \to \nu \int_0^1 \gamma(t)dt. \tag{16}$$

2. If $\lim_{n\to\infty} \gamma_{nj} \to \gamma_j$ for each fixed j as $n \to \infty$ then with probability one
$$\frac{1}{n}\sum_{j=1}^{n} \gamma_{nj}\epsilon_{nj} \to \nu \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \gamma_j. \tag{17}$$

The first case (16) can be used to give a limiting problem for (14). Writing (1) as

$$y_i = \mu(\boldsymbol{\beta}^*, t_i) + \epsilon_i, \tag{18}$$

and noting that (10) gives

$$(Y_c^T Y_c)_{ij} = \sum_{k=1}^{n-p} y_{k+i-1}y_{k+j-1},$$

then it follows almost surely that for the sampling regime appropriate for a transient signal and for fixed p

$$\frac{1}{n}Y_c^TY_c = \int_0^1 \mu(\boldsymbol{\beta}^*,t)^2dt\boldsymbol{e}\boldsymbol{e}^T + \sigma^2 I + o(1),\ n\to\infty, \tag{19}$$

where $\boldsymbol{e}^T = [1,1,\ldots,1]$. Note that

- the limiting contribution of the stochastic component (which here corresponds to the variance σ^2) is of the same order of magnitude as the contribution from the mean $\boldsymbol{\mu}$, and
- the limiting contribution from the mean is the integral term which gives almost no information on $\boldsymbol{\beta}^*$.

The inconsistency of Prony's method for this sampling regime is an immediate consequence of (19). It follows because the higher order contributions from the mean are truncation error terms in the estimate of the integral in (19) by $\sum_{k=1}^{n-p}\mu_{k+i-1}\mu_{k+j-1}$. These tend to zero like $O(n^{-1})$ while both the terms linear in the ϵ_{k+i-1} and the terms of the form $\epsilon_{k-i+1}\epsilon_{k-j+1}, i\neq j$ tend to zero like $n^{-1/2}N$ where N is normally distributed with finite variance. Thus the smaller stochastic terms have order in probability $O(n^{-1/2})$.

The second case (17) is used in carrying out the the corresponding calculation for the frequency estimation problem in which $h=1$. This gives the estimate valid almost surely for large enough n:

$$\begin{aligned}\frac{1}{n}Y_c^TY_c &= \frac{1}{2}\sum_{i=1}^{p}\alpha_i^2\begin{bmatrix} 1 & \cos(\beta_i^*) & \ldots & \cos(2p\beta_i^*)\\ \cos(-\beta_i^*) & 1 & \ldots & \cos((2p-1)\beta_i^*)\\ \vdots & \vdots & \ddots & \vdots\\ \cos(-2p\beta_i^*) & \cos(-(2p-1)\beta_i^*) & \ldots & 1\end{bmatrix}\\ &\quad +\sigma^2 I + o(1).\end{aligned} \tag{20}$$

Under the assumption of a correctly specified model, let $\boldsymbol{c}^*$ be the eigenvector of the mean contribution associated with the smallest (isolated) eigenvalue 0. It follows by applying standard perturbation theory to (20) that

$$\lim_{n\to\infty}\widehat{\boldsymbol{c}} = \boldsymbol{c}^* \tag{21}$$

provided the scale can be chosen so that the leading contribution from the stochastic part (the σ^2 term) can be cancelled in the necessary conditions (14) by the resulting Lagrange multiplier term. This requires

$$\sigma^2\boldsymbol{c}^* = \lim_{n\to\infty}\frac{\widehat{\lambda}}{n}\nabla\psi(\widehat{\boldsymbol{c}})^T, \tag{22}$$

and can be satisfied provided

$$\psi(\boldsymbol{c}) = \frac{1}{2}\boldsymbol{c}^T\boldsymbol{c}. \tag{23}$$

However, it does not appear that (21) would hold in general for other choices of scale. This specialised form of $\psi(\boldsymbol{c})$ is that appropriate to the Pisarenko form [2] of the algorithm, and leads to an eigenvalue problem for $\boldsymbol{c}$. It follows directly from (20), (22) that $\hat{\lambda}/n$ is a consistent estimator of σ^2 when the scale is chosen by (12), (23). Numerical results illustrating the above discussion are given in Tables 3 and 4 .

3 Maximum likelihood formulation

The problems with Prony's method stem from the uncritical use of least squares. Essentially, forming the difference equation has correlated the errors. Using (11) and noting that in the transient case the relation between $\boldsymbol{c}$ and $\boldsymbol{\beta}$ has a nontrivial dependence on h

$$\begin{aligned} \mathcal{E}\{\boldsymbol{r}(\boldsymbol{c}(\boldsymbol{\beta}^*))\boldsymbol{r}(\boldsymbol{c}(\boldsymbol{\beta}^*))^T\} &= X_c^T \mathcal{E}\{\,\boldsymbol{\epsilon}\,\boldsymbol{\epsilon}^T\}X_c \\ &= \sigma^2 X_c^T X_c \end{aligned} \tag{24}$$

so that the correct least squares formulation is

$$\min_{\boldsymbol{c},\psi(\boldsymbol{c})=1/2} \boldsymbol{c}^T Y_c^T (X_c^T X_c)^{-1} Y_c \boldsymbol{c}. \tag{25}$$

This appears to have been considered first for normal errors by Day and Osborne (Osborne [5]) who show that it corresponds to maximum likelihood estimation in this case. Osborne [6] gives an iterative reweighting algorithm for finding a point at which the necessary conditions for a minimum of (25) are satisfied (this will be referred to as algorithm GRA standing for Gradient condition Reweighting Algorithm). A simulation study is used to demonstrate that GRA can be very effective. Smyth in his thesis [12], and Osborne and Smyth [7], [8] show the consistency, scale independence, and asymptotic stability of the method under weak conditions on the probability distribution of the errors for the first sampling regime (transient signals). GRA is applied to the frequency estimation problem (the second sampling regime) in [3].

An alternative algorithm (ORA for Objective function Reweighting Algorithm) is suggested in [5] and also in Bresler and Macovski [1]. Here the reweighting is applied to the objective function (25) directly (that is in contrast to applying the reweighting in the statement of the necessary conditions for a minimum of (25) which is the key characteristic of GRA). Let

$$M_c(\boldsymbol{c}) = X_c^T X_c. \tag{26}$$

Then a step of ORA takes the form

$$\boldsymbol{c}_{k+1} = arg \min_{\boldsymbol{c},\psi(\boldsymbol{c})=1/2} \boldsymbol{c}^T Y_c^T M_c(\boldsymbol{c}_k)^{-1} Y_c \boldsymbol{c}. \tag{27}$$

The necessary conditions for this minimization are

$$Y_c^T M_c(\boldsymbol{c}_k)^{-1} Y_c \boldsymbol{c} = \lambda \nabla\psi(\boldsymbol{c})^T \tag{28}$$

where λ is the Lagrange multiplier associated with the scaling constraint. We will assume that

$$\nabla\psi(\boldsymbol{c})^T = \Phi_c \boldsymbol{c} \tag{29}$$

where Φ_c is positive (semi) definite. This choice has the advantage that efficient schemes based on inverse iteration are available to solve (28). For example, let

$$B_k = Y_c^T M(\boldsymbol{c}_k)^{-1} Y_c,$$

then one possible iterative scheme computes at the i'th step

$$\begin{aligned}
[B_k - \lambda_i \Phi_c]\, \boldsymbol{u}_{i+1} &= \Phi_c \boldsymbol{v}_i / \boldsymbol{s}_i^T \boldsymbol{v}_i, \\
[B_k - \lambda_i \Phi_c]\, \boldsymbol{v}_{i+1} &= \Phi_c \boldsymbol{u}_{i+1}, \\
\lambda_{i+1} &= \lambda_i - \frac{\boldsymbol{s}_{i+1}^T \boldsymbol{u}_{i+1}}{\boldsymbol{s}_{i+1}^T \boldsymbol{v}_{i+1}}
\end{aligned}$$

where $\boldsymbol{s}_i$ defines a scale for the process and typically might be chosen as $\Phi_c \boldsymbol{v}_i$. For any sensible choice the process is cubically convergent. The iteration is terminated when the change in λ relative to an estimate of σ^2 is very small (values of order $10^{-(15-\log(n))}$ have typically been used in our work).

It follows from (28) that the choice of scale is important in ORA. By comparison with the strict form of the necessary conditions used in GRA, the necessary conditions (28) omit the term obtained by differentiating $M_c(\boldsymbol{c})$ in calculating the gradient of (25). This leads to a simpler computational form for ORA. But it also means that ORA is not *a priori* computing the right quantity as it omits terms which are certainly not negligible.

In fact the convergence question for ORA is non trivial and will be discussed in a subsequent paper. Although experience in the exponential case has proved very satisfactory, ORA for rational fitting (for example the Michaelis-Menten equation [11]) is not acceptable. However, (28) is associated with a useful result. Let the scale be chosen such that

$$\boldsymbol{c}^T \Phi_c \boldsymbol{c} = 1\,. \tag{30}$$

Then it follows from (11), (28) that

$$\begin{aligned}
\widehat{\lambda} &= \widehat{\boldsymbol{c}}^T Y_c^T M_c^{-1} Y_c \widehat{\boldsymbol{c}} \\
&= \boldsymbol{y}^T P \boldsymbol{y}
\end{aligned} \tag{31}$$

where P is the $(n-p) \times (n-p)$ projection matrix

$$P = X_c M_c^{-1} X_c^T.$$

Let $(\widehat{\boldsymbol{c}}, \widehat{\lambda})$ solve (28), and let $\widehat{\boldsymbol{\beta}}$ be derived from $\widehat{\boldsymbol{c}}$ by (3), (4). If $\widehat{\boldsymbol{\beta}}$ is a consistent estimate of the true value $\boldsymbol{\beta}^*$ then $\widehat{\lambda}/n$ is a consistent estimate of σ^2. For the examples discussed in following sections this is a simple consequence of the estimates made; but its general validity is suggested strongly by the following argument. We show that if the constant vector $\widetilde{\boldsymbol{\beta}}$ is close to $\boldsymbol{\beta}^*$ (for example, equal to a member of a realisation of a sequence of consistent estimators of $\boldsymbol{\beta}^*$) then $\widetilde{\lambda}/n$, where $\widetilde{\lambda}$ is computed from (31), is close in probability to σ^2.

Let $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\beta}^*)$ be the vector of mean values (so $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$), and $\widetilde{\boldsymbol{\mu}} = \boldsymbol{\mu}(\widetilde{\boldsymbol{\beta}})$ so that

$$P\widetilde{\boldsymbol{\mu}} = 0.$$

Then

$$\begin{aligned}
\mathcal{E}\{\widetilde{\lambda}\} &= \mathcal{E}\{\boldsymbol{\epsilon}^T P \boldsymbol{\epsilon}\} + \boldsymbol{\mu}^T P \boldsymbol{\mu} \\
&= (n-p)\sigma^2 + n K_n \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|^2
\end{aligned} \tag{32}$$

where the norm is the Euclidean norm, and

$$K_n = O(1),\ n \to \infty.$$

Because P is a projection the estimate for K_n follows on noting that

$$\begin{aligned} \boldsymbol{\mu}^T P \boldsymbol{\mu} &= [\boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}}]^T P [\boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}}] \\ &\leq \|\boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}}\|^2 \\ &\sim n\|\boldsymbol{\beta}^* - \widetilde{\boldsymbol{\beta}}\|^2 \int_0^1 (\tfrac{d\mu}{ds})^2 dt \end{aligned} \tag{33}$$

where

$$\boldsymbol{s} = \frac{1}{\|\boldsymbol{\beta}^* - \widetilde{\boldsymbol{\beta}}\|}(\boldsymbol{\beta}^* - \widetilde{\boldsymbol{\beta}}).$$

To show that $\widetilde{\lambda}/n$ converges to its expectation in probability as $n \to \infty$ note that

$$\mathcal{V}\{\widetilde{\lambda}\} = \mathcal{E}\{(\boldsymbol{\epsilon}^T P \boldsymbol{\epsilon})^2\} + 4\boldsymbol{\mu}^T P \boldsymbol{\mu} \sigma^2 - (n-p)^2 \sigma^4, \tag{34}$$

and (using normality explicitly)

$$\begin{aligned} \mathcal{E}\{(\boldsymbol{\epsilon}^T P \boldsymbol{\epsilon})^2\} &= \sigma^4 \sum_{i,j,k,l} P_{ij} P_{kl} (\delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}), \\ &= \sigma^4 \left\{ (n-p)^2 + 2\|P\|_F^2 \right\}, \end{aligned} \tag{35}$$

where $\|.\|_F$ is the Frobenius norm. Now, as P is a symmetric projection

$$\|P\|_F^2 = n - p.$$

It follows that

$$\mathcal{V}\{\frac{\widetilde{\lambda}}{n}\} = O\left(\frac{\sigma^4}{n}\right). \tag{36}$$

An application of Chebyshev's inequality with fixed $\delta > 0$ gives

$$\mathcal{P}\{\frac{\widetilde{\lambda} - \mathcal{E}\{\widetilde{\lambda}\}}{n} > \delta\} \leq \frac{\mathcal{V}\{\widetilde{\lambda}/n\}}{\delta^2} \to 0,\ n \to \infty. \tag{37}$$

This verifies the convergence in probability and concludes the proof that $\hat{\lambda}/n$ is a consistent estimator of σ^2.

4 The consistency question for ORA (difference parameters)

In this section it is assumed that ORA applied to the modified Prony objective function in difference equation form converges. The consistency question for the resulting ORA estimates is addressed by considering in detail the simplest possible problem corresponding to

$$\mu(t) = \alpha_1 e^{-\beta_1^* t}. \tag{38}$$

For this problem the question of consistency can be addressed by elementary means. It is shown that there is a well defined scaling for which consistency can be demonstrated. Extension of results to more complex models requires a different approach, but preliminary calculations using the procedure employed by Osborne and Smyth [8] to analyse GRA suggest that the result that consistency holds for ORA only if the problem scale is chosen appropriately remains true. These predictions agree with the results of numerical simulations.

The necessary conditions for the ORA modification of Prony's method in its difference equation form solve the eigenvalue problem

$$Y_d^T M_d(\boldsymbol{d}_k)^{-1} Y_d \boldsymbol{d} = \lambda \Phi_d \boldsymbol{d}.$$

For one exponential (8) takes the form

$$d_1 h^{-1} \Delta y_i + d_2 y_i = r_i, \ i = 1, 2, \ldots, n-1 \tag{39}$$

so that

$$X_d^T = \begin{bmatrix} -d_1h^{-1} + d_2 & d_1h^{-1} & & & \\ & -d_1h^{-1} + d_2 & d_1h^{-1} & & \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ & & \cdots & -d_1h^{-1} + d_2 & d_1h^{-1} \end{bmatrix} \tag{40}$$

$$Y_d = \begin{bmatrix} h^{-1}\Delta y_1 & y_1 \\ h^{-1}\Delta y_2 & y_2 \\ \vdots & \vdots \\ h^{-1}\Delta y_{n-1} & y_{n-1} \end{bmatrix}, \tag{41}$$

and, setting $\tau = h^{-2}d_1^2 - h^{-1}d_1d_2$,

$$\begin{aligned} M_d &= X_d^T X_d \\ &= \begin{bmatrix} 2\tau + d_2^2 & -\tau & & & & \\ -\tau & 2\tau + d_2^2 & -\tau & & & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ & & & \cdots & -\tau & 2\tau + d_2^2 \end{bmatrix}. \end{aligned} \tag{42}$$

The matrix $Y_d^T M_d^{-1} Y_d$ evaluated at $\boldsymbol{d}(\boldsymbol{\beta}^*)$ is critical in determining if $\widehat{\boldsymbol{\beta}} \to \boldsymbol{\beta}^*$ as $n \to \infty$. To check this out we proceed as in the case of Prony's method and compute the limits as $n \to \infty$ of $\frac{1}{n} Y_d^T M_d^{-1} Y_d$. This separates into two parts:

1. the contribution of the mean term, and

2. the contributions of the stochastic terms arising from the error term in the model.

Writing out the matrix in detail gives

$$Y_d^T M_d^{-1} Y_d = \begin{bmatrix} (D(\boldsymbol{\mu} + \boldsymbol{\epsilon}))^T \\ (E_0(\boldsymbol{\mu} + \boldsymbol{\epsilon}))^T \end{bmatrix} M_d^{-1} \begin{bmatrix} D(\boldsymbol{\mu} + \boldsymbol{\epsilon}) & E_0(\boldsymbol{\mu} + \boldsymbol{\epsilon}) \end{bmatrix} \tag{43}$$

where $E_0 = \left[\begin{array}{cc} I & 0 \end{array} \right]$ and D is the matrix representation of the operator $h^{-1}\Delta$.The contribution from the mean is given by

$$\boldsymbol{\mu}^T E_0^T M_d^{-1} E_0 \, \boldsymbol{\mu} \left[\begin{array}{cc} h^{-2}(e^{\beta_1^* h} - 1)^2 & h^{-1}(e^{\beta_1^* h} - 1) \\ h^{-1}(e^{\beta_1^* h} - 1) & 1 \end{array} \right]. \tag{44}$$

Note that $\boldsymbol{d}(\boldsymbol{\beta}^*)^T + O(h) = \left[\begin{array}{cc} 1 & -\beta_1^* \end{array} \right]^T + O(h)$ is an eigenvector of (44) associated with the eigenvalue 0. As the terms in the matrix braces are each $O(1)$ as $n \to \infty$, the order of this term depends on the order of $\boldsymbol{\mu}^T E_0^T M_d^{-1} E_0 \, \boldsymbol{\mu}$ which is shown in the appendix to be $O(n)$. Thus

$$\lim_{n\to\infty} \frac{1}{n} \boldsymbol{\mu}^T E_0^T M_d^{-1} E_0 \boldsymbol{\mu} \left[\begin{array}{cc} h^{-2}(e^{\beta_1^* h} - 1)^2 & h^{-1}(e^{\beta_1^* h} - 1) \\ h^{-1}(e^{\beta_1^* h} - 1) & 1 \end{array} \right] \boldsymbol{d}(\boldsymbol{\beta}^*) = 0.$$

The important expectation in the stochastic contribution comes from terms which are quadratic in ϵ since terms such as $\boldsymbol{\mu}^T D^T M_d^{-1} D \boldsymbol{\epsilon}$ have zero expectation. Let

$$T = \left[\begin{array}{cc} \boldsymbol{\epsilon}^T D^T M_d^{-1} D \, \boldsymbol{\epsilon} & \boldsymbol{\epsilon}^T D^T M_d^{-1} E_0 \, \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon}^T E_0^T M_d^{-1} D \, \boldsymbol{\epsilon} & \boldsymbol{\epsilon}^T E_0^T M_d^{-1} E_0 \, \boldsymbol{\epsilon} \end{array} \right]. \tag{45}$$

Then it is shown in the appendix that

$$\lim_{n\to\infty} \frac{1}{n} T = \left[\begin{array}{cc} \sigma^2 & 0 \\ 0 & 0 \end{array} \right]$$

where the convergence is convergence in probability.

This brings us to much the same point as that reached in discussing the use of Prony's method to estimate frequencies in section 2. Again we have that the mean and stochastic components of the model contribute terms of similar magnitude. Thus, although the mean term has the correct limiting form, $\boldsymbol{d}(\boldsymbol{\beta}^*)$ can satisfy the limiting estimating equation to leading order terms only if the scale can be chosen so that

$$\Phi_d = \left[\begin{array}{cc} 1 & 0 \\ 0 & 0 \end{array} \right] + o(1) \tag{46}$$

In particular, this is satisfied provided

$$\psi(\boldsymbol{d}) = \frac{1}{2} d_1^2. \tag{47}$$

This scaling ensures that the difference equation (39) is necessarily non trivial. Choices of scaling for $\boldsymbol{d}$ that do not satisfy (46) will give estimates of $\boldsymbol{\beta}$ which are inconsistent.

The above results are illustrated by a small simulation which is reported in Table 1. In this, GRA and ORA for the consistent scale (47) are compared with ORA for the inconsistent scale $\psi(\boldsymbol{d}) = d_1^2 + d_2^2$. Data is generated using

$$y_i = \exp(-4\frac{i}{n}) + \epsilon_i, \; i = 1, 2, \ldots, n$$

where $\epsilon_i \sim N(0, 0.01)$. The first column indicates if the GRA or ORA procedure was used, and in the latter case with which scaling. Estimates of $\boldsymbol{\beta}$ are obtained from

100 simulated data sets and the table shows the means and standard deviations (in brackets and below) of three quantities obtained from each data set. These are the estimate of $\boldsymbol{\beta}$, the number of iterations taken (iter), and an estimate s of the standard deviation given by

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \mu(\widehat{\boldsymbol{\beta}}, t_i))^2}{n - 2p}. \tag{48}$$

This is used here for ORA rather than the formula based on (32), (36) in order to be able to make a direct comparison with GRA. The use of $\sigma = 0.1$ is large enough to make all methods work reasonably hard for the smaller values of n, and the inconsistent case essentially cannot cope. Certainly the predictions of our theory are born out. The interesting observation is that GRA and consistent ORA are strictly comparable in performance, so that the relative simplicity of ORA is likely to make it the method of choice. This comparability persists also for simulations we have carried out for two exponentials, but here numerical difficulties which arise due to the illconditioning of M for larger values of n become more evident.

5 The consistency question for ORA (recurrence parameters)

The recurrence form does not lend itself easily to the same techniques that proved successful for showing consistency for the difference form. The complication comes from the property that in the limit as $n \to \infty$ ($h \to 0$) the components of $\boldsymbol{c}$ tend to the coefficients of powers of E in $(E-1)^p$ which are signed multiples of the binomial coefficients and thus independent of $\boldsymbol{\beta}$. This makes it necessary to examine higher order terms. For this reason we study the equivalence between the two formulations discussed in Remark (1.1) and derive the results for the recurrence form from those derived for the difference form in the previous section. Let

$$T_c = \begin{bmatrix} -h^{-1} & 1 \\ h^{-1} & 0 \end{bmatrix}, \tag{49}$$

then, setting $E_1 = \begin{bmatrix} 0 & I \end{bmatrix}$,

$$\begin{aligned} Y_c\boldsymbol{c} &= [\ E_0\boldsymbol{y} \quad E_1\boldsymbol{y}\ ]\boldsymbol{c}, \\ &= Y_cT_cT_c^{-1}\boldsymbol{c}, \\ &= Y_d\boldsymbol{d}. \end{aligned}$$

Thus

$$Y_d = Y_cT_c,\ \boldsymbol{d} = T_c^{-1}\boldsymbol{c}. \tag{50}$$

Now, remembering that M is unchanged under the reparametrization, this permits the necessary conditions (28) to be put in the form (Remark 1.1)

$$T_c^TY_c^TM^{-1}Y_cT_cT_c^{-1}\boldsymbol{c} = \lambda T_c^T\Phi_cT_cT_c^{-1}\boldsymbol{c} \tag{51}$$

Number of data points	Algorithm and scale	$\hat{\beta}$	*iter*	s
$n = 10$	GRA	4.0835	4.390	0.097
		(1.0784)	(1.197)	(0.027)
	ORA	4.2241	3.300	0.098
	d_1^2	(1.1198)	(0.870)	(0.028)
	ORA	**	11.90	**
	$d_1^2 + d_2^2$	**	(7.31)	**
$n = 30$	GRA	4.0409	3.420	0.1010
		(0.4230)	(0.727)	(0.0135)
	ORA	4.0885	2.420	0.1010
	d_1^2	(0.4319)	(0.496)	(0.0135)
	ORA	**	14.00	**
	$d_1^2 + d_2^2$	**	(7.15)	**
$n = 100$	GRA	4.0044	3.040	0.1003
		(0.2360)	(0.567)	(0.0066)
	ORA	4.0181	2.0	0.1003
	d_1^2	(0.2375)	(0.0)	(0.0066)
	ORA	6.0134	13.07	0.1219
	$d_1^2 + d_2^2$	(0.6166)	(2.79)	(0.0127)
$n = 300$	GRA	4.0236	2.490	0.0999
		(0.1224)	(0.577)	(0.0037)
	ORA	4.0281	2.0	0.0999
	d_1^2	(0.1227)	(0.0)	(0.0037)
	ORA	5.9229	12.62	0.1204
	$d_1^2 + d_2^2$	(0.3204)	(1.41)	(0.0068)

Table 1: Simulation results for the model $y_i = \exp(-4i/n) + \epsilon_i, i = 1, \ldots, n$. Results are given for the GRA and ORA with consistent and inconsistent scaling for the difference parameters. ** means failure of the algorithm.

Comparing with (50) shows that identical results are obtained in the two parametrizations provided the scale transforms according to

$$\Phi_d = T_c^T \Phi_c T_c. \tag{52}$$

To see what this means consider the scale defined by

$$\Phi_c = \begin{bmatrix} \xi_1 & \\ & \xi_2 \end{bmatrix},\ \xi_1,\ \xi_2 \geq 0 \tag{53}$$

Multiplying out gives

$$\Phi_d = h^{-2} \left\{ \begin{bmatrix} \xi_1 + \xi_2 & 0 \\ 0 & 0 \end{bmatrix} + O(h) \right\}.$$

Comparing this with (46) shows that any nonzero scaling matrix of the form (53) is consistent. Examples of consistent scalings include

$$\psi(\boldsymbol{c}) = c_1^2,\ c_1^2 + c_2^2,\ \text{and } \mathrm{c}_2^2.$$

These would appear to include all cases of interest except the original Prony scaling which is not of this form. Another example of an inconsistent scaling is given by

$$\psi(\boldsymbol{c}) = (c_1 + c_2)^2.$$

For the example considered in the previous section, $y_i = \exp(-4i/n) + \epsilon_i$, the corresponding results in this case are given in Table 2. The results follow the same pattern as before showing the same sharp distinction between consistent and inconsistent scalings. The results for the two cases $\psi(\boldsymbol{d}) = d_1^2$, and $\psi(\boldsymbol{c}) = c_2^2$ are identical confirming (52).

6 The question of consistency in the frequency estimation case

The above calculations suggest that the effectiveness of ORA for the estimation of rate constants associated with transient signals can be quite adequate provided the correct choice of scale is made. However, the frequency case is associated with results of a rather different kind in maximum likelihood calculations, and these suggest parallel questions for ORA. First, the global maximum of the likelihood is known to give $O(n^{-3/2})$ accurate estimates of the frequency parameters, and this super convergence result contrasts with the $O(n^{-1/2})$ result usual in such calculations. The catch is that the likelihood surface in its original parametrization has many local minima (the number increasing with n) which are all within $O(n^{-1})$ of the correct frequency estimate, and that the corresponding amplitude estimates are inconsistent [10]. Second [3], the re-parametrization in GRA does not save it from the difficulties caused by multiple stationary points.

More elaborate calculations are involved if an attempt is made to analyse ORA in the frequency case following our previous procedures as the simplest example already

Number of data points	Algorithm and scale	$\hat{\beta}$	*iter*	s
$n = 10$	ORA	4.2241	4.430	0.0978
	c_2^2	(1.1198)	(1.103)	(0.0275)
	ORA	3.9999	3.920	0.0977
	$c_1^2 + c_2^2$	(1.0362)	(1.061)	(0.0275)
	ORA	**	12.58	**
	$(c_1 + c_2)^2$	**	(8.11)	**
$n = 30$	ORA	4.0885	3.190	.1010
	c_2^2	(0.4319)	(0.563)	(0.0135)
	ORA	4.0023	3.020	0.1010
	$c_1^2 + c_2^2$	(0.4172)	(0.681)	(0.0135)
	ORA	**	14.79	**
	$(c_1 + c_2)^2$	**	(7.52)	**
$n = 100$	ORA	4.0181	2.570	.1003
	c_2^2	(0.2375)	(0.497)	(0.0066)
	ORA	3.9919	2.430	0.1003
	$c_1^2 + c_2^2$	(0.2347)	(0.498)	(0.0066)
	ORA	6.1642	13.87	0.1243
	$(c_1 + c_2)^2$	(0.6517)	(3.11)	(0.0137)
$n = 300$	ORA	4.0281	2.0	0.0999
	c_2^2	(0.1227)	(0.0)	(0.0037)
	ORA	4.0194	1.970	0.0999
	$c_1^2 + c_2^2$	(0.1222)	(0.300)	(0.0037)
	ORA	6.0564	13.29	0.1225
	$(c_1 + c_2)^2$	(0.3361)	(1.50)	(0.0771)

Table 2: Simulation results for the model $y_i = \exp(-4i/n) + \epsilon_i, i = 1, \ldots, n$. Results are given for the ORA with consistent and inconsistent scaling for the recurrence parameters. ** means failure of the algorithm.

involves two exponential terms. For this reason we have resorted to a numerical approach. We consider the signal

$$\mu(t) = \cos(.01t + 1),$$

and data is generated using

$$y_i = \mu(i) + \epsilon_i,\ i = 1, 2, \ldots, n$$

where, as before, $\epsilon_i \sim N(0, .01)$. Again the data reported are the means and standard deviations of the results of 100 replications for each of $n = 10, 30, 100, 300$, and follows the same basic format as before, but the estimates are sorted in increasing value of their imaginary part before averaging. Also, this time the results for Prony's method are quoted as well. The inconsistency of Prony's method for other than the Pisarenko scaling is clear, but the performance of the other methods is interesting. No special assumptions were made about the Prony parameters so it was not assumed in advance that the rate constants were pure imaginary. The calculations were started at the true parameter values in each case, and this proves close enough to the actual computed estimates for GRA to avoid the convergence problems resulting from multiple stationary points outlined above. Calculations have been carried out both for the recurrence and difference form of ORA, but the results are very similar so that only the recurrence results are reported. ORA encounters some trouble when $n = 10$, and it was necessary to remove redundancy in parametrization by requiring that $Im\,\beta \in [-\pi, \pi]$. When more than 40 iterations occured in any of the experiments the computation is terminated and the corresponding results are starred. Only the Prony results are given for $n = 300$ as the errors in the ORA calculation of the rate constants proves to be in the sixth decimal place. Also values of $(\widehat{\lambda}/n)^{1/2}$ for the ORA and Prony calculations are reported. These show a tendency to underestimate σ when the estimation procedure is consistent. The results are given in Tables 3 and 4. This shows two very interesting aspects of our calculations:

1. Here the choice of scale needed to ensure consistency for ORA is clearly much less important in both the recurrence and difference forms. This could be compared with the discussion of Prony's method where nothing could be done in the transient case, but there proved to be a consistent scaling in the frequency case.

2. The ORA results exhibit the same super convergence results as GRA.

One point which does not show up here could prove to be at least as important. Other testing that has been carried out suggests that ORA appears much less sensitive than does GRA to the choice of starting values for large n. The multiple stationary values of GRA can be explained by noting that the gradient of the objective function (25) has poles at the zero eigenvalues of X^TX, and there are lots of these in the frequency case. On the other hand, in ORA, forming the necessary conditions from the objective function does not involve the difficult term $(X^TX)^{-1}$. This permits the small size of the $X\boldsymbol{c}$ to be more important in minimizing the quadratic form, providing a kind of smoothing mechanism.

Algorithm and scale	$\hat{\beta}_1$	$\hat{\beta}_2$	*iter*	$\sqrt{\widehat{\lambda}/n}$
$n = 10$				
GRA	$-0.1028 - 0.2237i$	$0.0934 + 0.2237i$	4.78	
	(0.2970)	(0.2887)	(1.1063)	
Prony	$-0.2067 - 0.3212i$	$0.1815 + 0.2270i$		0.1515
$c_1^2 + c_2^2 + c_3^2$	0.9804	0.3481		0.0639
ORA (rec)	$-0.1045 - 0.1849i$	$0.1065 + 0.1849i$	3.66	0.1011
$c_1^2 + c_2^2 + c_3^2$	(0.2373)	(0.2305)	(0.9663)	(0.0300)
Prony	$0.8928 - 1.2591i$	$0.8711 + 0.0967i$		0.1402
c_3^2	(1.7370)	(0.9207)		(0.0295)
ORA (rec)	$-0.0516 - 0.2242i$	$0.0901 + 0.2242i$	4.87	0.0800
c_3^2	(0.2106)	(0.2190)	(1.0508)	(0.0246)
Prony	$-0.0017 - 1.9758i$	$0.2749 + 1.0019i$		0.2010
$(c_1 + c_2 + c_3)^2$	(1.1596)	(1.9723)		(0.0258)
ORA (rec)	$0.0361 - 1.5879i$	$0.2102 + 0.6454i$	24.45 *	0.2040
$(c_1 + c_2 + c_3)^2$	(1.1220)	(1.8136)	(11.005)	(0.0714)
$n = 30$				
GRA	$0.0008 - 0.0992i$	$0.0008 + 0.0992i$	2.64	
	(0.0118)	(0.0118)	(0.523)	
Prony	$-0.0489 - 0.0980i$	$0.0426 + 0.0665i$		0.0953
$c_1^2 + c_2^2 + c_3^2$	(0.4252)	(0.1628)		(0.0164)
ORA (rec)	$0.0006 - 0.0994i$	$0.0006 + 0.0994i$	2.37	0.0933
$c_1^2 + c_2^2 + c_3^2$	(0.0117)	(0.0117)	(0.501)	(0.0118)
Prony	$1.4076 - 2.5761i$	0.5852		0.1471
c_3^2	(1.5714)	(1.3104)		(0.0142)
ORA (rec)	$0.0028 - 0.0980i$	$0.0028 + 0.0980i$	2.92	0.0933
c_3^2	(0.0122)	(0.0122)	(0.393)	(0.0018)
Prony	$-0.7064 - 2.0001i$	$-1.0246 + 1.12147i$		0.6965
$(c_1 + c_2 + c_3)^2$	(0.7537)	(1.7296)		(0.0162)
ORA (rec)	$0.0271 - 0.1174i$	$-0.0271 + .1174i$	5.42	0.1142
$(c_1 + c_2 + c_3)^2$	(0.0105)	(0.0105)	(0.843)	(0.0204)

Table 3: Simulation results for frequency estimation, $n = 10$ and $n = 30$. The true values of β_1 and β_2 are $\pm.1i$ and the final column should estimate $\sigma = .1$. Prony's method is inconsistent except for the scaling $c_1^2 + c_2^2 + c_3^2$. ORA is less sensitive to scaling for this frequency problem. * implies that the algorithm failed.

Algorithm and scale	$\hat{\beta}_1$	$\hat{\beta}_2$	*iter*	$\sqrt{\widehat{\lambda}/n}$
$n = 100$				
GRA	$0.0000 - 0.1000i$	$0.0000 + 0.1000i$	2.08	
	(0.0007)	(0.0007)	(0.394)	
Prony	$-0.0085 - 0.0907i$	$-0.0019 + 0.0907i$		0.0989
$c_1^2 + c_2^2 + c_3^2$	(0.0526)	(0.0347)		(0.0100)
ORA (rec)	$0.0000 - 0.1000i$	$0.0000 + 0.1000i$	1.96	0.0980
$c_1^2 + c_2^2 + c_3^2$	(0.0007)	(0.0007)	(0.197)	(0.0066)
Prony	$1.7969 - 3.0788i$	0.1172		0.1528
c_3^2	(0.7838)	(0.7427)		(0.0081)
ORA (rec)	$0.0000 - 0.1000i$	$0.0000 + 0.1000i$	2.37	0.0980
c_3^2	(0.0007)	(0.0007)	(0.506)	(0.0066)
Prony	$0.0993 - 1.7775i$	$0.0993 + 1.7775i$		0.6866
$(c_1 + c_2 + c_3)^2$	(0.1063)	(0.1063)		(0.0099)
ORA (rec)	$-0.0000 - 0.1004i$	$-0.0000 + 0.1004i$	2.89	0.0984
$(c_1 + c_2 + c_3)^2$	(0.0007)	(0.0007)	(0.315)	(0.0067)
$n = 300$				
Prony	$0.0005 - 0.0993i$	$0.0005 + 0.0993i$		0.1001
$c_1^2 + c_2^2 + c_3^2$	(0.0109)	(0.0109)		(0.0054)
Prony	$1.6933 - 3.1416i$	0.0152		0.1545
c_3^2	(0.2709)	(0.0012)		(0.0048)
Prony	$-0.1284 - 1.7837i$	$-0.1284 + 1.7837i$		0.6943
$(c_1 + c_2 + c_3)^2$	(0.0587)	(0.0587)		(0.0058)

Table 4: Simulation results for frequency estimation, $n = 100$ and $n = 300$. The true values of β_1 and β_2 are $\pm.1i$ and the final column should estimate $\sigma = .1$. GRA and ORA give correct answers for $n = 300$. Prony's method is inconsistent except for the scaling $c_1^2 + c_2^2 + c_3^2$.

7 Conclusion

The consistency of estimates given by the Prony, GRA, and ORA algorithms as $n \to \infty$ has been discussed in the context of the simplest possible applications. Not much can be done for Prony in the transient case, but a distinguished choice of scale (corresponding to Pisarenko's method) gives consistency in the frequency estimation problem. In the transient case ORA requires a particular choice of scale to be consistent for the difference parameters. But this permits us to deduce a useful family of consistent scales for the recurrence parameters. Here GRA and ORA appear to be strictly comparable methods. In an interesting contrast to the transient case, it appears to be much easier for ORA to be consistent in the frequency estimation problem. But more importantly, it appears to be much less sensitive to choice of starting value than does GRA while still producing super convergent estimates of the frequency parameters. Thus ORA would appear to be the superior method in this case.

References

[1] Y. Bresler and A. Macovski, *Exact maximum likelihood parameter estimation of superposed exponential signals in noise*, IEEE Trans. Acoust., Speech, Signal Processing **34** (1986), 1081-1089.

[2] S.M. Kay and S.L. Marple, *Spectrum analysis - a modern perspective*, Proc. IEEE, **69** (1981), 1380-1419.

[3] M.S. Mackisack, M.R. Osborne, and G.K. Smyth, *A modified Prony algorithm for estimating sinusoidal frequencies*, submitted for publication.

[4] E. Malinvaud, *The consistency of nonlinear regression*, Ann. Math. Stat. **41** (1970), 956-969.

[5] M.R. Osborne, *A class of non-linear regression models*, Data Representation, edited by R.S. Anderssen and M.R. Osborne, University of Queensland Press, (1970),pp 94-101.

[6] M.R. Osborne, *Some special nonlinear least squares problems*, SINUM **12** (1975), 571-592.

[7] M.R. Osborne and G.K. Smyth, *A modified Prony algorithm for fitting functions defined by difference equations*, SISSC **12** (1991), 362-382.

[8] M.R. Osborne and G.K. Smyth, *A modified Prony algorithm for exponential fitting*, (accepted for publication).

[9] R. Prony, *Essai éxperimental et analytique: sur les lois de la dilatabilité de fluides élastique et sur celles de la force expansive de la vapeur de l'alkool, à différentes températures*, J. de l'École Polytechnique **1** (1795), 24-76.

[10] J.A. Rice and M. Rosenblatt, *On frequency estimation*, Biometrika **75** (1988), 477-84.

[11] D.V. Roberts, *Enzyme Kinetics*, Cambridge Chemistry Texts, Cambridge University Press, 1977.

[12] G.K. Smyth, *Coupled and Separable Iterations in Nonlinear Estimation*, Ph.D thesis, Australian National University, 1987.

[13] W.F. Stout, *Almost Sure Convergence*, Academic Press, 1974.

8 Appendix

The first result to be shown here is that the term $\boldsymbol{\mu}^T E_0^T M_d^{-1} E_0 \boldsymbol{\mu}$ in equation (44) is $O(n)$ as $n \to \infty$. To do this we use the eigen decomposition of M_d. We have

$$M_d \boldsymbol{v}_i = \upsilon_i \boldsymbol{v}_i \ \ (MV = V\Upsilon)$$

where

$$\upsilon_i = 4\tau \sin(\frac{i\pi}{2n})^2 + d_2^2, \ i = 1, 2, \ldots, n-1, \tag{54}$$

$$(\boldsymbol{v}_i)_j = \sqrt{\frac{2}{n}} \sin(\frac{i\pi j}{n}), \ j = 1, 2, \ldots, n-1, \ i = 1, 2, \ldots, n-1. \tag{55}$$

It is known that $\upsilon_i - d_2^2 = d_1^2 \gamma_i i^2$ where $1 < \gamma_i < \pi^2$ so that, in particular, $\sum_{i=1}^{n-1} \frac{1}{\upsilon_i} < 2$, $\forall n$. We will also need the singular value decomposition for D the matrix representation of the operation $h^{-1}\Delta$. This is

$$D\boldsymbol{u}_i = \lambda_i \boldsymbol{v}_i \ \left(D^T = U\Lambda V^T\right)$$

where

$$\lambda_i = 2h^{-1} \sin(\frac{i\pi}{2n}), \ i = 1, 2, \ldots, n-1, \tag{56}$$

$$(\boldsymbol{u}_i)_j = \sqrt{\frac{2}{n}} \cos(\frac{i\pi}{n}(j - \frac{1}{2})), \ j = 1, 2, \ldots, n, \ i = 1, 2, \ldots, (n-1). \tag{57}$$

We use the eigen decomposition to obtain

$$\begin{aligned} \boldsymbol{\mu}^T E_0^T M_d^{-1} E_0 \, \boldsymbol{\mu} &= \sum_{i=1}^{n-1} \frac{1}{\upsilon_i} \left\{ \sum_{j=1}^{n-1} \sqrt{\frac{2}{n}} \sin(\frac{i\pi j}{n}) e^{\frac{-\beta_1^* j}{n}} \right\}^2 \\ &= \frac{2}{n} \sum_{i=1}^{n-1} \frac{1}{\upsilon_i} I_i^2 \end{aligned} \tag{58}$$

where

$$I_i = \sum_{j=1}^{n-1} \sin(\frac{i\pi j}{n}) e^{\frac{-\beta_1^* j}{n}}$$

is essentially a decreasing function of i. As

$$I_k \approx n \int_0^1 \sin(k\pi t) e^{-\beta_1^* t} dt = O(n)$$

for small k, and

$$\sum_{i=1}^{n-1} \frac{1}{\upsilon_i} = O(1) \; \forall \, n.$$

Thus it follows that

$$\boldsymbol{\mu}^T E_0^T M_d^{-1} E_0 \, \boldsymbol{\mu} = O(n). \tag{59}$$

The second result to be proved here is that the expectation of the stochastic matrix

$$T = \begin{bmatrix} \boldsymbol{\epsilon}^T D^T M_d^{-1} D \boldsymbol{\epsilon} & \boldsymbol{\epsilon}^T D^T M_d^{-1} E_0 \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon}^T E_0^T M_d^{-1} D \boldsymbol{\epsilon} & \boldsymbol{\epsilon} E_0^T M_d^{-1} E_0 \boldsymbol{\epsilon} \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}$$

tends to the limiting form $\begin{bmatrix} O(n) & O(1) \\ O(1) & O(1) \end{bmatrix}$ as $n \to \infty$. To show this it is convenient to define two extra variables as follows,

$$\boldsymbol{\epsilon}^{\#} = U^T \, \boldsymbol{\epsilon} \tag{60}$$

and

$$\boldsymbol{\epsilon}^{*} = V^T E_0 \, \boldsymbol{\epsilon}, \tag{61}$$

then $\epsilon_i^{\#}, \;\; i = 1, 2, \ldots, (n-1)$, and $\epsilon_i^{*}, \;\; i = 1, 2, \ldots, (n-1)$ are independent and $\sim N(0, \sigma^2)$. To calculate the expectations we have:

1. $\mathcal{E}\{T_{22}\}$:

$$\begin{aligned} \mathcal{E}\{T_{22}\} &= \operatorname{trace} \mathcal{E}\{\mathrm{M_d^{-1} E_0} \, \boldsymbol{\epsilon} \, \boldsymbol{\epsilon}^{\mathrm{T}} \mathrm{E_0^T}\}, \\ &= \operatorname{trace} \mathcal{E}\{\Upsilon^{-1} \, \boldsymbol{\epsilon}^{*} \, \boldsymbol{\epsilon}^{*\mathrm{T}}\}, \\ &= \sigma^2 \sum_{i=1}^{n-1} \frac{1}{\upsilon_i}, \\ &= O(1). \end{aligned} \tag{62}$$

2. $\mathcal{E}\{T_{21}\}$:

$$\begin{aligned} \mathcal{E}\{T_{21}\} &= \operatorname{trace} \mathcal{E}\{\mathrm{M_d^{-1} D} \, \boldsymbol{\epsilon} \, \boldsymbol{\epsilon}^{\mathrm{T}} \mathrm{E_0^T}\}, \\ &= \operatorname{trace} \mathcal{E}\{\Upsilon^{-1} \Lambda \, \boldsymbol{\epsilon}^{\#} \, \boldsymbol{\epsilon}^{*\mathrm{T}}\}, \\ &= \sigma^2 \frac{2}{n} \sum_{i=1}^{n-1} \frac{\lambda_i}{\upsilon_i} \sum_{j=1}^{n-1} \sin(\frac{i\pi j}{n}) \cos(\frac{i\pi}{n}(j - \frac{1}{2})), \\ &= \frac{-\sigma^2}{2n} \sum_{i=1}^{n-1} \frac{\lambda_i^2}{\upsilon_i}, \\ &= O(1). \end{aligned} \tag{63}$$

3. $\mathcal{E}\{T_{11}\}$

$$\begin{aligned} \mathcal{E}\{T_{11}\} &= \operatorname{trace} \mathcal{E}\left\{\mathrm{M_d^{-1} D} \, \boldsymbol{\epsilon} \, \boldsymbol{\epsilon}^{\mathrm{T}} \mathrm{D^T}\right\}, \\ &= \operatorname{trace} \mathcal{E}\{\Upsilon^{-1} \Lambda \, \boldsymbol{\epsilon}^{\#} \, \boldsymbol{\epsilon}^{\#\mathrm{T}} \Lambda\}, \\ &= \sigma^2 \sum_{i=1}^{n-1} \frac{\lambda_i^2}{\upsilon_i}, \\ &= O(n) \end{aligned} \tag{64}$$

as it follows from (54), (56) that all the terms in the summation are $O(1)$. It also follows from section 3 that $\lim_{n\to\infty} \frac{1}{n}\mathcal{E}\{T_{11}\}$ must be σ^2.

We can show the convergence as $n \to \infty$ of $\frac{1}{n}T$ to its expectation readily. The pattern of the argument is illustrated by T_{11}.

$$\begin{aligned} T_{11} &= \sum_{i=1}^{n-1} (\epsilon_i^{\#})^2 \frac{\lambda_i^2}{\upsilon_i} \\ &= \mathcal{E}\{T_{11}\} + \sum_{i=1}^{n-1} \frac{\lambda_i^2}{\upsilon_i} \left\{ (\epsilon_i^{\#})^2 - \sigma^2 \right\} \\ &\to n\left(\frac{\mathcal{E}\{T_{11}\}}{n} + o(1) \right) \ a.s.,\ n \to \infty \end{aligned}$$

by the strong law of large numbers (the second case (17) is relevant here).

The remaining stochastic terms become negligible in comparison with the mean term as $n \to \infty$. We consider the term

$$\begin{aligned} Z &= \boldsymbol{\epsilon}^T E_0^T M^{-1} E_0 \, \boldsymbol{\mu} \\ &= \boldsymbol{\epsilon}^{*T} \Upsilon^{-1} V^T E_0 \, \boldsymbol{\mu} \\ &= \sqrt{\frac{2}{n}} \sum_{i=1}^{n-1} \frac{1}{\upsilon_i} \, \boldsymbol{\epsilon}_i^* I_i \end{aligned}$$

(terms involving $D\,\boldsymbol{\mu}$ add nothing new). The law of large numbers doesn't help here (Z/n is a sum of terms of the form $O(\frac{1}{\sqrt{n}i^2})\,\boldsymbol{\epsilon}_i^*$); but the Chebyshev inequality permits a simple estimate in probability. The relevant variance calculations give

$$\begin{aligned} \mathcal{V}\{Z\} &= \mathcal{V}\{\, \boldsymbol{\epsilon}^T E_0^T M^{-1} E_0 \, \boldsymbol{\mu}\} \\ &= \mathcal{E}\{\, \boldsymbol{\mu}^T E_0^T V \Upsilon^{-1} \, \boldsymbol{\epsilon}^* \, \boldsymbol{\epsilon}^{*T} \Upsilon^{-1} V^T E_0 \, \boldsymbol{\mu}\}, \\ &= \sigma^2 \sum_{i=1}^{n-1} (\frac{\boldsymbol{v}_i^T E_0 \, \boldsymbol{\mu}}{\upsilon_i})^2, \\ &= \sigma^2 O(n); \end{aligned}$$

Then for $\delta > 0$, fixed,

$$\mathcal{P}\{\frac{Z}{n} > \delta\} \leq \frac{\mathcal{V}\{Z/n\}}{\delta^2} \to 0,\ n \to \infty.$$