Information Loss with Transform Methods in System Identification: A New Set of Transforms with High Information Content

STEPHEN W. PROVENCHER AND ROBERT H. VOGEL
European Molecular Biology Laboratory, Postfach 10.2209, D-6900 Heidelberg, Federal Republic of Germany

Received 15 October 1979; revised 12 February 1980

ABSTRACT

Linear transform methods like moments, modulating functions, and Laplace transforms are widely used for parameter estimation in system identification problems because they can reduce a large set of overdetermined equations to a small set of linear and nonlinear equations, which often have a very simple form and a unique solution. However, the effects of noise in the data are neglected in deriving these equations. We show (in terms of Fisher's information measure, the generalized variance, and simulations) that these methods can lead to very large errors in the estimates. We develop a new set of transforms based on the idea of maximizing their Fisher information content. The robustness of these new transforms, in contrast to the others, is illustrated by simulations of nanosecond fluorescence decay and multicomponent exponential decay.

1. INTRODUCTION

Many system identification problems in the biosciences are reduced at some stage to parameter estimation problems by postulating a model of the form

$$\hat{y}(t,\boldsymbol{\theta}) = \sum_{j=1}^{N_\lambda} \alpha_j f(\lambda_j, t), \tag{1.1}$$

where $f(\lambda,t)$ is specified and the $2N_\lambda \times 1$ vector $\boldsymbol{\theta}$ has the α_j and λ_j parameters as components. The problem is then to estimate the actual parameter values $\bar{\boldsymbol{\theta}}$ from the noisy data

$$y_k = \hat{y}\left(t_k, \bar{\boldsymbol{\theta}}\right) + \varepsilon_k, \qquad k = 1, \ldots, N_y, \tag{1.2}$$

or, in terms of $N_y \times 1$ vectors, $\mathbf{y} = \hat{\mathbf{y}}(\bar{\boldsymbol{\theta}}) + \boldsymbol{\varepsilon}$. For simplicity, we assume throughout that the noise, $\boldsymbol{\varepsilon}$, has a normal distribution with zero means and with covariance matrix M_y known to within a scale factor.

52 Vanderbilt Ave., New York, NY 10017 0025-5564/80/060251+12$02.25

The dynamic response of a linear system to a sudden perturbation, for example, can often be described by (1.1) with $f(\lambda,t)=e^{-\lambda t}$. If the impulse response of the measuring instrument or the rise or decay time of the perturbation is comparable to the response time of the system, then $f(\lambda,t)$ must be written as a convolution with the impulse response or the perturbation, as in nanosecond fluorescence decay [1–8].

A weighted least-squares analysis of (1.2) yields the maximum-likelihood estimates for the parameters $\bar{\boldsymbol{\theta}}$, and this would therefore seem to be the best procedure to use. However, such a nonlinear regression involves the computation of $\hat{\mathbf{y}}(\boldsymbol{\theta})$ and its partial derivatives at each iteration. When $N_y \gg 1$ and $f(\lambda,t)$ is a convolution, this can be computationally expensive, especially since the analysis should be repeated from many starting points in parameter space to have a good chance of finding the global optimum and not just a local one.

Partly because of this, there has been great interest in linear transform methods that apply a linear operator to reduce the data to $2N_\lambda$ points:

$$\eta_i = \sum_{k=1}^{N_y} A_{ik} y_k, \qquad i=1,\ldots,2N_\lambda, \tag{1.3}$$

or $\boldsymbol{\eta}=A\mathbf{y}$, where A is a $2N_\lambda \times N_y$ matrix. Often A is originally an integral operator, but the numerical quadrature over the discrete data takes the form of (1.3). The corresponding reduced model and data can be written $\hat{\boldsymbol{\eta}}(\boldsymbol{\theta})=A\hat{\mathbf{y}}(\boldsymbol{\theta})$ and $\boldsymbol{\eta}=A\hat{\mathbf{y}}(\bar{\boldsymbol{\theta}})+A\boldsymbol{\varepsilon}$. Neglecting the noise and fitting the reduced data to the model yields

$$\boldsymbol{\eta}=A\hat{\mathbf{y}}(\boldsymbol{\theta}), \tag{1.4}$$

$2N_\lambda$ nonlinear equations in the $2N_\lambda$ parameters $\boldsymbol{\theta}$. In special cases, (1.4) can have a very simple form and a unique solution (ignoring the sensitivity of this solution to the noise). Examples are the methods of moments [1,2], modulating functions [3], and Laplace transforms [4] when $f(\lambda,t)$ is a convoluted or simple exponential.

Hartley [9] has shown that in general no sufficient statistics for parameters associated with nonlinear functions exist; i.e., information is lost in going from $\mathbf{y}$ to $\boldsymbol{\eta}$, and accuracy is lost in solving (1.4) instead of (1.2). In Sec. 2 we quantify this information loss in terms of Fisher's information measure [10, p. 41] and the generalized variance [11, p. 27], which are closely related to the uncertainties in the parameter estimates. We show (in Sec. 5) that the use of the above transforms can lead to a disastrous loss of information and accuracy. In Secs. 3 and 4, we develop a set of transforms, called "information transforms," that are based on the idea of maximizing their Fisher information content. In Sec. 5 we illustrate with numerical tests the effectiveness of these transforms.

In Sec. 4 we show how the data reduction in solving (1.4) with information transforms rather than (1.2) permits a more thorough search for a global optimum. The solution of (1.4) is then used as a starting estimate of $\bar{\boldsymbol{\theta}}$ in a full least-squares solution of (1.2). The convergence is usually very rapid because the starting estimate is usually very good. Since least-squares procedures are used throughout, parameter constraints and extra corrections for such effects as scattered light in fluorescence decay are straightforward to include [5, 6]. A user-oriented FORTRAN IV program implementing the method will be available on request.

Often N_λ is also unknown and it is also necessary to estimate the minimum N_λ that is consistent with the data. We will not discuss this problem here, but most methods require the value of the likelihood function for each alternative N_λ; e.g., see [12–14]. Therefore, the estimation of $\bar{\boldsymbol{\theta}}$ is a prerequisite to estimating N_λ.

Another approach to estimating N_λ, as well as $\bar{\boldsymbol{\theta}}$, is to generalize the model to

$$\hat{y}(t) = \int \alpha(\lambda) f(\lambda, t)\, d\lambda \tag{1.5}$$

and solve for $\alpha(\lambda)$ [14–16]. This approach is interesting because it converts a nonlinear problem to a linear (albeit ill-posed) problem with each λ_j in principle automatically detected by a peak in $\alpha(\lambda)$ at $\lambda = \lambda_j$. Using improved methods for solving for $\alpha(\lambda)$, this approach has proved very useful when N_λ is so large that the individual λ_j cannot be resolved and $\alpha(\lambda)$ becomes effectively continuous [17–19]. However, when the λ_j can be resolved using (1.1), we have found the accuracy and resolving power of the information-transform method to be far superior. In practice, it is often useful to analyze a set of data in terms of both (1.1) and (1.5) to help decide which model is more appropriate [17–19].

2. FISHER INFORMATION AND GENERALIZED VARIANCE

A very useful measure of the information about $\bar{\boldsymbol{\theta}}$ contained in the reduced data $\boldsymbol{\eta}$ is given by Fisher's information matrix [10, p. 41; 20], the $2N_\lambda \times 2N_\lambda$ matrix $I(\bar{\boldsymbol{\theta}}, A)$ with elements

$$\left[I(\bar{\boldsymbol{\theta}}, A) \right]_{jk} = \int \left[\frac{\partial \ln L(\boldsymbol{\eta}, \boldsymbol{\theta})}{\partial \theta_j} \frac{\partial \ln L(\boldsymbol{\eta}, \boldsymbol{\theta})}{\partial \theta_k} \right]_{\boldsymbol{\theta} = \bar{\boldsymbol{\theta}}} L(\boldsymbol{\eta}, \bar{\boldsymbol{\theta}})\, d\boldsymbol{\eta}, \tag{2.1}$$

where the integration is over all possible $\boldsymbol{\eta}$ and, since $\mathbf{y}$ and therefore $\boldsymbol{\eta}$ are normally distributed, the likelihood function is

$$L(\boldsymbol{\eta}, \boldsymbol{\theta}) = \left[(2\pi)^{2N_\lambda} \det(M_\eta) \right]^{-1/2} \exp\left[-\tfrac{1}{2} (\boldsymbol{\eta} - \hat{\boldsymbol{\eta}})^{\mathrm{T}} M_\eta^{-1} (\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}) \right], \tag{2.2}$$

where from (1.3) it is easy to verify that the covariance matrix of $\boldsymbol{\eta}$ is

$$M_\eta = AM_y A^T, \tag{2.3}$$

where A^T denotes the transpose of A. Substituting (2.2) into (2.1), it is straightforward to show

$$I(\bar{\boldsymbol{\theta}}, A) = G_\eta^T M_\eta^{-1} G_\eta, \tag{2.4}$$

where G_η is the $2N_\lambda \times 2N_\lambda$ matrix with elements

$$[G_\eta]_{ij} = \left[\frac{\partial \hat{\eta}_i}{\partial \theta_j} \right]_{\boldsymbol{\theta} = \bar{\boldsymbol{\theta}}}. \tag{2.5}$$

$I(\bar{\boldsymbol{\theta}}, A)$ is just the normal-equations matrix that would occur at convergence to $\boldsymbol{\theta} = \bar{\boldsymbol{\theta}}$ of a least squares analysis of (1.4) using the usual Gauss approximation to the Hessian [21, p. 97]. The usual approximation to the covariance matrix of the estimates of $\bar{\boldsymbol{\theta}}$ is simply $I^{-1}(\bar{\boldsymbol{\theta}}, A)$; this is also the Cramer-Rao lower bound for the covariance of an unbiased estimate of $\bar{\boldsymbol{\theta}}$ [10, pp. 41,77; 21, p. 41]. The generalized variance is defined as [20]

$$D(\bar{\boldsymbol{\theta}}, A) = \det\left[I^{-1}(\bar{\boldsymbol{\theta}}, A) \right]. \tag{2.6}$$

This determinant is a measure of the uncertainties in the parameter estimates as well as the correlations between them. In particular, it is the product of the variances of the canonical variables [21, p. 174], and $[D(\bar{\boldsymbol{\theta}}, A)]^{N_\lambda}$ is proportional to the volumes of the usual approximate hyperellipsoidal confidence regions for the parameter estimates [20].

In parameter estimation problems, the criterion of choosing an experimental design [11, p. 63] or a test input signal that minimizes the generalized variance is referred to as "D-optimality" and is generally considered to be the best criterion [20]. It therefore would be natural to look for a transformation matrix A that minimizes $D(\bar{\boldsymbol{\theta}}, A)$. Furthermore, the sizes of $D(\bar{\boldsymbol{\theta}}, A)$ for the A-matrix for each of the transform methods mentioned in Sec. 1 provide useful criteria for comparing the methods and for optimizing any degrees of freedom in them.

3. INFORMATION TRANSFORMS

In practice the data are preweighted so that M_y is an identity matrix. Thus, if the original data vector is $\mathbf{z}$ with nonsingular covariance M_z, we make the transformation

$$\mathbf{y} = VE^{-1/2}V^T\mathbf{z}, \tag{3.1}$$

where $M_z = VEV^T$ is the eigenvalue decomposition of M_z. (For uncorrelated noise in $\mathbf{z}$, $M_z = E$ is already diagonal and no eigenvalue decomposition is necessary.) From (2.3) and (2.4) we then have

$$I(\bar{\boldsymbol{\theta}}, A) = G_\eta^T (AA^T)^{-1} G_\eta. \tag{3.2}$$

First consider the problem of minimizing $D(\bar{\boldsymbol{\theta}}, A)$ when $N_\lambda = 1$ and we only want to estimate λ_1 from one transform. [In the least-squares analysis of (1.4) or (1.2), we eliminate the linear parameters α_j as implicit functions of the λ_j [22, 23, 14]. For simplicity we do not do so here.] Then $I(\bar{\boldsymbol{\theta}}, A)$ is a scalar, the reciprocal of $D(\bar{\boldsymbol{\theta}}, A)$, and from (1.3), (2.5), and (3.2) we obtain

$$I(\bar{\lambda}_1, \mathbf{a}^T) = (\mathbf{a}^T \hat{\mathbf{y}}')^2 / (\mathbf{a}^T \mathbf{a}), \tag{3.3}$$

where $\mathbf{a} = A^T$, and $\hat{\mathbf{y}}'$ is also an $N_y \times 1$ vector with elements

$$\hat{y}'_k = \left[\frac{\partial \hat{y}(t_k, \lambda_1)}{\partial \lambda_1} \right]_{\lambda_1 = \bar{\lambda}_1} = \alpha_1 \left[\frac{\partial f(\lambda_1, t_k)}{\partial \lambda_1} \right]_{\lambda_1 = \bar{\lambda}_1}. \tag{3.4}$$

Clearly, $I(\lambda_1, \mathbf{a}^T)$ will be maximized with the transform $\mathbf{a} = \gamma \hat{\mathbf{y}}'$, where γ is an arbitrary scalar.

This is just the transformation made in forming the normal equation in a standard least-squares analysis of (1.2) at convergence to $\lambda_1 = \bar{\lambda}_1$. Thus, as expected, the transform in least squares is the maximum-information one. However, since $\bar{\lambda}_1$ is not known *a priori*, we must define a new transform at each iteration of the least squares analysis, and it is this time-consuming step that we wish to avoid.

Therefore, we maximize the expectation value

$$\langle I(\mathbf{a}^T) \rangle = \int p(\lambda_1) I(\lambda_1, \mathbf{a}^T) \, d\lambda_1, \tag{3.5}$$

where the intergration is over all possible values of λ_1, and $p(\lambda_1)$ is the *a priori* probability density for λ_1. [In Sec. 5 we assume the usual case of no *a priori* knowledge of $\boldsymbol{\theta}$ and take $p(\boldsymbol{\theta})$ as a constant that can be ignored.[1]] We restrict $\mathbf{a}$ to a unit vector, since (3.3) is independent of the scale of $\mathbf{a}$. Then combining (3.3)–(3.5), expanding, and integrating term by term, we

[1]We assume that the observable range of λ_1 is finite, and therefore that such a uniform distribution $p(\lambda_1)$ exists. This finite range is usually implied by the finite measurement range of t [14]. This assumption is also necessary for setting up the interpolation grid in Sec. 4. If the range of λ_1 really were infinite, then a change to a new variable would be necessary.

obtain the quadratic form

$$\langle I(\mathbf{a}^T)\rangle = \mathbf{a}^T B \mathbf{a}, \tag{3.6}$$

where B is the positive semidefinite $N_y \times N_y$ matrix with elements

$$B_{jk} = \alpha_1^2 \int p(\lambda_1) \frac{\partial f(\lambda_1, t_j)}{\partial \lambda_1} \frac{\partial f(\lambda_1, t_k)}{\partial \lambda_1} d\lambda_1. \tag{3.7}$$

(The constant α_1^2 can be ignored here.) The eigenvalues Λ_k of B are nonnegative, and we arrange them in decreasing order with increasing k. Let ψ_k be the corresponding orthonormalized eigenvectors. Then

$$\langle I(\psi_k^T)\rangle = \Lambda_k, \tag{3.8}$$

and $\langle I(\mathbf{a}^T)\rangle$ is maximized by $\mathbf{a} = \psi_1$ [24]. We therefore define the information transform as $A = \psi_1^T$.

Of all $\mathbf{a}$ orthogonal to ψ_1, the one maximizing $\langle I(\mathbf{a}^T)\rangle$ is ψ_2, and of all $\mathbf{a}$ orthogonal to ψ_1 and ψ_2, ψ_3 is optimal, etc. [24]. Since the ψ_k are orthogonal, the information carried by each of their $\eta_k = \psi_k^T \mathbf{y}$ is additive. Thus the total $\langle I \rangle$ carried by the first N reduced data points η_k is

$$\langle I \rangle_N = \sum_{k=1}^{N} \Lambda_k. \tag{3.9}$$

If $N = N_y$, then $\boldsymbol{\eta}$ would just be an orthogonal transformation of $\mathbf{y}$; $\boldsymbol{\eta}$ would contain all the information in $\mathbf{y}$, and least-squares analyses of $\boldsymbol{\eta}$ and $\mathbf{y}$ would be identical. Often, the Λ_k rapidly decrease with increasing k, and the fraction of the total $\langle I \rangle$ carried by the first N reduced data η_k obeys

$$\langle I \rangle_N / \langle I \rangle_{N_y} \approx 1 \tag{3.10}$$

for some $N \ll N_y$. This is analogous to the efficiency of the first N principal components or Karhunen-Loeve transforms in linear problems, except for two complications due to the nonlinearity: first, we are forced to use expectation values, and second, even a maximum-likelihood estimate using all the data generally only achieves full efficiency in the asymptotic limit of large N_y [10, p. 77].

When $N_\lambda > 1$, we need $2N_\lambda$ transforms for the $2N_\lambda$ parameters in (1.3). Something like the above procedure would require optimizing an expression containing a determinant, which is very difficult [11, Chapter 3]. Because of (3.9) and (3.10) and the fact that the λ_j take the same form in each term in the sum in (1.1), we take the transforms used for (3.9) and (3.10) to be our information transforms; i.e., the $2N_\lambda$ rows of A are simply $\psi_1^T, \psi_2^T, \ldots, \psi_{2N_\lambda}^T$.

While certainly not optimum, we have found these transforms to be robust and reliable in a wide range of applications. Some examples are given in Sec. 5.

4. IMPLEMENTATION

Two computational problems must be overcome. First, the eigenvector decompositions of M_z and B would require too much computation for $N_y \gtrsim 100$. Therefore, the program reduces these matrices to $n_y \times n_y$ by selecting approximately only every $[N_y/n_y]$th row and column of M_z and only every $[N_y/n_y]$th t_k in forming B in (3.7). The full ψ-vectors are then approximated by linearly interpolating between the n_y elements in the eigenvectors actually calculated. An n_y between 40 and 80 has always been sufficient. This was tested by repeating the entire analysis with a larger n_y and comparing the results. This is to be expected, since the full ψ are suboptimal anyway, and reasonable approximations of the general forms of the full ψ should be sufficient.

The second problem is that the least-squares analysis of (1.4) requires $\hat{\eta}_i$ and $\partial\hat{\eta}_i/\partial\lambda_j$, $i,j=1,\ldots,2N_\lambda$, to be evaluated at each iteration. From (1.1) and (1.4),

$$\hat{\eta}_i(\boldsymbol{\theta}) = \sum_{j=1}^{N_\lambda} \alpha_j s_i(\lambda_j), \tag{4.1}$$

$$s_i(\lambda) = \sum_{k=1}^{N_y} A_{ik} f(\lambda, t_k). \tag{4.2}$$

Thus, although the number of data is reduced from N_y to $2N_\lambda$ in going from (1.2) to (1.4), each $s_i(\lambda)$ requires N_y functional evaluations, and the amount of computation is the same. However, the program tabulates the $s_i(\lambda)$ and $ds_i(\lambda)/d\lambda$ at 125 equally spaced grid points covering the allowed region [14] of the λ-axis, once and for all, and does all subsequent evaluations rapidly and accurately with seven-point Lagrange and three-point Hermite interpolation [25], respectively. If N_y is large and $f(\lambda, t)$ complicated, this can increase the speed by orders of magnitude. This permits a much more thorough analysis with an elaborate grid search of parameter space and many separate least squares analyses from different starting points in λ-space. The details of the least-squares procedure are given elsewhere [14].

5. NUMERICAL TESTS

Several hundred simulated data sets were analyzed; a few representative results are given here. The first three cases are Examples F, H, and N of

McKinnon et al. [7] simulating fluorescence decay data where a two-component exponential decay is convoluted with the exciting lamp intensity, $F(\tau)$:

$$f(\bar{\lambda}_j, t_k) = \int_0^{t_k} F(\tau)\exp\left[-\beta_j(t_k-\tau)\right]d\tau, \qquad j=1,2, \tag{5.1}$$

$$F(\tau) = 5.802\tau^2 e^{-0.4\tau}, \tag{5.2}$$

where $t_k = k$, $k = 1,\ldots,160$, $\beta_1 = 0.08$, and we use $\lambda = \ln\beta$ to better cover the wide range of β [14, 16]. For Examples F, H, and N, $\beta_2 = 0.05, 0.02, 0.05$ and $\bar{\alpha}_1/\bar{\alpha}_2 = 10, 1, 0.1$, respectively. The $\bar{\alpha}_j$ were adjusted so that $\hat{y}(t,\bar{\boldsymbol{\theta}}) = 10^4$ at its maximum. Under these conditions, we always have $\hat{y}(t_k,\bar{\boldsymbol{\theta}}) > 100$, and we therefore use pseudorandom uncorrelated zero-mean normal ε_k with variance $\hat{y}(t_k,\bar{\boldsymbol{\theta}})$ as a good approximation to Poisson noise.

Example 4 involves a sum of four exponentials:

$$f(\bar{\lambda}_j, t_k) = \exp\left[-\beta_j t_k\right]. \tag{5.3}$$

For $j = 1,\ldots,4$, $\alpha_j = 1$ and $\beta_j = 0.004, 0.02, 0.1, 0.5$. There are nine groups of data points with 40 in the first group and 15 in the others. In group m, $t_k - t_{k-1} = 0.1\times 2^m$. The range covered is $t_1 = 0.2$ to $t_{160} = 1538$. This type of logarithmic sampling is often used in dynamic studies of processes covering a wide time range [26]. The ε_k were pseudorandom independent zero-mean normal deviates with a standard deviation of 0.01.

The following transforms were compared:

(1) *Information transforms* (INF): The integration limits in (3.7) were from $-\infty$ to ∞; in Examples F, H, and N, (3.7) was approximated by the trapezoidal rule over a sufficiently wide finite range of λ.

(2) DISCRETE *transforms* (DIS):

$$A_{nk} = \cos\left[\pi(n-1)\frac{\ln(t_k/t_1)}{\ln(t_{N_y}/t_1)}\right], \qquad n \text{ odd},$$

$$A_{nk} = \sin\left[\pi n\frac{\ln(t_k/t_1)}{\ln(t_{N_y}/t_1)}\right], \qquad n \text{ even}. \tag{5.4}$$

These are based on the fact that the main features of the eigenfunctions of the kernel (5.3) in the λ and $\ln t$ variables are well represented by (5.4) [14]. The transforms are the basis of DISCRETE [14, 16], a widely used program for the automatic analysis of multicomponent exponential decay data.

(3) *Modulating function transforms* (MOD) [3, 7]:

$$A_{nk} = w_k t_k^{\gamma_n}(t_{N_y} - t_k)^{\gamma_n}, \tag{5.5}$$

where $\nu_n = 4,4,4,4,5,8,14,32$ and $\gamma_n = 32,14,8,5,4,4,4,4$, $n = 1,\ldots,8$. In Examples F, H, and N, only $n \leqslant 4$ are used. In (5.5)–(5.7), A is a discrete approximation to an integral transform, and the $w_k = t_k - t_{k-1}$ (with $t_0 \equiv 0$) are just the approximate trapezoidal-rule weights.

(4) *Laplace transforms* (LAP) [4,7]:

$$A_{nk} = w_k \exp[-(n-1)s' t_k], \tag{5.6}$$

where $s' = 0.01$ for Examples F, H, and N. In Example 4, $s' = 0.0178$ gave the best results of several s'-values tried.

(5) *Moment transforms* (MOM):

$$A_{nk} = \frac{w_k t_k^{n-1}}{\sum_{i=1}^{N_y} w_i t_i^{n-1} y_i}. \tag{5.7}$$

This is the usual form [1,2,7] except that the transforms are normalized to one to prevent numerical ill-conditioning in (1.4) due to large variations in the magnitude of the unnormalized moments.

A greater numerical stability in DIS, MOD, LAP, and MOM could be obtained by working with a set of uncorrelated unit-variance transforms, $M_\eta^{-1/2}\boldsymbol{\eta}$, analogous to (3.1). In INF, M_η is already an identity matrix. In DIS the data are preweighted as in (3.1). No statistical weighting of the data is possible with modulating functions, moments, or Laplace transforms.

From the above specifications, the generalized variance $D(\bar{\boldsymbol{\theta}},A)$ was evaluated from (2.6) for the five transforms and also for the case of no data reduction, i.e., when A is the $N_y \times N_y$ identity matrix and all the information is used by analyzing the full data vector $\mathbf{y}$. We call this last generalized variance $D_{\min}(\bar{\boldsymbol{\theta}})$. The ratio $D(\bar{\boldsymbol{\theta}},A)/D_{\min}(\bar{\boldsymbol{\theta}})$ is a good measure of the information lost by the data-reduction transformation A; the reciprocal of this ratio is analogous to the efficiency [10, p. 39] of a single-parameter estimator. Another measure is $R(\bar{\boldsymbol{\theta}},A)/R_{\min}(\bar{\boldsymbol{\theta}})$, the relative increase in the estimated root-mean-relative-squared error:

$$R(\bar{\boldsymbol{\theta}},A) = \left[\frac{1}{2N_\lambda}\sum_{j=1}^{N_\lambda}\frac{\sigma^2(\alpha_j)}{(\bar{\alpha}_j)^2} + \frac{\sigma^2(\beta_j)}{(\bar{\beta}_j)^2}\right]^{1/2}, \tag{5.8}$$

where the variance estimates $\sigma^2(\cdot)$ are the diagonal elements of the covariance matrix for the β_j and α_j parameters, and $R_{\min}(\bar{\boldsymbol{\theta}})$ is the $R(\bar{\boldsymbol{\theta}},A)$ when there is no data reduction. The D-ratio directly accounts for correlations in

TABLE 1

$D(\bar{\theta},A)/D_{\min}(\bar{\theta})$, the Relative Increase in the Generalized Variance, and (in Parentheses) $R(\bar{\theta},A)/R_{\min}(\bar{\theta})$, the Relative Increase in the Estimated Root-Mean-Relative-Squared Error, Due to the Information Lost by the Data-Reduction Transforms

	Examples			
Transform	F	H	N	4
Information	1.47 (1.12)	1.18 (1.04)	1.70 (1.19)	14 (2.5)
DISCRETE	5.83 (2.14)	3.83 (1.73)	3.20 (1.43)	10 (1.7)
Modulating functions	13.84 (2.57)	3.55 (1.39)	3.81 (1.64)	10^{19} (10^{4})
Laplace	2.09 (1.44)	1.07 (1.02)	1.11 (1.04)	294 (2.8)
Moments	1.13 (1.05)	1.72 (1.25)	1.17 (1.07)	$>10^{30}$

the estimates as well as errors and is generally preferred [20]. Note that both ratios are invariant to multiplying the noise vector ε by a constant.

These two ratios are given in Table 1. In the four-parameter examples, F, H, and N, the information loss is not very serious, especially for INF, LAP, and MOM, where the generalized variance is increased by at most a factor of two. However, the information content deteriorates drastically with more parameters. In Example 4, MOM and MOD are very poor, and LAP required a careful choice of s'.

For each example, five sets of simulated data were analyzed. As expected, the observed $D(\bar{\theta},A)$ and $R(\bar{\theta},A)$ were consistent with Table 1. In all five cases in Example 4, MOM, LAP, and MOD failed to converge because of numerical ill-conditioning in (1.4), in spite of 18-significant-figure arithmetic.

INF was consistently reliable in hundreds of other simulations as well. DIS has worked well in thousands of analyses of multicomponent exponentials but is not as robust as INF in other cases. In ranking the other three, it is important to note that MOM and LAP as implemented here have no cutoff errors, whereas the usual methods of moments [1,2] and Laplace transforms [4] do. Thus, while MOM and LAP did better here with fluorescence decay than MOD,[2] others [7,8] have found the opposite with the usual forms of MOM and LAP. Furthermore, the extra flexibility in choosing the functional form and parameters for modulating functions means that a more effective set could yet be discovered. The criteria of

[2]The usual form of the method of modulating functions also uses successive derivatives of the modulating functions as transforms. When all of these are also included in A, then $D(\bar{\theta},A)/D_{\min}(\bar{\theta})$ in Table 1 is improved to 1.90, 1.87, 1.52, and 10^{12} for Examples F, H, N, and 4, respectively.

minimizing $D(\bar{\theta},A)$ (perhaps averaged over representative $\bar{\theta}$) is natural for choosing a set of modulating functions. This might be worthwhile if the experimental design were nearly constant, i.e., with the same t_k and a reasonably reproducible lamp flash. These modulating functions might then provide a quick and simple means of preliminary data appraisal.

6. CONCLUSIONS

When the noise statistics are known to within a scale factor, least squares (or maximum likelihood for nonnormal statistics) is unquestionably superior to any linear data-reduction transform method, which inevitably must lose information. However, for a given amount of computation time, the more thorough search from several starting points in parameter space permitted by the faster transform methods might be more effective in finding the global optimum, especially for complicated $f(\lambda,t)$ or large N_y. In this case information transforms could be useful.

Transforms like Laplace, moments, and modulating functions are popular because they can lead to very simple equations, some with unique solutions. However, this is misleading because noise is neglected in deriving these equations, and these solutions can therefore be very poor. Although these transforms can be useful in analyzing one- or two-component fluorescence decay data over a narrow time range, the straightforward extension to other cases can lead to very large errors and is not recommended.

The Fisher information measure and the generalized variance provide useful quantitative criteria for assessing transform methods and for choosing the optimum set of transforms from a parametrized family.

REFERENCES

1 I. Isenberg, R. D. Dyson, and R. Hanson, Studies on the analysis of fluorescence decay data by the method of moments, *Biophys. J.* 13:1090–1115 (1973).

2 J. Eisenfeld, S. R. Bernfeld, and S. W. Cheng, System identification problems and the method of moments, *Math. Biosci.* 36:199–211 (1977).

3 B. Valeur, Analysis of time-dependent fluorescence experiments by the method of modulating functions with special attention to pulse fluorimetry, *Chem. Phys.* 30:85–93 (1978).

4 A. Gafni, R. L. Modlin, and L. Brand, Analysis of fluorescence decay curves by means of the Laplace transformation, *Biophys. J.* 15:263–280 (1975).

5 A. Grinvald and I. Z. Steinberg, On the analysis of fluorescence decay kinetics by the method of least squares, *Anal. Biochem.* 59:583–598 (1974).

6 A. Grinvald, The use of standards in the analysis of fluorescence decay data, *Anal. Biochem.* 75:260–280 (1976).

7 A. E. McKinnon, A. G. Szabo, and D. R. Miller, The deconvolution of photoluminescence data, *J. Phys. Chem.* 81:1564–1570 (1977).

8 D. V. O'Connor, W. R. Ware, and J. C. Andre, Deconvolution of fluorescence decay curves. A critical comparison of techniques, *J. Phys. Chem.* 83:1333–1343 (1979).
9 H. O. Hartley, Exact confidence regions for the parameters in non-linear regression laws, *Biometrika* 51:347–353 (1964).
10 S. D. Silvey, *Statistical Inference*, Halsted Press, New York, 1975.
11 V. V. Fedorov, *Theory of Optimal Experiments*, Academic, New York, 1972.
12 H. Akaike, A new look at the statistical model identification, *IEEE Trans. Automatic Control* AC-19:716–723 (1974).
13 T. Söderström, On model structure testing in system identification, *Internat. J. Control* 26:1–18 (1977).
14 S. W. Provencher, An eigenfunction expansion method for the analysis of exponential decay curves, *J. Chem. Phys.* 64:2772–2777 (1976).
15 D. G. Gardner, J. C. Gardner, G. Laush, and W. W. Meinke, Method for the analysis of multicomponent exponential decay curves, *J. Chem. Phys.* 31:978–986 (1959).
16 S. W. Provencher, A Fourier method for the analysis of exponential decay curves, *Biophys. J.* 16:27–41 (1976).
17 S. W. Provencher, Inverse problems in polymer characterization: Direct analysis of polydispersity with photon correlation spectroscopy, *Makromol. Chem.* 180:201–209 (1979).
18 S. W. Provencher, J. Hendrix, L. De Maeyer, and N. Paulussen, Direct determination of molecular weight distributions of polystyrene in cyclohexane with photon correlation spectroscopy, *J. Chem. Phys.* 69:4273–4276 (1978).
19 S. W. Provencher and V. G. Dovi, Direct analysis of continuous relaxation spectra, *J. Biochem. Biophys. Meth.* 1:313–318 (1979).
20 R. K. Mehra, Optimal input signals for parameter estimation in dynamic systems—survey and new results, *IEEE Trans. Automatic Control* AC-19:753–768 (1974).
21 Y. Bard, *Nonlinear Parameter Estimation*, Academic, New York, 1974.
22 G. H. Golub and V. Pereyra, The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate, *SIAM J. Numer. Anal.* 10:413–432 (1973).
23 J. A. Jacquez, A global strategy for nonlinear least squares, *Math. Biosci.* 7:1–8 (1970).
24 R. Bellman, *Introduction to Matrix Analysis*, McGraw-Hill, New York, 2nd ed., 1970, p. 113.
25 F. B. Hildebrand, *Introduction to Numerical Analysis*, McGraw-Hill, New York, 1956, pp. 60, 314.
26 R. H. Austin, K. W. Beeson, S. S. Chan, P. G. Debrunner, R. Downing, L. Eisenstein, H. Frauenfelder, and T. M. Nordland, Transient analyzer with logarithmic time base, *Rev. Sci. Instrum.* 47:445–447 (1976).