ELSEVIER J. Biochem. Biophys. Methods 36 (1998) 157–173

JOURNAL OF biochemical and biophysical methods

# Some pitfalls in curve-fitting and how to avoid them: A case in point

Richard I. Shrager[a,*], Richard W. Hendler[b]

[a]*Mathematical and Statistical Computing Laboratory, Division of Computer Research and Technology, National Institutes of Health, Bethesda MD 20892, USA*

[b]*National Heart Lung and Blood Institute, National Institutes of Health, Bethesda MD 20892, USA*

Received 17 December 1997; accepted 18 March 1998

**Abstract**

When curve-fitting is used to support a complex nonlinear model containing several exponential terms, some of which have closely-spaced time constants, a particular burden of proof must be assumed. Most important, the uniqueness of the solution must be explored and discussed. Statistical tests for the degree of error and independence of the parameters should be provided, as well as information relating to the steps actually used in the fitting procedures. As an example of the need for the procedures we recommend in this communication, we have chosen an important case in point that has been published recently, and which deals with the kinetics of electron transfer from fully-reduced cytochrome oxidase to $O_2$, analyzed by the method of SVD-based least squares. The problems we deal with in this case are applicable to a wide variety of other cases that involve curve-fitting to mathematical models. 

*Keywords:* Curve fitting; Degree of error; Independence of parameters; Testing

## 1. Introduction

Two recent publications, Bose et al. [1] and Sucheta et al. [2] came to different conclusions, even though both groups were investigating the same enzyme using the same type of preparation and the same type of analysis, namely SVD-based least

*Corresponding author.

0165-022X/98/$19.00 
PII: S0165-022X(98)00007-4

squares. While the biochemical issues in these papers are of considerable importance, they are not the subject of this paper, which is devoted to methodological issues. These issues are general, in that they apply to many cases of fitting models to data, especially models which are a sum of similar terms like exponentials or Gaussians.

The plan of this paper is as follows. To begin, we present the particular problem at hand, and explain some of the methods by which it is analyzed. This is followed by an exploration of the case in point, namely, the solution presented by Sucheta et al. [2]. Finally, we explore some pitfalls which may be encountered when dealing with such problems, and suggest procedures to avoid them.

## 2. Presentation of the problem

Optical absorbance spectra are taken at several values of time, and recorded in successive columns of a matrix **A**, so that each element of **A**, namely $a_{i,j}$, is the absorbance at the $i$th wavelength $w_i$, $i = 1$ to $m$. and the $j$th time $t_j$, $j = 1$ to $n$. Each $i$th row of **A** is a time course measured at the fixed wavelength $w_i$.

The kinetic process that produces these spectra is assumed to be first order, and is modelled by a sum of exponentials:

$$Y_{i,j} = y_i(t_j) = c_{i,1}\exp\left(\frac{-t_j}{\tau_1}\right) + c_{i,2}\exp\left(\frac{-t_j}{\tau_2}\right) + \cdots + c_{i,Q}\exp\left(\frac{-t_j}{\tau_Q}\right) + c_{i,Q+1}$$

where the final terms $c_{i,Q+1}$ are base levels. Each $\tau_q$, $q=1$ to $Q$, is called a time constant, which is the time that exponential term has dropped to $1/e$ of its value at $t=0$. Each row of **A** is described by a sum of Q exponential terms plus base level, but the $\boldsymbol{\tau}$'s are in common for all rows, i.e. there are only $Q$ $\boldsymbol{\tau}$'s in the entire problem. In contrast, each row has its own set of $Q+1$ coefficients, so that the matrix **C**, with elements $c_{i,j}$, is of size $m$ by $Q+1$. Therefore, the number of fitted parameters is $Q$ (the number of $\boldsymbol{\tau}$'s) plus $m(Q+1)$ (the number of coefficients).

Let **X** be the $Q+1$ by $n$ matrix of exponentials $x_{j,q}=\exp(-t_j/\tau_q)$, with $x_{j,Q+1}$ for the base level term. The model Y may now be expressed entirely in matrix terms:

$$\mathbf{Y} = \mathrm{C}\mathrm{X}^{\mathrm{T}}$$

where superscript T denotes the matrix transpose. A curve-fitting procedure is used to refine first estimates of $\boldsymbol{\tau}$ by a series of iterations. In each iteration, the current iterate $\boldsymbol{\tau}$ determines the entries in **X**, then **C** is given by $\mathbf{C}=\mathbf{A}(\mathbf{X}^T)^+$, where superscript + denotes the Moore-Penrose pseudoinverse. In this way, the entries in **C** are adjusted to their best (least-squares) values given the current iterate $\boldsymbol{\tau}$. Because the least-squares **C** is directly computable from $\boldsymbol{\tau}$ and the data, the curve-fitting code is usually set up to find $\boldsymbol{\tau}$ alone, and **C** is computed as described above, rather than use the entries of both $k$ and **C** as explicit parameters [3]. The aim of the iterations is to minimize $\mathbf{A}-\mathbf{Y}$ in the least squares sense. Because the matrix **Y** is being fitted to the matrix **A**, the process is called matrix least squares.

For various reasons (e.g. [4–6]), the direct fitting of **A** is often avoided. Rather, **A** is decomposed into three factors:

$$\mathbf{A} = \mathbf{USV}^{\mathrm{T}}$$

called the singular value decomposition (SVD) of **A**, allowing a smaller problem with clearer patterns for modeling. SVD programs are available in most mathematical and statistical program libraries. One then proceeds to use **V** rather than **A** as data, weighting each *j*th column of **V** by $s_{j,j}^2$, the *j*th diagonal element of **S**. Equivalently, one can use the matrix **VS** as data in an unweighted fit. Considerable efficiency is gained by ignoring all but the first few columns of **V**, which contain most of the signal. The factors $s_{j,j}$ are required to make the fitting of **V** equivalent to the fitting of **A** [5]. The goal of the fit is now to minimize $\mathbf{VS}-\mathbf{Y}$ in the least squares sense, where the model **Y** is redefined as $\mathbf{Y}=\mathbf{XC}$ and **C** is computed by $\mathbf{C}=\mathbf{X}^{+}\mathbf{VS}$. Because one is fitting to the output of an SVD program, the procedure is called SVD-based least squares. The same kind of curve-fitting procedure can be used whether one chooses to fit to **VS** or directly to **A**. SVD-based least squares was introduced by us in 1982 [4], with further refinements (mainly weights) in 1986 [5], and a tutorial in 1994 [6].

## 3. The case in point

The case in point is the six-exponential result given on page 558 of [2]. The reported time constants and standard errors for these exponentials in microseconds (μs) are:

$$1.2\pm0.2,\ 10\pm5,\ 26\pm5,\ 32\pm5,\ 86\pm10,\ 1300\pm100$$

We will call this set of parameters p-set A. P-set A is not the solution to the data set whose first six ***u*** and ***v*** vectors are shown in [2]. Rather, p-set A is an average of parameter sets from fits to repeated experiments, and the reported standard errors are also derived from those parameter sets (private communication). The 42 measurement times are exponentially spaced from near 0.05 μs to near 50 ms, covering six orders of magnitude with about seven measurements per decade of time. This means that only 14 of the 42 time values are between 3 μs and 300 μs, where four of the six exponentials exert their primary influence. Further, two of those four exponentials have time constants of 26 and 32 μs, which is a very close spacing for resolving exponentials from such sparse biological data. Further still, the signal-to-noise ratio of the data in [2] is only about 10:1 or less through most of the wavelength range, as shown in [2] Fig. 1. The good ranges appear to be in narrow bands around 480 and 605 nm. But [2] Fig. 4 shows considerable error in the 605 nm band, possibly indicating large error in the time direction at that wavelength. Yet, the reported standard errors for the 26 and 32 μs time constants are quite modest, less than 20%.

It is not clear how it is possible to obtain a result as precise as the one described above. We are concerned that other investigators, noting the stated precision of the results in [2], will be encouraged to engage in modeling projects of similar detail and precision. We believe that the descriptions of the various fitting techniques in [2] are inadequate to enable a reader to do similar work. All manner of small systematic error can thwart or invalidate such results. For example, if the model does not completely explain the data, then by definition, small systematic error exists, and Sucheta et al. [2]

admit that ''Although the residual spectra are reasonably good, some differences between the data and the fit are observed on ~3 μs and ~10 ms time scales.'' Also, looking at [2] Fig. 4, systematic error is evident in the wavelength range 580–620 nm.

To illustrate the dangers of such a detailed model, we used a data set similar in character to the example shown in [2] Fig. 3. The six v-vectors in [2] Fig. 3 were enlarged by a factor of 4.24 using a CANON model NP-6650 II copying machine. The $X$ and $Y$ coordinates were carefully measured with a millimeter ruler. The measurements were then converted to time and magnitude values, which comprised the six columns of our $\mathbf{V}$ matrix. The associated singular values $s_{j,j}$, $j=1$ to 6, are required as weights, but these were not reported in [2]. Instead, we used the averages of singular values obtained in our own studies on the same system [1]. These estimated singular values were:

18.697, 2.3308, 0.3195, 0.1128, 0.0498, 0.0332

The data scanned from [2] Fig. 3 together with our estimated singular values will be called data set B.

Data set B is probably not an exact match to the data in [2] Fig. 3, nor does anyone claim that p-set A is a solution to either of these two data sets. Nevertheless, p-set A provides a fit to data set B that appears to be the same as the fits shown in [2] Fig. 3 (Table 1, panel 1 and Fig. 1, panels A–F). The only major difference between data set B and the data presented in [2] lies in the (likely) difference in the singular values, an issue that will be addressed below.

When p-set A was used as a first estimate, allowing a Nelder-Meade curve-fitting procedure to adjust all six parameters, a result much different from p-set A was obtained (Table 1, panel 2 and Fig. 1, panels A–F). The Nelder-Meade function minimizer `fmins` provided by MATLAB was used in this and all subsequent curve fits in this paper. `fmins` is the same program that was used by Sucheta et al. (private communication). The four middle time constants all became indistinguishable from a value of ~26 μs. The resulting asymptotic standard errors and dependencies (discussed later in this paper) became huge, indicating lack of uniqueness in the fitted parameters.

The apparent separation of the middle four time constants in Table 1 panel is due to our use of minimum-separation constraints throughout this study to avoid singularities. Successive $\boldsymbol{\tau}$'s were required to be in a ratio of at least 1.05. Since p-set A exceeds these requirements, they should not preclude us from duplicating p-set A. In Tables 1 and 2, time constants affected by the constraints are marked with asterisks. The use of this kind of constraint will be discussed later.

Sucheta suggested that we try first estimates of $\boldsymbol{\tau}$ equally spaced on a log scale (private communication). At first, we tried separation by a factor of 10, namely:

$\boldsymbol{\tau} = (0.16, 1.6, 16, 160, 1600, 16000)$

selected to keep the extreme $r$'s within the time range of the data. The result was a solution quite distant from p-set A, yet yielding a substantially lower sum of squares, 14% below that of p-set A (Table 1, panel 3). Next, the first estimates were separated by a factor of 5 rather than 10, namely:

$\boldsymbol{\tau} = [1, 5, 25, 125, 625, 3125]$

Table 1
Fittings to data set B using six exponentials

| Panel | $\tau$(μs) | Error | Dependency | Sum sqs. |
|---|---|---|---|---|
| 1 | 1.20 | 0.78 | 1.2 | 0.35532 |
| | 10.0 | 7.8 | 1.2 | |
| | 25.0 | 83 | 6.0 | |
| | 32.0 | 89 | 6.6 | |
| | 86.0 | 87 | 2.1 | |
| | 1300 | 60 | 1.2 | |
| 2 | 1.76 | 1.06 | 1.3 | 0.34484 |
| | 23.9* | 24 000 | 1100 | |
| | 25.2* | 74 000 | 3400 | |
| | 26.5* | 78 000 | 3500 | |
| | 27.9* | 27 000 | 1200 | |
| | 1370 | 53 | 1.1 | |
| 3 | 0.104 | 0.055 | 1.2 | 0.30307 |
| | 0.879 | 0.41 | 1.3 | |
| | 14.0 | 2.4 | 1.7 | |
| | 29.5 | 3.9 | 1.7 | |
| | 1253 | 63 | 1.3 | |
| | 31 959 | 29 000 | 1.2 | |
| 4 | 1.55 | 0.77 | 1.3 | 0.29969 |
| | 15.1 | 3.7 | 1.8 | |
| | 58.9* | 480 | 46 | |
| | 62.0* | 520 | 46 | |
| | 804 | 140 | 2.0 | |
| | 4215 | 1200 | 1.6 | |

*These time constants are separated by constraints.

yielding a converged solution distinct from all the others, with the lowest sum of squares we found (Table 1, panel 4).

Table 2 shows fittings to data set B using less than six exponentials. To generate first estimates for five-exponential fits, we used p-set A and the other three parameter sets in Table 1, removing one exponential from each set as follows. In panel 1 (p-set A), the time constants 25 and 32 were replaced by the single time constant 28 μs. In Table 1, panel 2, the middle four time constants were replaced by the three values 13, 22, and 31 μs. Both of these first estimates yielded the same converged solution (Table 2, panel 1 and Fig. 1, panels G–L). In Table 1, panel 3, the largest $\tau$, which had a large relative error, was removed. The converged result is in Table 2, panel 2. In Table 1, panel 4, the two middle $r$ values were replaced by a single time constant of 60 μs. The converged result is in Table 2, panel 3. Finally, some effort was made to reduce the number of exponentials to four. The solution shown in Table 2, panel 4 was found from a first estimate using Table 2, panel 3, but replacing the $\tau$ values 18.2 and 23.8 with a single $\tau$ of 21 μs. The converged result is in Table 2, panel 4 and Fig. 1, panels G–L. This result is also duplicated from a considerable variety of other first estimates.

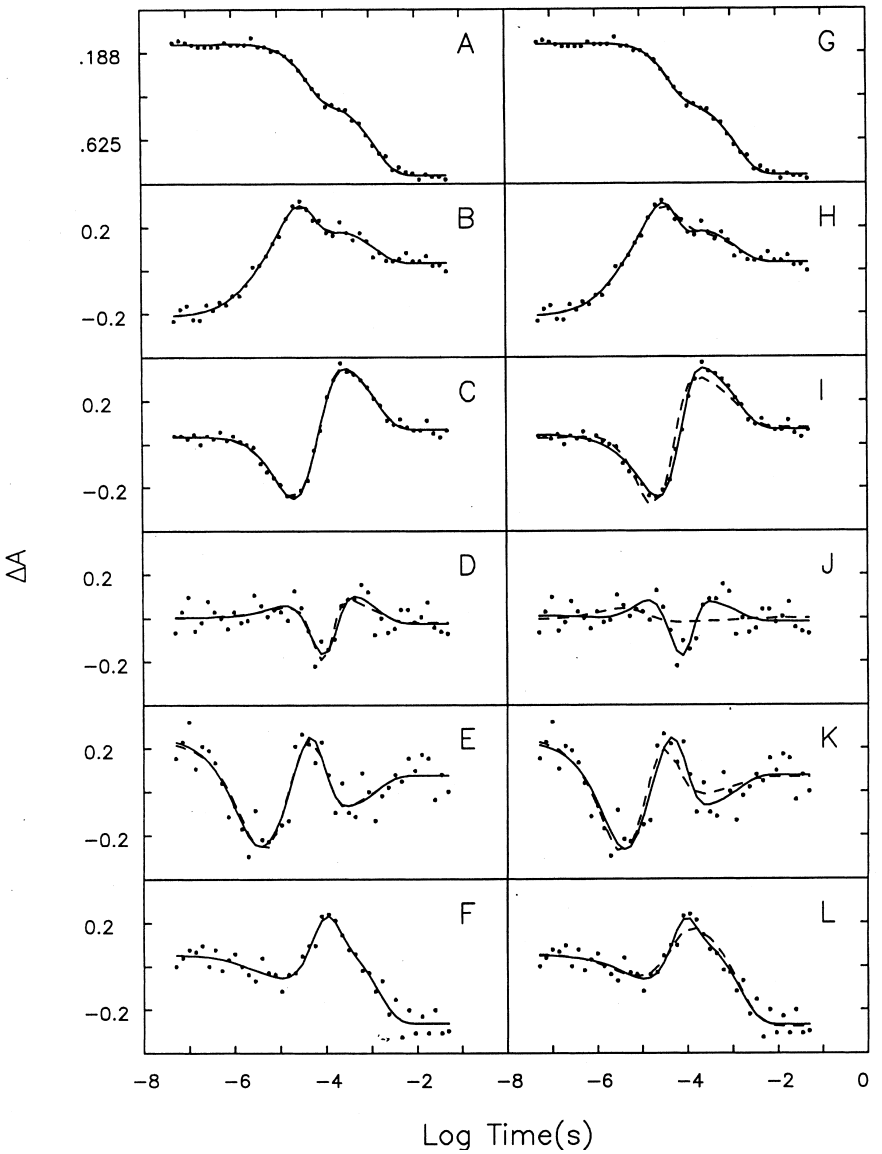


Fig. 1. Fittings to the data of Sucheta et al. The data shown were obtained from [2] Fig. 3 as described in the text. The panels from top to bottom represent the first six columns of the matrix V. In panels A–F, dots are the measured data, solid lines are the sit-exponential fitted curves using p-set **A** (Table 1, panel 1), and dashed lines are the fitted curves using the solution that converged from p-set **A** (Table 1, panel 2). In panels G–L, dots are the measured data (same as panels A–F), solid lines are the five-exponential fitted curves from Table 2, panel 1, and dashed lines are the four-exponential fitted curves from Table 2, panel 4. These four cases are shown because each case was converged from the previous case, with possible removal of an exponential. See text under the headings ‘A case in point’ and ‘How many exponentials?’

Table 2
Fittings to data set B using fewer than six exponentials

| Panel | $\tau$(μs) | Error | Dependency | Sum sqs. |
|---|---|---|---|---|
| 1 | 1.71 | 0.80 | 1.3 | 0.34904 |
| | 29.4* | 1400 | 180 | |
| | 31.0* | 3000 | 345 | |
| | 32.6* | 1600 | 167 | |
| | 1387 | 53 | 1.1 | |
| 2 | 0.105 | 0.060 | 1.3 | 0.33805 |
| | 0.857 | 0.44 | 1.4 | |
| | 13.0 | 2.0 | 1.5 | |
| | 32.2 | 3.6 | 1.5 | |
| | 1435 | 48 | 1.1 | |
| 3 | 1.73 | 0.79 | 1.3 | 0.33410 |
| | 18.2 | 10 | 4.9 | |
| | 23.8 | 11 | 4.8 | |
| | 1232 | 67 | 1.3 | |
| | 25 075 | 19 000 | 1.2 | |
| 4 | 1.66 | 0.84 | 1.3 | 0.36988 |
| | 15.3 | 4.3 | 2.3 | |
| | 28.6 | 5.4 | 2.1 | |
| | 1427 | 50 | 1.1 | |

*These time constants are separated by constraints.

Methods of comparing models with different numbers of parameters will be discussed later. At this point, we observe that neither p-set A (i.e. the averaged result from [2]) nor its converged solution (Table 1, panels 1 and 2) are the best solutions for data set B in the least squares sense. We also observe that all of the five-exponential fits in Table 2 have lower sums of squares than p-set A.

Although we most likely did not use the same singular values as [2], it has been observed that the converged solutions are sometimes not sensitive to those values. See e.g. the example in [5]. Nevertheless, we did not rest on the assumption that the singular values we chose were close enough to get a best comparison to p-set A. Noting that any chosen set of singular values will produce parameters by curve-fitting **VS**, one can try to optimize the singular values by minimizing the difference between the resulting parameters and p-set A in a weighted least squares sense. The weighted sum of squares SS we minimized by adjusting the singular values was:

$$\mathrm{SS} = \left(\frac{\tau_1 - 1.2}{0.2}\right)^2 + \left(\frac{\tau_2 - 10}{5}\right)^2 + \cdots + \left(\frac{\tau_6 - 1300}{100}\right)^2$$

The process was started from various geometric series as the first estimates for the singular values:

$$s_{j,j} = \rho_{j-1}\,, \quad j = 1{:}6$$

where $\rho$ is the ratio for the series of first estimates, but not for subsequent iterates. One can use $s_{1,1}=1$ as a constant value without loss of generality, because the fitted parameters depend only on the ratios of weights, not on their absolute values. For each chosen set of singular values, p-set A was used as a first estimate in the subsequent fit. In no case could we get time constants all of which were within three standard deviations of their counterparts in p-set A. For example, the closest of our solutions to p-set A in the least-squares sense was:

$$\boldsymbol{\tau} = [1, 9, 9.5, 36, 82, 954]$$

Instead of $\tau_3$ and $\tau_4$ being close, as in p-set A, here, $\tau_2$ and $\tau_3$ are close, so close that a constraint was required to keep $\tau_2$ and $\tau_3$ from merging. Also, $\tau_6$ is less than 3/4 the value of that in p-set A. Such solutions offer no support for the idea that this data is a sample from a distribution whose underlying solution is p-set A.

For completeness, some additional observations must be noted about the treatment in [2].

1. Sucheta et al. chose six exponentials by the process of adding one exponential at a time until the fit failed to improve. Their criterion for stopping is not spelled out in the paper. They cite 'The double difference map, along with SVD results' (See [2] Fig. 2 for the double difference map), but not how such procedures were coupled to draw conclusions. In general, using small improvements in the fit to justify additional parameters is not safe. Minor reduction in the sum of squares can be more than offset by a loss in degrees of freedom. Some objective statistical criteria for balancing these opposing effects will be discussed later.
2. The analysis in [2] uses the first six columns of **V**, where the proper data for a six-exponential process contains (at least) seven columns of **V**, one for each exponential transition, with an additional column for the presence of a non-zero final spectrum as shown in [2] Fig. 1. (This 'extra rank' phenomenon is shown in some detail in [6] under the heading 'SVD examples', where a titration of three chromophores is analyzed using four singular vectors.) Several things could explain this omission. The seventh column of **V** could have been so noisy that it didn't seem worth including. (This might also indicate that the noise level was too high to allow resolution of such a model, but in any case, including extra columns of **V** does no harm when proper weights are used.) The spectra involved might be linearly dependent. Finally, the process might in fact require less than six exponentials, even though p-set A with its reported small standard errors would seem to indicate six.
3. In [2], the results discussed above were further used, along with additional spectra from the literature, to deduce microscopic rate constants (thus nearly doubling the number of deduced rate constants) and the associated spectra. Fig. 5 in [2] shows the additional fits that were used to justify the additional rate constants. These fits are mostly poor. Except for [2] Fig. 5, panel e, the experimental spectra are all seen to be of similar shape, with a trough near 480 nm and a peak near 605 nm and two less prominent peaks between. To declare a successful fit data of this kind, one should

account for the subtle differences between the various data curves. Nothing approaching this is seen in [2] Fig. 5. The misfit is such that the theoretical (dotted) curves in panels d and f seem to have been interchanged.

At this point, we end the concerted part of our discussion of the procedures described in [2], although we will refer to [2] occasionally when the subject warrants. The remainder of the paper will describe various difficulties that one must be aware of when fitting exponentials to data, with some suggestions about the avoidance of pitfalls.

## 4. What is a solution?

For linear models, where each parameter is a first power multiplier of a term with no other parameters, a best solution can always be found, in the sense that the sum of squares is the lowest possible. The main difficulty arising in linear problems is that there may be an infinite number of equally good solutions arranged in a linear subspace of parameter space, in which case the problem is called singular. This means that if you have any two distinct solutions, all parameter sets lying on the (infinite) line between them are also solutions. There is also a kind of gray area, where the problem is ill-conditioned or almost singular, which makes it difficult to locate a solution precisely. Whether a linear problem is well-conditioned, ill-conditioned, or singular, there are no isolated local (inferior) minima of the sum of squares. There is only a best (though possibly nonunique) minimum. In any of these cases, it is always possible to produce a solution with an (almost) optimal sum of squares. While uniqueness and precision may be in doubt, optimality is not.

Solutions involving nonlinear models may also be of the above classes, but there are other possibilities as well. Unlike linear problems, where a solution is found essentially in one step, nonlinear problems are solved iteratively. There must be a criterion for stopping. If that criterion is too loose for the current situation, the fitting process will stop before any solution is reached. This premature convergence is possible in any problem where the downhill gradient in the sum of squares is locally very shallow. Also unlike linear problems, nonlinear problems may have several solutions in the form of local minima of the sum of squares. Solving the problem several times with various first estimates of the parameters is a guard against accepting an inferior local solution, but it is not foolproof. There may be an as yet undiscovered better solution. Also, if several minima with comparable sums of squares are found, the current marginally-best solution may not be best in a future trial. If the fit is good, the best one can say is that the current best solution provides the best fit yet found to the data, proving that a model of the proposed form can indeed fit, and that the resulting parameters are viable candidates.

## 5. Error estimation

There are errors in the data from which the fitted parameters are derived. Therefore, from experiment to experiment, the parameters themselves are random variables with

some statistical distribution. One way to estimate standard errors of the parameters is to do several experiments and use the resulting parameter sets as statistical samples of the underlying distribution. Indeed, this would seem to be the most direct method of error estimation, provided the number of experiments is sufficient, and the parameters are well-determined. However, when the parameters have large or infinite error, a stable curve-fitting method can give the appearance of small error. For example, we and Sucheta et al. and many others use the Nelder-Meade simplex method, which is a hill-climber, hunting for a local minimum of the sum of squares, which it may find even when that minimum is not locally unique, i.e. when there is an entire continuum of equally-good solutions. Further, the simplex method can stop short of finding even a local solution when the best downhill direction is so flat that the promised improvement is below some tolerance. The effect of this imposed stability is that the observed variation in the parameters from experiment to experiment appears much smaller than is really the case, producing extremely optimistic error estimates. Diabolically, this stabilizing effect is absent in the best cases when the solution is unique and well-conditioned with small errors, and it is strongest in the worst cases, when the errors in the parameters ought to be huge.

One can illustrate this effect using a two-exponential model:

$$f(t) = c_1 \exp(-t/\tau_1) + c_2 \exp(-t/\tau_2)$$

where $\tau_1$ and $\tau_2$ are to be found by the above fitting procedure. The data points ($t$, $y$) are (0, 1) and (1, 1/e), i.e. two points from a single exponential, with a first-order time constant of 1.0. For these data, the model is too rich because (a) the model contains two exponentials where the data are adequately explained by one, and (b) there are more parameters than data points. We generated 100 data sets by adding normal pseudo-random noise to the two $y$ values with a standard error of 0.01. Using the same simplex method used by Sucheta et al., we fitted those 100 data sets using the same first estimates of the parameters, computing the standard errors of the parameters from the 100 samples. We repeated the entire procedure with a variety of first estimates where $\tau_1$ and $\tau_2$ were placed symmetrically about 1.0, and differed by no more than a factor of 3. The apparent standard errors of the fitted parameters were less than 2% despite the fact that the parameters were underdetermined by the data. But more important, the mean values of $\boldsymbol{\tau}$ were within 2% of the first estimates, regardless of what the first estimates were. In other words if your first estimate is close to the solution you are hoping for, this kind of procedure may simply tell you what you want to hear.

## 6. Ill-conditioning

The fitted function depends on the fitted parameters, which in turn depend on the data. However, the fitted parameters may be extremely sensitive to the data, a situation called ill-conditioning. A small change in the data quite naturally leads to a small change in the fitted function. Yet in a seeming paradox, these small changes in the data and in the fitted function can be accompanied by large, often huge, changes in the fitted parameters. This can only happen when the effect of a change in one parameter can be nearly cancelled

by changes in the other parameters. Sums-of-exponentials models become ill-conditioned when two or more time constants are close in magnitude, because a change in the coefficient of one exponential can be cancelled by the changes in coefficients of the nearby exponentials. When the time constants become too close, the pseudoinverse $\mathbf{X}^+$ becomes impossible to compute without rank deflation, which we seek to avoid, because deflating the rank of $\mathbf{X}$ essentially removes one of the exponentials without permission. One option is to place constraints on the exponential rate constants, so that nearby rate constants are kept apart by at least some ratio near (but not equal to) unity, say $\tau_j/\tau_{j-1}=1.05$. Since the user chooses the ratio, an element of subjectivity is introduced, but not necessarily a serious one. The user is not claiming that a particular ratio is valid, but rather is using that ratio as a device for fitting generic 'close' rate constants that may be justified by a superior fit. Once it is determined that, say, two closely-spaced time constants are essential, one can then rephrase the model to define one time constant as a multiple of its close neighbor, thus eliminating a parameter without eliminating the required term. For example, with 10% separation, the two-exponential model cited above would become:

$$c_1 \exp[-t/\tau_1] + c_2 \exp[-t/(1.1\tau_1)]$$

which might allow resolution of the average of the two time constants, but not their ratio. (However, if this kind of constraint is used, one must say so.)

## 7. Asymptotic error analysis

One advantage of the Levenberg-Marquardt curve-fitting approach over the simplex method is its ability to provide asymptotic statistics at very little additional cost. A discussion of asymptotic statistics is given in [7]. In brief, two kinds of parameter error estimate can be easily produced: the standard error when all parameters are fitted (call that error $\sigma(\tau_j)$), and the standard error when only one $\tau_j$ is fitted, all other parameters being regarded as known fixed constants (call that error $\sigma_1(\tau_j)$ or intrinsic standard error). The nondimensional ratio $\sigma(\tau_j)/\sigma_1(\tau_j)$ is what we call the dependency of $\tau_j$, the factor by which the intrinsic standard error is magnified by the presence of the other parameters. Large dependency values provide a reliable sign of ill-conditioning. (Dependency is closely related to the variance inflation factor discussed in [8].) High dependency warns that perturbing the value of this parameter, or possibly removing it from the fit entirely, would likely be compensated almost completely by adjustments in the remaining parameters, with no significant degradation of the fit. A large dependency, say, 10 or higher, begins to be worrisome, 100 or higher being a strong indication that too many parameters are being used to explain the current set of data. Of course, in a nonlinear problem, it might be possible to converge on another solution that would not have this problem, and one should certainly explore this option. As in any approximation, one must be careful about the interpretaion of asymptotic statistics, being linearized approximations applied to nonlinear parameters. However, among such statistics, very high error estimates and dependencies send a relatively reliable warning

of some problem that needs correcting. Usually that problem is too many parameters, or not enough associated data, or both.

## 8. Computing asymptotic errors and dependency

Asymptotic statistics are generated by approximating the nonlinear model with an associated linear model. The term 'asymptotic' refers to a limit as the experiment is repeated many times and averaged. In theory, the standard errors will diminish by the factor $N^{-1/2}$ where $N$ is the number of experiments. With increasingly high $N$, the relevant region of parameter space will become small enough that the linear approximation will be accurate. Asymptotic errors are estimates of this limiting behavior, usually for the case $N=1$. The matrix of partial derivatives:

$$B = \{b_{ij} = \partial f_i / \partial \tau_j\}$$

is a function of the fitted parameters $\boldsymbol{\tau}$. But for error estimation, **B** is evaluated at the solution $\boldsymbol{\tau}$, and treated as constant. (In a linear problem, **B** is constant, namely, the matrix of constant coefficients of the parameters.) The two-term Taylor's expansion of the non-linear model, which is:

$$f(\boldsymbol{\tau} + \Delta\boldsymbol{\tau}) = f(\boldsymbol{\tau}) + \mathbf{B}\Delta\boldsymbol{\tau}$$

then becomes a set of linear equations in $\Delta\boldsymbol{\tau}$, which can be solved by linear least squares. In addition, statistics associated with the linear fit can be used as approximations in the nonlinear case. Dependency estimates $dep(j)$ for parameters $\tau_j$ are generated as follows:

$$\mathbf{H} = \mathbf{B}^{\mathrm{T}}\mathbf{B} + \mathbf{D}\ ;\ \mathbf{R} = \mathbf{H}^{-1}\ ;\ \mathrm{dep}(\tau_j) = (h_{jj}r_{jj})^{1/2}$$

where $h$ and $r$ denote elements of **H** and **R**, and **D** is a diagonal matrix for stabilizing the computation of **R**. The asymptotic standard errors and intrinsic standard errors are given by:

$$\sigma(\tau_j) = \sigma r_{j,j}^{1/2}$$

$$\sigma_1(\tau_j) = \sigma/\left(h_{j,j}^{1/2}\right)$$

where $\sigma$ with no argument or subscript denotes the standard error of the fit:

$$\sigma = (\mathrm{SOS})/(\mathrm{DOF})$$

The quantity SOS is the sum of squares of the residuals $\mathbf{VS}-\mathbf{Y}$, and degrees of freedom (DOF) is the number of residuals minus the number of parameters. Choosing the proper DOF is somewhat subtle. For purposes of computing standard errors when the linear parameters are subsumed in the fitting function, the number of parameters is $m+Q(m+1)$, where $m$ is the number of $\boldsymbol{\tau}$'s. When the linear and nonlinear parameters are all fitted explicitly, then the number of parameters is $m+Q(m+1)$, where $m$ is the number of columns in **V**. Also, when subjecting $\sigma$ to certain statistical tests, as we will do shortly, the full number of parameters, linear and nonlinear, is required.

Ideally, degrees of freedom, standard errors, and goodness-of-fit tests should be computed from the original data, not from the SVD, but since the raw data from [2] were not at our disposal in this work, we used data set B. If the data matrix **A** were used, then $m$ (above) would be the number of rows in **A**. Computing **R** is a risky operation, because **Q** might be singular or ill-conditioned. To insure that **Q** is invertable, its main diagonal elements $q_{jj}$ are increased by very small amounts $d_{ij}$, enough to insure that the calculation of **R** will proceed regardless of the condition of **Q**. Of course, when **Q** is ill-conditioned or singular, making small changes in **Q** will make large (possibly infinite) changes in **R**, but only in the sense that huge errors and dependencies will shrink to very large errors and dependencies. In other words, the same severe warning flags will be raised, smaller in magnitude to be sure, but not nearly small enough to regard as passable.

## 9. Which estimate of error?

Estimating standard error by sampling repeated experiments can lead to under-estimated error in ill-conditioned cases. One must check the condition of the problem. Asymptotic errors are linear approximations applied to a nonlinear problem, and as such, almost always have built-in biases. For example, we estimate from simulations that asymptotic analysis of p-set A and data set B will produce error estimates too large by a factor of two. Also, the presence of constraints complicates the treatment of all estimates of error. See e.g. [9]. In Tables 1 and 2, the error estimates do not account for the presence of those constraints that are indicated by asterisks. Our goal in that table is to illustrate the difficulty of resolving nearby exponentials, so our statistics are estimates of what they would be if the tabulated solution had been reached with no constraints.

Since each method has its disdvantages, the best approach is to use both error estimates, to check one against the other. But supposc only one experiment has been done? Perhaps repeating it would be prohibitive in time or expense. If the experiment is thorough, so that loss of a single spectrum would not prevent a solution, one can use a jackknife estimate of the type described in [10]. This estimate allows one to get several estimates of the parameters from a single data set, from which sample means and errors can be computed. However, for nonlinear models, this means solving a series of nonlinear problems, which can prove computationally intensive, perhaps multiplying the effort of the original problem many times over. Often, one elects simply to show the asymptotic estimates as qualitative descriptors of resolvability rather than provide more precise estimates of error.

## 10. Underdetermined solutions

In our experience, iterative methods will produce one of three types of solution when faced with an underdetermined fit to a sum of exponentials:

1. Type one is the situation discussed above, namely a non-unique solution close to the initial estimate.
2. Type two is that some exponentials will be merged, in the sense that their exponents will become almost equal. In this case, one may be able to discard some of the merged exponentials.
3. Type three is that some of the exponentials will drift out of significance, by acquiring very small coefficients, or time constants that are beyond anything indicated by the data (e.g. a time constant of 0.1 or 100 in data that runs from 1 to 10). A type three solution may entail an exponential with an exceedingly small coefficient $c_j$ or with a time constant so far below the first non-zero $t$ that it will cease to have any effect on the computed function, e.g. $c_j = 0$, or e.g. $\tau_1 = 0.001$ when $t_1 = 1$.

The particular type of solution will depend on the model, the experimental design (the choice of observations $t$), the noise pattern, and the weights. In any of these cases, either the error estimates or the dependencies of the offending exponents will be huge or infinite. In our work, we have encountered all three types of solution from time to time, but the first type is the most dangerous, because it gives the impression of a successful fit, until the asymptotic statistics tell us otherwise.

## 11. Weights

If the elements of **A** are uncorrelated and have a common variance, then a curve-fit directly to **A** should have uniform weights. But even in this ideal case, if one decides to use SVD-based least squares, the corresponding curve fit to **V** will require varying weights [5]. The elements in each column **V** col $j$ have a variance proportional to $s_{jj}^{-2}$, requiring a weight of $s_{jj}^{2}$ to compensate for increased noise in the less significant columns. In other words, the sum of squares of the weighted residuals given by $\mathbf{VS} - \mathbf{Y}$ (rather than the unweighted residuals $\mathbf{V} - \mathbf{Y}$) should be minimized. Failing this, the noise in those less significant columns of **V** may be raised to the level of signal, giving the impression that there are more features in the data (hence requiring a richer model) than are really there. In other words, the first and most misleading type of underdetermined solution is encouraged.

## 12. How many exponentials?

Sucheta et al. [2] chose a six-exponential model by adding one exponential at a time until the sum of squares failed to improve. They report that using six exponentials rather than five 'significantly improved the residual spectra', but they offer only a qualitative discussion of why they believe this. Our own exploration of data set B provides no evidence that the six exponentials in [2] are supported by the data.

An objective procedure for deciding when the improvement in the residuals is significant is given in [11], where it is referred to as 'the extra sum of squares principle.' Two tests are implied by this principle. These tests require four quantities. Let the sum

of squared residuals (SS) be SS1 for the simpler model (model 1) and SS2 for the more complex model (model 2). Recall that the degrees of freedom (DOF) for a fit is given by:

$$\text{DOF} = (\text{number of observations}) - (\text{number of parameters})$$

and that the number of parameters must include the 'hidden' parameters, i.e. the coefficients of the exponentials, which are different from column to column of **V**. Let the DOF's be DOF1 for the model 1 and DOF2 for model 2.

The easier of the two tests to apply is the variance-of-fit (VOF) test, where VOF is SS/DOF, or specifically, VOF1 is SS1/DOF1 and VOF2 is SS2/DOF2. When two competing models differ in the number of parameters, it is usually the case that SS2<SS1, but it is always the case that DOF2<DOF1, so it may be that VOF2>VOF1 despite model 2 having more parameters. When VOF1 is no larger then VOF2, one can reject the added complexity of model 2 with at least 50% confidence, and more stringent tests are unnecessary.

The *F* ratio test is somewhat more involved, but it allows testing at lower confidence levels. Therefore, if model 2 passes the variance-of-fit test, proceed to the *F* ratio test. Define the following:

$$\text{dSS} = \text{SS1} - \text{SS2},\ \text{dDOF} = \text{DOF1} - \text{DOF2},\ F = (\text{dSS}/\text{dDOF})/[\text{SS2}/\text{DOF2}]$$

Consult the appropriate *F* ratio table for the desired confidence level, using dDOF and DOF2 in that order as the tabular DOF's. If the computed *F* ratio is less than the table entry, model 2 may be rejected at that confidence level. When DOF2 is greater than the largest finite table entry, one should use the 'infinity' entry for DOF2. Computer programs are available for the *F* ratio test, e.g. the MATLAB statistical toolbox from MathWorks, Inc.

The numbers of parameters for the exponential models in question are m base levels, *Q* rate constants, and *mQ* coefficients, where *m* is the number of retained columns of **V**, yielding a total of $m+Q+mQ$ parameters. The corresponding DOF is $mn-m-Q-mQ$, where *n* is the number of rows in **V**. Thus the DOF is decreased by $m+1$ for every added exponential. Some of the panels in Tables 1 and 2 will serve as examples:

| Table | Panel | SS | exps. | DOF | VOF |
|---|---|---|---|---|---|
| 1 | 1 | 0.3552 | 6 | 204 | 0.001742 |
| 1 | 2 | 0.3448 | 6 | 204 | 0.001690 |
| 2 | 1 | 0.3490 | 5 | 211 | 0.001654 |
| 2 | 4 | 0.3699 | 4 | 218 | 0.001697 |

One can see that the SS and VOF columns do not suggest the same order of merit. Note that the lowest VOF is for the five exponential fit. Note also the near-equivalence of the second and fourth examples.

## 13. Check list for reporting the fitting process

Non-linear curve-fitting can require painstaking work, especially in cases where similar terms (e.g. exponentials) threaten to overlap or merge, or where the process is otherwise non-routine in some way. The process is difficult, not only for the investigator, but for the reader, who should be able to repeat the process upon reading the paper. It is not sufficient to leave the details in the references, unless those references are followed very closely. The reader should not have to guess which parts of which references were used and which parts were not. It is far better for the author to describe what was in fact done, especially with regard to the following issues:

1. What kind of data were used?
2. What kind of noise was present, especially if the noise was severely non-normal in some way, e.g. noise spikes. What measures were taken to compensate for that non-normality.
3. What was the model equation?
4. What were the fitted parameters within the model?
5. Were any constraints placed on the fitted parameters? For example, were the $\boldsymbol{\tau}$'s required to remain apart by exactly some factor, or a minimum of some factor?
6. What method of curve-fitting was used? In MATLAB, for example, there are several methods available, depending on which toolbox is available.
7. What efforts were made to insure the existence, local uniqueness, and optimality of the solutions?
8. What statistical weights were used? This issue is of special concern when using SVD-based least squares.
9. What were the singular values in the example? Without these values one cannot judge how certain the rank of the problem is, nor what the proper weights should be.
10. How were the standard errors of the fitted parameters computed? There are several methods of estimation, each with its merits and drawbacks.
11. Were the experiments repeated? If so, how many repeats were there, and how good was the agreement between experiments?
12. Were the results of fitting checked by any other means? For example, were there any independently-derived spectra to compare with some of the fit-derived spectra?
13. Was an overall goodness-of-fit test used? If so was it applied A? **V**? Rows? Columns? If the fitting process is in several stages, at which stages was the test applied?

Omitting many of these details has become a common practice that does no harm to the reader's understanding in routine cases, and, frankly, improves the readability of the work. But there are cases in which such details must be told, especially when they are optional in general, yet indispensable for reproducing the results at hand. The case in point, Sucheta et al., is anything but a routine fitting procedure, yet it is notable for its absence of computational details. At the very least, such details should be put in a first paper of a series, or an appendix, or archived in some publicly-available source that. reflects the author's practice, The author should not refer to some standard regimen that in practice has been abridged, augmented, or otherwise altered without comment.

## 14. Summary

We have discussed some common and serious problems encountered in curve-fitting complex mathematical models to data. As a case in point, we have used a recently-published paper on the kinetics of electron transfer from cytochrome oxidase to $O_2$. Using a data set similar to that of Sucheta et al. [2], we have illustrated problems dealing with non-unique solutions, closely-spaced exponential terms, and overparameterization. We have suggested several specific procedures that can help to avoid these problems.

## References

[1] Bose S, Hendler RW, Shrager RI, Chan SI, Smith PD. Multichannel analysis of single-turnover kinetics of cytochrome $aa_3$ reduction of $O_2$. Biochemistry 1997;36:2439–49.
[2] Sucheta A, Geordiadis KE, Einarsdottir O. Mechanism of cytochrome c oxidase-catalyzed reduction of dioxygen to water: evidence for peroxy and ferryl intermediates at room temperature. Biochemistry 1997;36:554–65.
[3] Golub GH, Pereyra V. The differentiation of pseudoinverses and nonlinear least squares problems whose variables separate. SIAM J Numer Analysis 1973;10:413–32.
[4] Shrager RI, Hendler RW. Titration of individual components in a mixture with resolution of difference spectra, pKs, and redox transitions. Anal Chem 1982;54:1147–52.
[5] Shrager RI. Chemical transitions measured by spectra and resolved using singular value decomposition. Chemometr Intell Lab Sys 1986;1:59–70.
[6] Hendler RW, Shrager RI. Deconvolutions based on singular value decomposition and the pseudoinverse: a guide for beginners. J Biophys Biochem Meth 1994;28:1–33.
[7] Bard Y. Nonlinear parameter estimation. New York: Academic Press, 1974:176–9.
[8] Stuart J, Ord JK. Kendall's Advanced Theory of Statistics. 5th ed. New York: Oxford University Press, 1991:1066.
[9] Shrager RI. Constraint analysis in model building. In: Vogt WG, Mickle MH, editors. Modeling and simulation 5 part 2. Pittsburgh: Instrument Society of America, 1975:991–6.
[10] Efron B. The jackknife, the bootstrap, and other resampling plans. Philadelphia: SIAM, 1982.
[11] Draper NR, Smith H. Applied regression analysis. New York: John Wiley, 1966.