Biophysical Chemistry, 36 (1990) 167–176
Elsevier

BIOCHE 01447

Variability of estimated binding parameters

J. Boiden Pedersen and Ole Knudsen

Fysisk Institut, Odense Universitet, DK-5230 Odense M, Denmark

Received 19 September 1989
Revised manuscript received 28 December 1989
Accepted 17 January 1990

Variability; Binding parameters; Binding analysis; Experimental design

The standard deviation is often used as a measure of the accuracy or reliability of estimated binding parameters and this implies that the parameter values are normally distributed. This may not be the case and we show that the unknown distribution of acceptable parameter values associated with a specific model and a particular set of experimental data can be calculated easily. This can be done for any binding model, linear or non-linear, and the method is very robust and accurate. The effect of the magnitude of the experimental error and the distribution of data points on the variability of the parameters is readily investigated. This makes the method useful for the practical design of experiments in terms of the number and range of concentrations (or doses) which need to be studied in order to obtain the desired acccuracy.

1. Introduction

The critical step in the analysis of ligand-binding experiments is the determination of the binding model that describes the data most appropriately and the estimation of the values of the binding parameters. This requires a binding model to be fitted to the experimental data and a non-linear least-squares method is commonly applied, e.g., the iterative Marquardt algorithm [1,2]. Other more robust methods based on Monte Carlo techniques are, however, sometimes preferable. The goodness of the fit is described by a value which indicates the overall deviation of the theoretical fit from the data. In a least-squares fit, this is the r.m.s. (root mean square) value of the residuals. If the model is correct the deviations (residuals) are due to random experimental error (noise) and the experimental points are then randomly scattered around the theoretical binding curve.

Having obtained a good fit, the next question to arise concerns how reliable the estimated values are. This can be rephrased as: how much can the values of the parameters be varied without any noticeable change in the r.m.s. value? This is the variability of the parameters. Clearly, the larger the experimental errors the greater the variability, but otherwise there is no general rule. Different models and in fact different parameters in a given model may have quite different variability. There are several approximate ways to estimate the variability of a parameter. One method consists of keeping the investigated parameter fixed at a selected value while fitting the data to the model with the other parameters being allowed to vary. If the fit so obtained is approximately as good as the best fit, then this parameter value is also as good as the best-fit value. This method is actually more useful than one would expect. Another way to estimate the variability of the parameters is to use the formulae from linear statistics, e.g., as done by Roughton et al. [3]. This implies that the model is approximately linear in an appropriate

Correspondence address: J. Boiden Pedersen, Fysisk Institut, Odense Universitet, DK-5230 Odense M, Denmark.

region about the best-fit value and that the experimental errors are normally (Gaussian) distributed. By this method one can estimate the standard deviations of the parameters, i.e., the width of the assumed normal distribution of the parameter. The correlations between the parameters can also be estimated. The usefulness of these results depends on the validity of the assumption of the linearity of the model, which cannot be precisely calculated although useful rules of thumb do exist.

The best quantitative description of the variability of binding parameters for a given data set is obviously the complete probability distribution of the parameter values associated with the data. From this one can calculate all average values of interest and the distribution gives us directly the probability that a given value of a parameter is consistent with the data. The detailed form of the distribution is also important, e.g., if the distribution has a flat maximum then a range of values would be equally probable, and this may lead to convergence problems in the fitting procedure.

In the present work we discuss and compare two different methods for calculating the full distribution of parameter values associated with a given data set. One of these methods is commonly applied and is discussed in detail in a recent book [2]. It is, however, a time-consuming and approximate method and so we advocate a different method which uses Bayes statistics and the Monte Carlo technique. This method is superior because it is both quicker and generally applicable.

2. Calculation of variability

The experimental data D consist of a set $\{x_i, y_i\}$ of paired values of the free ligand activity (concentration) x_i and the corresponding saturation y_i. We want to describe the data by a binding model which gives a theoretical relation between activity and binding (saturation) as

$$y = f(x,k) \tag{1}$$

where $f(x,k)$ is a known model function which depends on a set of binding parameters $k = \{k_j\}$. By fitting the data to the model function these parameters are assigned some fixed values but we want to know what other values of the parameters can describe the data. Below we discuss two different methods for calculating this variability.

2.1. Method A: The artificial data method

If the model is correct then the residuals, i.e., the deviations between the binding curve and the data, are due solely to experimental error. Thus, one might take the calculated points $\{x_i, f(x_i)\}$ as representing the 'true' data points. From these true data points one can then generate artificial data representing observed experimental data by adding Gaussian noise with a standard deviation equal either to the standard deviation of the experimental error, if known, or alternatively to the r.m.s. value determined by the fitting procedure. A large number (≈ 1000) of artificial data sets is constructed. Each data set is fitted to the original model function $f(x)$ in order to determine the corresponding parameter values. The distribution of parameter values thus obtained is taken to represent the distribution of parameter values of the original data set. Obviously, the method can only be used if a model function has been found that describes the data correctly. This is not always possible and sometimes one is interested in the variability of the parameters of a model that is not necessarily correct or even the best possible.

2.2. Method B: Method based on Bayes statistics

We wish to calculate the probability $p(f(k) \mid D)$ that the model function f with associated parameter values k describes a given data set D. This probability cannot be directly calculated. We can, however, calculate the 'inverse' quantity, the sampling probability $p(D \mid f(k))$ that the data D will be produced by a specific model $f(k)$. The problem is therefore to connect these two probability distributions. The solution to this problem is given by Bayes theorem. With prior information I, data D, and experimental error σ, the probability that the model $f(k)$ describes the data is given by Bayes theorem as

$$p(f(k) \mid D,\sigma,I) = p(f(k) \mid I)\frac{p(D \mid f(k),\sigma,I)}{p(D \mid \sigma,I)} \tag{2}$$

A more detailed discussion of Bayes theorem and its applications is available in a recent review [4]. The denominator $p(D\,|\,\sigma,I)$ is the probability that the observed data is consistent with the prior information I and the experimental error σ. Being independent of the model $f(k)$, this is just a normalization constant and is not needed. Thus, we are left with an estimation of the prior probability $p(f(k)\,|\,I)$ and the sampling distribution $p(D\,|\,f(k),\sigma,I)$.

The prior probability $p(f(k)\,|\,I)$ is the probability of the model and the model parameters based on some prior information I. A variety of prior information can be included. For the present application we will take the prior information to be that all the parameters must be non-negative which is an obvious requirement that association constants must fulfill. The allowed range of values of k (the sampling space) is always restricted upwards on physical grounds, since an infinitely large association constant has no meaning. If one has specific knowledge of the upper bounds of the parameters then this information should be included directly. However, in most cases this is not necessary, since such bounds may be automatically implemented during the computation (see below).

The sampling distribution $p(D\,|\,f(k),\sigma,I)$ is the probability of generating the data D by a specific model with specified parameter values, experimental noise, and prior information. For a given set of parameter values k, consistent with the prior information, the probability that the model function will generate the observed data is equal to the probability that the experimental errors will make up the residuals (the difference between the calculated values $f(x_i,k)$ and experimentally observed values y_i). Consequently, for Gaussian-distributed errors

$$p(D\,|\,f(k)\sigma,I)=\prod_{i=1}^{m}\frac{1}{\sqrt{2\pi\sigma_i^2}}e^{-\{(y_i-f(x_i,k))^2/2\sigma_i^2\}^{-1}} \tag{3}$$

where σ_i is the experimental error of y_i and m the number of data points. The probability that the model describes the data is now given by eqs 2 and 3. Considering only the k dependence of this expression for a fixed model f we use the shorthand notation $p(k)$ for the normalized distribution $p(f(k)\,|\,D,\sigma,I)$, i.e.

$$p(k)=N^{-1}\prod_{i=1}^{m}\frac{1}{\sqrt{2\pi\sigma_i^2}}e^{-\{(y_i-f(x_i,k))^2/2\sigma_i^2\}^{-1}} \tag{4}$$

where $k>0$ and N is a normalized factor, such that $\int p(\vec{k})\,\mathrm{d}\vec{k}=1$. If all data have the same experimental error, i.e., $\sigma_i=\sigma$, then eq. 4 can be simplified to

$$p(k)=N^{-1}\left(\frac{e^{-\mathrm{r.m.s.}^2(k)/2\sigma^2}}{\sqrt{2\pi\sigma^2}}\right)^m \tag{5}$$

where r.m.s. is the root mean square of the residuals corresponding to the set of parameter values k, i.e.

$$\mathrm{r.m.s.}^2(k)=\frac{1}{m}\sum_{i=1}^{m}(y_i-f(x_i,k))^2 \tag{6}$$

This is an n-dimensional distribution where n is the number of model parameters k_j. From this we can calculate all the quantities of interest, e.g., mean values, variances and covariances of the parameters:

$$\bar{k}_i=\int\dots\int k_i p(k)\,\mathrm{d}k_1\dots\mathrm{d}k_n \tag{7}$$

$$\sigma_{k_i}^2=\int\dots\int\left(k_i-\bar{k}_i\right)^2 p(k)\,\mathrm{d}k_1\dots\mathrm{d}k_n \tag{8}$$

$$\mathrm{covar}(k_i,k_j)=\int\dots\int\left(k_i-\bar{k}_i\right)\left(k_j-\bar{k}_j\right)p(k)\,\mathrm{d}k_1\dots\mathrm{d}k_n \tag{9}$$

One can also calculate the one-dimensional distribution of a single parameter as

$$p(k_i)=\int\dots\int p(k)\,\mathrm{d}k_1\dots\mathrm{d}k_{i-1}\,\mathrm{d}k_{i+1}\dots\mathrm{d}k_n \tag{10}$$

and this gives the distribution of that particular parameter irrespective of the values of the other parameters.

The range of integration is equal to the sampling space and is determined by the prior information as discussed above. In the present case, the lower limit of integration is zero and the upper limit is either determined by the prior information or is more or less arbitrarily set. Strictly speaking, the upper limit of integration cannot be infinity, since the model function given below by eq. 11 tends to a finite value as k goes to infinity and hence the integral does not exist. Unless explicitly required, the upper limit need not be precisely specified, since the probability $p(k)$ is non-vanishing only in a narrow range of order σ_k around $\bar{k}$ and the numerical calculation may be automatically restricted to this range. That range is typically found by performing an initial computation of $p(k)$. Based on the initial results the range is either contracted or expanded to cover the region where $p(k)$ is non-vanishing. This is the most economical choice for the range in the final calculation and this region thus becomes the integration range. Note, however, that if the prior information specifies a range of parameter values which is narrower than that allowed by the experimental error then this range must be used explicitly as the integration range.

As can be seen from the equations all that is needed is an integration over the sampling space. This can be done in a variety of ways, the simplest being with a rectangular grid. However, for dimensions larger than 4 there is a clear advantage in using a Monte Carlo method. Although in many cases the number of parameters is 4 or less, e.g., the binding of oxygen to haemoglobin, we have written a general program which can be used for any number of parameters. The program uses a Monte Carlo technique, discussed in details by Tarantola [5], and calculates the various integrals by sampling from $p(k)$ in parameter space. Typically we used 10^6 sampling points, which for 4 dimensions is more than sufficient to produce a high-quality, continuous picture of the distribution.

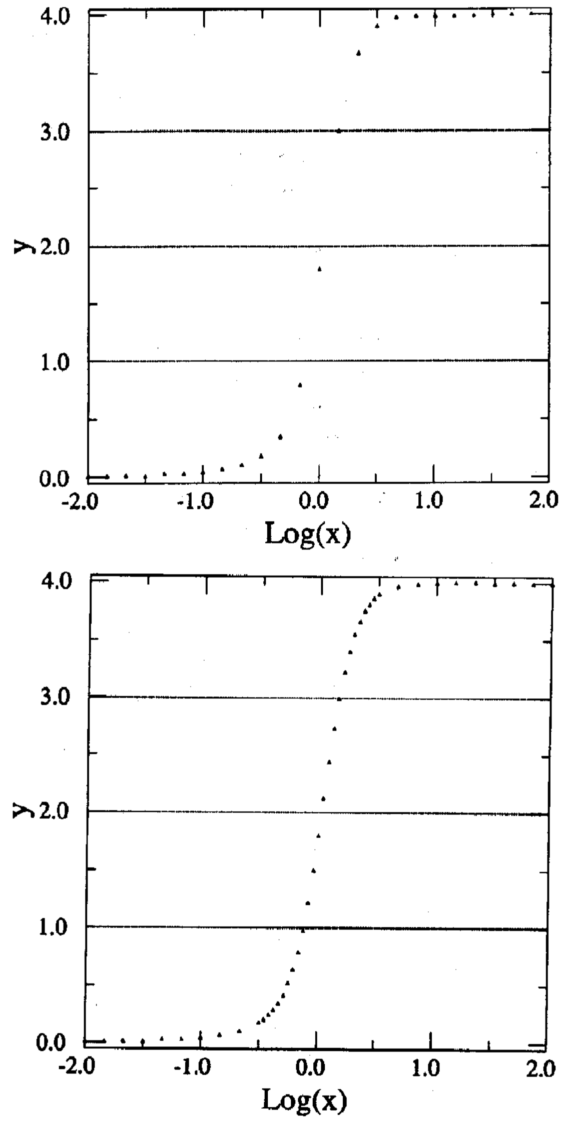

Fig. 1. The standard (top) and extended (bottom) data sets. Both sets were constructed from the Adair equation with the parameter values given in section 3, i.e., $(\beta_1 \ldots \beta_4)$ have the values (0.535, 0.0353, 0.0168, 1.000). Gaussian noise with a standard deviation of 0.004 was added to represent small experimental errors.

3. Data

We have used a constructed data set which corresponds to a highly cooperative case resembling the binding of oxygen to haemoglobin. The original data, i.e. those containing no experimental errors, are described by the Adair equation

$$f(x) = \frac{\sum_{i=1}^{4} i\beta_i x^i}{1 + \sum_{i=1}^{4} \beta_i x^i} \tag{11}$$

where the Adair constants ($\beta_1 \ldots \beta_4$) have the values (0.535, 0.0353, 0.0168, 1.000). The 'observed' ligand activity consists of 25 values which are equidistant on a logarithmic scale in the range of -2.0 to $+2.0$ (cf. fig. 1, top). This data set is identical to that used by Gill et al. [6] in a study using method A.

In addition to this data set we have also used the data set shown in fig. 1 (bottom), which differs from the first data set only in the addition of 18 more densely spaced data points in the middle region where the binding curve is very steep and therefore contains fewer points. This set is called the extended data set.

By adding Gaussian noise to these original data we constructed data representing experimental data. In order to investigate the effect of the magnitude of the experimental error we have used two different noise levels. The low noise with a standard deviation of 0.004 corresponds to the error found in accurate optical absorbance measurements. The data sets displayed in fig. 1 have this noise level. Such small experimental errors are barely visible (cf. fig. 1). The high noise had a standard deviation 5-times higher, i.e., $\sigma = 0.020$.

4. Data analysis

We have analyzed the variability of the three data sets mentioned above: the standard set with small experimental errors, the extended set which has more points in the middle region and the noisy data set which has 5-fold greater experimental error.

By the artificial data method (method A) we have generated 1000 new data sets from the fitted values of the Adair constants, i.e., those obtained by fitting the original data set. All the new data sets were fitted to the Adair equation and the distribution and average values of the parameter values were calculated. This procedure took about 1 h on an Apollo 4000 workstation. The resolution and accuracy of the one-parameter distribution is obviously low for such a 'small' number of data sets, since the distribution is calculated as frequencies of parameter values in finite ranges. One must therefore choose between low resolution with reasonable accuracy or higher resolution with poor accuracy.

Method B based on Bayes statistics required integrations over the parameter space which were performed by a Monte Carlo method using 10^6 sampling points. This took about 3 min on the Apollo 4000 workstation.

In addition to the statistical quantities of mean (m), standard deviation (S.D.) and correlation coefficients between parameters (C_{ij}) we have also calculated %bias which is a measure of the nonlinearity of the model and %error which is a measure of the relative accuracy of a parameter. The latter two quantities are defined as: %bias = $100(\beta_i - m_i)/\beta_i$ and %error = $100 \cdot$ S.D.$/m_i$.

5. Results

Table 1 displays the results of the calculations with the standard data set. In most cases the two methods gave similar results. However, it was noticeable that method A gave smaller values of %bias and larger values of %error than method B. In particular, a large discrepancy existed for the two parameters β_2 and β_3. Investigation of this

Table 1

Variability of estimated Adair parameters for the standard data set

The values denoted fitted are the best-fit values for the specific data set used in the analysis. m, mean value; S.D., standard deviation of the allowed parameter values. The measure of the non-linearity of the model %bias is defined as $100(\beta_i - m_i)/\beta_i$ and the percentage relative error %error is defined as $100 \cdot$ S.E.$/m_i$, where the fitted β values are used.

	Fitted	Variability		%bias	%error
		m	S.D.		
Method A					
β_1	0.517	0.520	0.0160	−0.6	3.1
β_2	0.0565	0.0507	0.0195	10.3	38.5
β_3	0.0050	0.0091	0.0111	−81.8	122.7
β_4	1.002	1.002	0.0061	0.0	0.6
Method B					
β_1	0.517	0.524	0.0156	−1.4	3.0
β_2	0.0565	0.0425	0.0179	24.9	42.2
β_3	0.0050	0.0148	0.0102	−193.9	68.5
β_4	1.002	1.001	0.0060	0.1	0.6

problem revealed that the main cause was the slow convergence of the fitting procedure used in method A. The starting guesses were always the fitted values, since the new fitted values are expected to be similar to these values. However, the distributions of the parameters β_2 and β_3 were rather broad and flat (cf. figs 2 and 3). Method B showed that the most probable value of β_3 was in fact zero. All values of β_3 between the 'true' value and zero were roughly equally probable, i.e., they had similar r.m.s. values. The fitting procedure necessary in method A may therefore stop at any value in this range.

Figs 2 and 3 display the single-parameter distributions calculated by the two methods. It is quite clear that method B gave a more detailed and

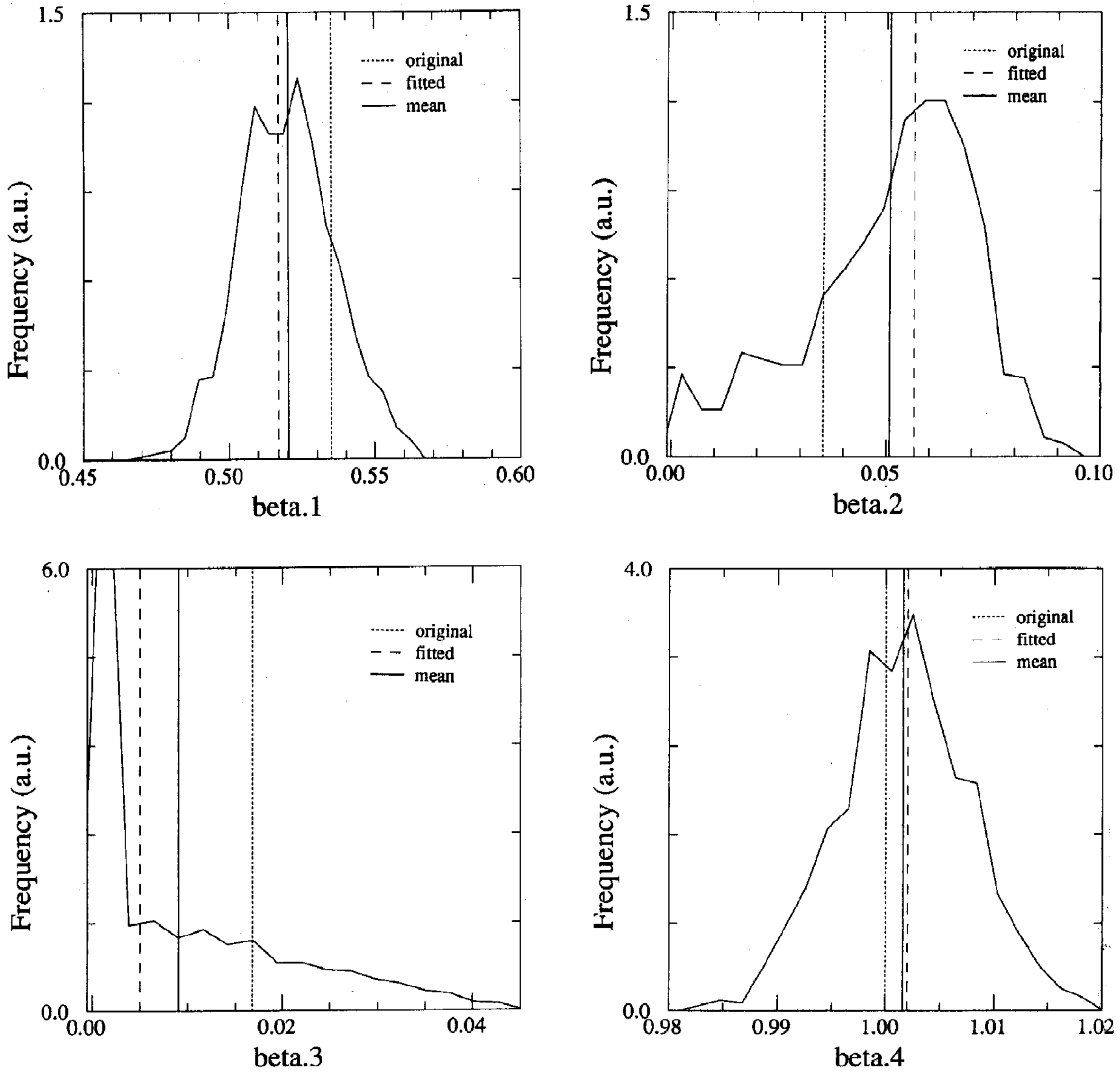

Fig. 2. Single-parameter distributions, corresponding to the standard data set, calculated by method A. The fitted parameter values displayed in table 1 were used to generate the 1000 new data sets.

accurate distribution. The main difference between the two methods was, however, the ability to calculate broad and flat distributions such as those for β_2 and β_3. Method A gave a rather poor picture of these distributions. Method B, on the other hand, calculated all the distributions without any problems. It can be seen that β_3 was not significantly different from zero. In order to see how this was related to the experimental error and the distribution of data points, we repeate›l the analysis for the two other data sets, the extended set and the noisy set.

Table 2 displays the results for the noisy data set which has an experimental error 5-times that of the standard set used above. Again the calculated %bias was smaller with method A than with method B, indicating that the two methods do not calculate the same distribution. The standard devi-

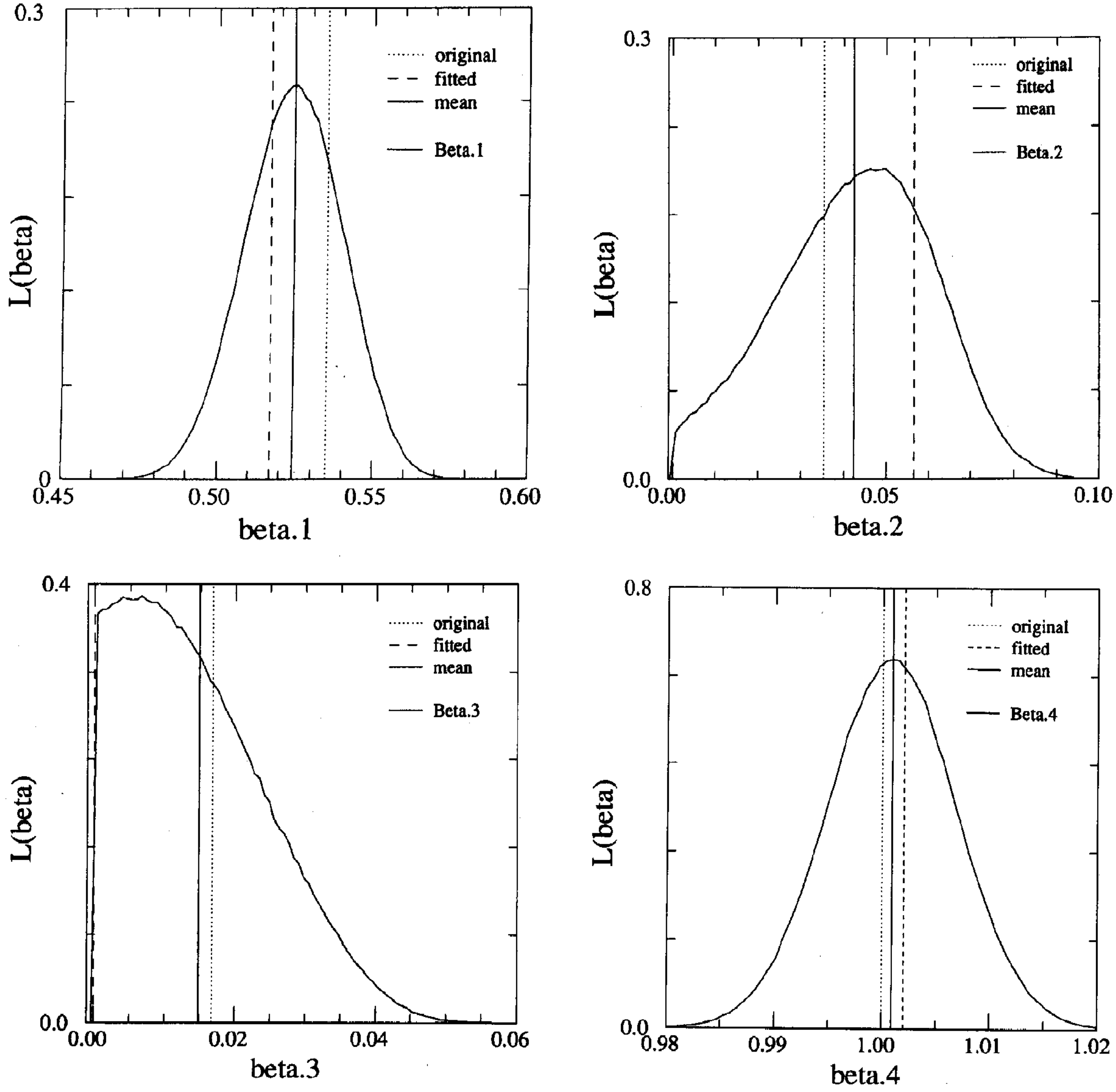

Fig. 3. Single-parameter distributions calculated by method B from the standard data set.

Table 2

Variability of estimated Adair parameters for the noisy data set

See table 1 for explanation.

	Fitted	Variability		%bias	%error
		m	S.D.		
Method A					
β_1	0.537	0.537	0.0633	0.0	11.8
β_2	0.0599	0.0552	0.0567	7.9	102.8
β_3	0.0255	0.0316	0.0359	−23.9	113.5
β_4	0.988	0.986	0.0286	0.2	2.9
Method B					
β_1	0.537	0.523	0.0568	2.6	10.9
β_2	0.0599	0.0643	0.0451	−7.4	70.0
β_3	0.0255	0.0397	0.0282	−55.7	71.0
β_4	0.998	0.979	0.0275	0.9	2.8

ations and the correlation coefficients (not shown) calculated by the two methods were similar, however. As a result of the higher experimental error the estimated value of β_2 was no longer significantly different from zero. Thus, for this degree of experimental error, neither β_2 nor β_3 was significantly different from zero.

The results of the calculations for the extended data set, which has low experimental error and many data points, are displayed in table 3 and the distribution of the parameters β_2 and β_3 for all three data sets are displayed in figs 4 and 5. As a result of the additional data points in the middle region the parameter value of β_3 was now estimated to be significantly different from zero, as is evident from the graphical illustration of the single-parameter distribution in fig. 5 (bottom); the width of the distribution was unchanged but the peak was shifted away from zero. Comparison with table 1 shows that %error was roughly the same while %bias was reduced substantially as a

Table 3

Variability of estimated Adair parameters for the extended data set

See table 1 for explanation.

	Fitted	Variability		%bias	%error
		m	S.D.		
Method A					
β_1	0.536	0.536	0.0142	−0.1	2.6
β_2	0.0368	0.0357	0.0189	3.1	53.0
β_3	0.0150	0.0159	0.0117	−5.9	73.9
β_4	1.000	1.000	0.0036	0.0	0.4
Method B					
β_1	0.536	0.537	0.0125	−0.3	2.3
β_2	0.0368	0.0336	0.0158	8.9	47.2
β_3	0.0150	0.0173	0.0098	−15.5	56.9
β_4	1.000	1.000	0.0038	0.0	0.4

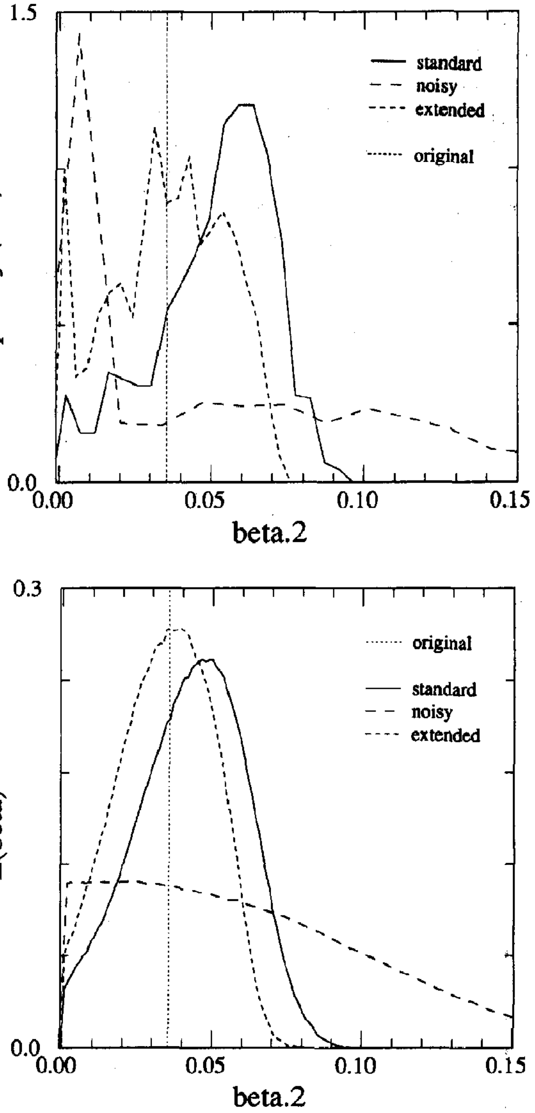

Fig. 4. Single-parameter distributions of β_2 calculated by method A (top) and method B (bottom) for all three data sets as indicated.

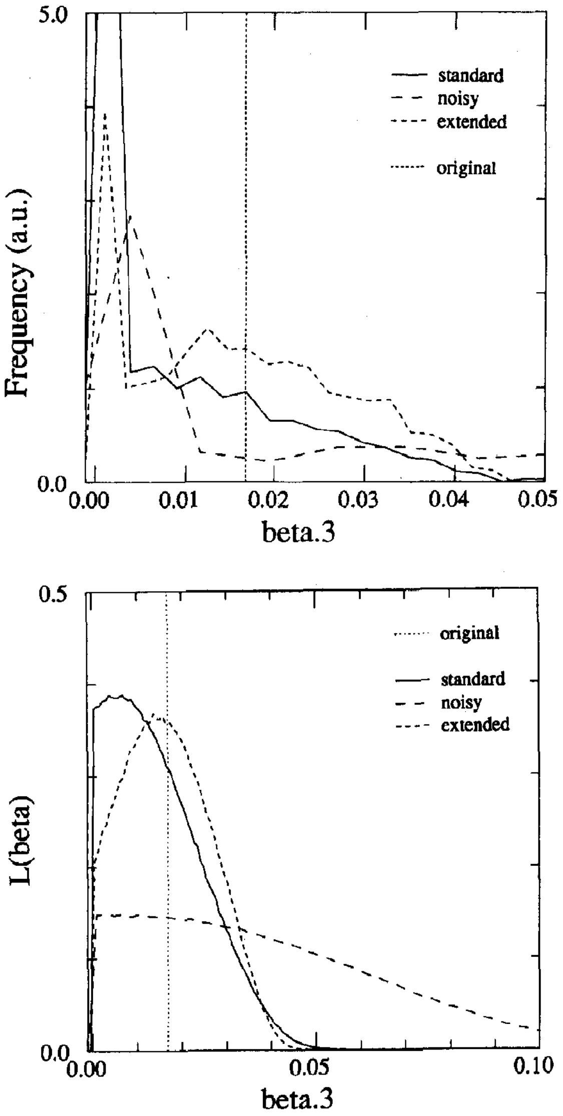

Fig. 5. Single-parameter distributions of β_3 calculated by method A (top) and method B (bottom) for all three data sets as indicated.

result of the extra data points. Only method B gave results of quality sufficient to allow such conclusions (cf. figs 4 and 5).

6. Discussion

It has been shown above that method B is superior to method A in all respects: it is faster, more accurate and robust. We have never observed a case which could not be treated satisfactorily by method B. On the other hand, there are several instances in which method A gives the wrong results. One such case arises when the model function does not describe the data correctly, which may be difficult to decide and which occurs quite often, e.g., when oxygen-binding data are fitted to the MWC model [7] rather than to the Adair equation. Since method A generates new artificial data sets from the best-fit values of the parameters and the assumed model function, it is not surprising that the distribution thus calculated of parameter values will generally be symmetrical about the best-fit value. This distribution merely reflects the way data sets are generated and not the manner in which the experimental data points are distributed around the model curve.

Another serious drawback of method A is its inability to calculate very broad and flat distributions. Method B, on the other hand, is equally useful for any model, correct or not. It will calculate the distribution of parameters in the specified model associated with that particular data set. Even if the model is not the best possible these distributions are very useful in an investigation of the applicability of a model. A necessary condition for the applicability of a model to several data sets obtained under varying conditions may be that some of the parameters are constants and others depend on external parameters in a specified way. By application of method B it is easy to see whether these required dependences are consistent with the data.

The high-quality graphs calculated by method B in this work were obtained by application of 10^6 sampling points. Although this took only a few minutes on our computer one can reduce this time even further by using fewer sampling points. Distributions with an acceptable resolution can be obtained with 10^5 or even 10^4 sampling points. Method A requires a fitting to at least 10^3 data sets to give an acceptable representation of the distribution and this already takes a substantial amount of computer time (1 h). More serious are the problems associated with this method in calculating broad and flat distributions. It should be noted that it is essential to restrict the allowed parameter values to those being physically acceptable. In method A this can be done by transforming the variables. For example, β_i may be written as $\exp(\gamma_i)$ which ensures that $\beta_i \geq 0$ irrespective of the sign of γ_i. In method B such restrictions are directly implemented by restricting the available parameter space. More complicated restrictions can also readily be implemented in method B, e.g., one parameter may be known to have a value in a specified range. Such a restriction would be rather difficult to incorporate in method A. In fact, one of the advantages of Bayes statistics and the present method is the ease with which prior information can be taken into account.

In experimental design the decisions about what concentrations (or doses) to use and what experimental errors to accept are very important. If, a priori, one has a good idea of the parameter values or, better still, results from an initial trial or experiment, then these parameter values can be used to generate an artificial data set (like those used in method A) and then by method B one can investigate the accuracy obtained for the parameters as a function of experimental error and the placement of experimental data. This approach will assist the choice of the range (and number) of concentrations (or doses) to be investigated and in this way optimize the results with respect to a given amount of experimental effort.

References

1 D.W. Marquardt, J. Soc. Indust. Appl. Math. 11 (1963) 431.

2 W.H. Press, B.P. Flannery, S.A. Teukolsky and W.T. Vetterling, Numerical recipes. The art of scientific computing (Cambridge University Press, Cambridge, 1986) ch. 14, p. 529.

3 F.J.W. Roughton, A.B. Otis and R.L.J. Lyster, Proc. Roy. Soc. B 144 (1955) 29.

4 E.T. Jaynes, in: Maximum-entropy and Bayesian methods in inverse problems, eds C.R. Smith and W.T. Grandy, Jr (Reidel, Dordrecht, 1985) p. 21.

5 A. Tarantola, Inverse problem theory. Methods for data fitting and model parameter estimation, ch. 3 (Elsevier, Amsterdam, 1987).

6 S.J. Gill, P.R. Connelly, E. Di Cera and C.H. Robert, Biophys. Chem. 30 (1988) 133.

7 J. Monod, J. Wyman and J.P. Changeux, J. Mol. Biol. 12 (1965) 88.