Available online at www.sciencedirect.com

SCIENCE DIRECT®

Analytical Biochemistry 343 (2005) 106–115

ANALYTICAL BIOCHEMISTRY

www.elsevier.com/locate/yabio

Statistical error in isothermal titration calorimetry: Variance function estimation from generalized least squares

Joel Tellinghuisen *

Department of Chemistry, Vanderbilt University, Nashville, TN 37235, USA

Received 8 March 2005
Available online 6 May 2005

Abstract

The method of generalized least squares (GLS) is used to assess the variance function for isothermal titration calorimetry (ITC) data collected for the 1:1 complexation of Ba^{2+} with 18-crown-6 ether. In the GLS method, the least squares (LS) residuals from the data fit are themselves fitted to a variance function, with iterative adjustment of the weighting function in the data analysis to produce consistency. The data are treated in a pooled fashion, providing 321 fitted residuals from 35 data sets in the final analysis. Heteroscedasticity (nonconstant variance) is clearly indicated. Data error terms proportional to q_i and q_i/v are well defined statistically, where q_i is the heat from the ith injection of titrant and v is the injected volume. The statistical significance of the variance function parameters is confirmed through Monte Carlo calculations that mimic the actual data set. For the data in question, which fall mostly in the range of $q_i = 100$–2000 μcal, the contributions to the data variance from the terms in q_i^2 typically exceed the background constant term for $q_i > 300$ μcal and $v < 10$ μl. Conversely, this means that in reactions with q_i much less than this, heteroscedasticity is not a significant problem. Accordingly, in such cases the standard unweighted fitting procedures provide reliable results for the key parameters, K and ΔH° and their statistical errors. These results also support an important earlier finding: in most ITC work on 1:1 binding processes, the optimal number of injections is 7–10, which is a factor of 3 smaller than the current norm. For high-q reactions, where weighting is needed for optimal LS analysis, tips are given for using the weighting option in the commercial software commonly employed to process ITC data.

Keywords: ITC; Data analysis; Generalized least squares; Nonlinear least squares; Monte Carlo

In the method of isothermal titration calorimetry (ITC)[1], reaction enthalpy q is estimated for sequential addition of one reactant (titrant) to the other (titrate), producing a titration curve of q versus extent of reaction. Analysis of such data provides estimates of the enthalpy change ΔH° and the equilibrium constant K for the reaction. Although titration calorimetry dates back more than 40 years [1], the current burst of interest in this method can be traced to the more recent development of instruments that permit relatively rapid generation of titration curves for small samples (~1 ml volume, ~1 mM concentration of titrate) [2–7]. Such instruments are now widely used to study biochemical binding processes. As an indication of the breadth of interest in this method, a recent check of the Science Citation Index for articles citing "isothermal titration calorimetry" in the titles, abstracts, or keywords turned up works in more than 60 different journals in the year 2004 alone.

ITC data are analyzed by least squares (LS) fitting, with most workers using nonlinear algorithms provided by the manufacturers of the instruments. By default, these packages use unweighted fitting, which tacitly assumes that the statistical error in the data is constant.

* Fax: +1 615 343 1234.
E-mail address: joel.tellinghuisen@vanderbilt.edu.

[1] *Abbreviations used:* ITC, isothermal titration calorimetry; LS, least squares; GLS, generalized least squares; MC, Monte Carlo.

0003-2697/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.ab.2005.04.026

In recent works, I have noted that uncertainty in the delivered volume of titrant from the syringe can be a significant source of random error, leading to noise proportional to q in the data [8,9]. When this type of error predominates, the fitting should employ weights, $w_i \propto q_i^{-2}$, or in one possible model of the volume error, correlated fitting. In a subsequent extensive study of the Ba^{2+} complexation with 18-crown-6 ether [10], the data were fitted in turn to the three limiting models: constant error, proportional error, and correlated error. The results showed a preference for the proportional error model, but it was acknowledged that this model could not be completely valid because every experimental technique is limited by constant error in the small-signal limit.

In the current work, I return to the data from [10] for a more extensive examination of the statistical error. To this end, I use the method of generalized least squares (GLS), in which the variance function itself is estimated as a part of the analysis [11–13]. Usually the goal of a GLS analysis is optimal estimation of the adjustable parameters in the LS fit model, but here the goal is the variance function itself. Confident knowledge of the nature of the statistical error in ITC data can be used to optimize the design of experiments, which was the underlying purpose of my original theoretical investigation [8]. For example, one important result of that study was the realization that as few as five to seven injections may be optimal for 1:1 complexation ($X + M \rightleftarrows MX$) when constant error dominates. Because ITC experiments typically take 5–10 min per injection, this result can mean a tremendous boost to throughput in production run work, where 20–30 injections seems to be the current norm.

In the GLS method, the variance function is estimated in a sort of back-and-forth iterative analysis of the data. The LS residuals from one analysis are themselves subjected to an LS fit to the variance model, the results of which are then used to redefine the weights in the data fit for the next cycle. Convergence typically occurs in four to eight cycles. The capabilities of this method have been fairly well characterized for linear LS fit models, but less well for nonlinear ones, so the GLS method itself is examined in the current work through Monte Carlo (MC) calculations on an ITC model representative of the actual experimental data. The method is applied to the data from [10] for several different proposed variance functions that include constant and proportional error. The results support the expectation of both constant and proportional contributions to the data error. However, for the data used in this study (which involved q values mostly in the range of 100–2000 μcal and injection volumes of 4–40 μl), a simple scale error turns out to be a more important source of proportional error than the volume error identified in [8].

In following sections, I first review the essential aspects of nonlinear LS needed here and then do the same for the GLS method. The latter is tested for a representative ITC data set through MC methods typically involving 30 nine-point data sets at a time. The 37 actual Ba^{2+}/crown ether data sets are then analyzed in a pooled fashion, under the assumption they all are characterized by the same variance function. This analysis is repeated for a half-dozen different proposed variance functions, of which the preferred one to emerge consists of a sum of three variances: one constant and two proportional to q^2. The implications of these results for experiment design are then considered briefly, with an updating of some key results found earlier [8]. Particularly important is the result already noted, namely that heat-starved reactions are better done with few injections.

Materials and methods

Nonlinear least squares and Monte Carlo

The nonlinear LS calculations are done using standard methods, including some that are readily available [14,15]. The specific codes are written in FORTRAN and are similar to those that have been described in studies of bias and non-Gaussian parameter distributions in linear and nonlinear LS [16,17]. At the heart of the calculations is the evaluation of the matrix **A**, given by

$$\mathbf{A} = \mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}, \qquad (1)$$

where the matrix **X** contains elements $X_{ij} = (\partial F_i/\partial \beta_j)$, in which F expresses the fit function in terms of the fit variables (x and y) and the adjustable parameters $\boldsymbol{\beta}$. In the current work, no correlated models are considered, so the weight matrix **W** is diagonal, with elements $W_{ii} = w_i = \sigma_i^{-2}$.

If we take the view that the data error σ_i is known, we can define an a priori variance–covariance matrix $\mathbf{V}_{\mathrm{prior}}$:

$$\mathbf{V}_{\mathrm{prior}} = \mathbf{A}^{-1}. \qquad (2)$$

$\mathbf{V}_{\mathrm{prior}}$ is exact for linear fit models conducted under the usual assumptions of normal unbiased data with an error-free independent variable. For example, it can be used to check a MC code, the output of which should agree (statistically) with $\mathbf{V}_{\mathrm{prior}}$ for a linear model. We can similarly define an "exact" $\mathbf{V}_{\mathrm{prior}}$ for exactly fitting data in a nonlinear model. This nonlinear $\mathbf{V}_{\mathrm{prior}}$ is not truly exact, because nonlinear LS parameters are not normally distributed and might not even have finite variance. However, examination of a number of representative nonlinear problems has led to a 10% rule of thumb that seems to be generally reliable: if the relative error in a parameter is less than 10%, the confidence limits can be assessed within 10% using the normal approximation [16]. In the earlier study of ITC data, this rule was found to hold [8].

The more common approach to parameter variance estimation in physical science is the a posteriori one, where the fit itself is used to assess the data error. This $\mathbf{V}_{\text{post}}$,

$$\mathbf{V}_{\text{post}} = \frac{S}{\nu}\mathbf{A}^{-1}, \tag{3}$$

differs from $\mathbf{V}_{\text{prior}}$ only by the prefactor, where S is the sum of weighted squared residuals and ν is the degrees of freedom, equal to the number of data points minus the number of adjustable parameters. If the data errors σ_i are known a priori, S is an estimate of χ^2 for the fit and S/ν is an estimate of the reduced χ^2, χ^2_ν. For correct application of $\mathbf{V}_{\text{post}}$, the data errors must be known to within a scale factor. Then S/ν is distributed as a scaled χ^2 variate. Otherwise, $\mathbf{V}_{\text{post}}$ is simply wrong.[2] It is useful to note that χ^2_ν has average value 1 and variance $2/\nu$. This means that variances estimated by sampling have constant relative error (for given ν), hence absolute error proportional to their magnitude.

The MC calculations were done as described previously [8,9,16,17] by adding normally distributed random error of appropriate magnitude to the points on an exact curve and then fitting the synthetic data and accumulating statistical and distributional information as needed.

Generalized least squares

In GLS, the residuals δ_i from an LS fit to a model are themselves fitted to obtain an estimate of the variance function [11–13]. The residual δ_i is the difference between the measured and fitted (or calculated) values of the dependent variable, here the heat q_i. The residuals are closely related to sampling estimates of variance because the latter is estimated from a sum of the former's squares. Accordingly, the scale of the residuals contains information about the variance. The sum of weighted residuals is exactly zero in many fit models and is approximately zero in others. Thus, one must fit some power of the absolute value of the residuals. Standard approaches involve fitting either ln $|\delta|$ or δ^2. Like sampling estimates of variance, squared residuals have constant relative error (for fixed ν). Recalling that the fitting weights should be proportional to the inverse variance for the fitted quantity, we see that the log fit is properly an unweighted one, whereas the fit of δ^2 should employ weights proportional to the inverse square of the variance function.

These points can be clarified through an example involving one of the variance functions tried here:

$$\sigma^2_{\mathrm{E},i} = \sigma^2 g^2(\text{vars}, \boldsymbol{\theta}) = \sigma^2[1 + a^2 q_i^2 + b^2(q_i/v_i)^2]. \tag{4}$$

Here, $\boldsymbol{\theta}$ represents the adjustable parameters in the variance function (a and b in this case), and vars represents any dependences on the fit variables or other quantities. Thus, in this case $\boldsymbol{\theta}$ contains the two parameters, a and b, and the variables enter as the heat q and the titrant volume v. The quantity g^2 suffices for proper weighting in the analysis of the data ($w_i \propto g^{-2}$), but σ^2 must be included in the fits of the residuals (and may be included in w_i for the purpose of quantitative comparisons in the data fit). In principle, the computation of g^2 should employ the adjusted (estimated true) values of q_i (and v_i, although this quantity is treated as error free in the fit model), but here the relative errors are so small that this makes little difference. Thus, in one cycle, the data sets are fitted, yielding residuals that are then fitted to Eq. (4), yielding new weights to be used in the next fit of the data sets.

There is one potentially significant flaw in fitting the residuals as just described: the statistics of residuals do not actually replicate the inherent statistical error in the data. This is the problem associated with the common awareness that LS fits are particularly sensitive to the end points, and it can be very significant for small data sets, where many of the points will be fitted much better than implied by their inherent statistical error.[3] It can be shown [12] that the residuals have variance proportional to $1 - H_{ii}$, where $\mathbf{H}$ is commonly called the "hat" matrix and is here given by

$$\mathbf{H} = \mathbf{X}\mathbf{A}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}. \tag{5}$$

To correct for this effect, one uses in the variance function fit "studentized" residuals, taken as $\delta_i/(1 - H_{ii})^{1/2}$. There are other corrections that can be incorporated into residuals fitting, but this one appears to suffice in the current work, as illustrated below through the MC calculations.

Fit models

The model used to fit the data is the perfusion model used in the earlier work [7–10], with incorporation of one additional parameter to correct for backlash [18] and with correction for volume error [19]. In a perfusion system, injection of volume v_i from the syringe results in the expulsion of the same volume of material up the stem of the cell. It is assumed that (i) the system is uniform and at equilibrium prior to each injection and that

[2] This means, for example, that parameter errors returned from an unweighted fit algorithm are unreliable if the data are actually heteroscedastic.

[3] Consider, for example, a data set consisting of N points and p parameters. The expectation value of $\chi^2 = \Sigma(\delta_i/\sigma_i)^2$ is $\nu = N - p$, so as p approaches N, clearly some of the residuals δ_i must be much less than σ_i.

(ii) the expelled material is of this equilibrium composition and is not involved in either the mixing or the heat production in the active volume V_0 of the cell. Within this framework, the model fits to the incremental production of heat for each injection of titrant, no matter how large that injection. This is to be contrasted with a differential approximation that was employed in some early work [5].

I consider just the 1:1 complexation of titrant X with titrate M, and for the current work the titrant injection volume v_i is constant (v) for all injections that lead to analyzable residuals in the data for a given experiment. After the ith injection, the total concentrations (free and complexed) [7,8] of X and M are given by

$$[\mathrm{X}]_{0,i} = [\mathrm{X}]_0(1 - d^i) \quad \text{and} \quad [\mathrm{M}]_{0,i} = [\mathrm{M}]_0 d^i, \tag{6}$$

where $[\mathrm{X}]_0$ is the concentration of titrant in the syringe and $[\mathrm{M}]_0$ is the starting concentration of titrate M in the reaction vessel. The dilution factor $d = 1 - v/V_0$. At equilibrium, the concentrations of reactants and product satisfy the equilibrium expression,

$$\frac{[\mathrm{MX}]_i}{([\mathrm{X}]_{0,i} - [\mathrm{MX}]_i)([\mathrm{M}]_{0,i} - [\mathrm{MX}]_i)} = K \equiv K^{\circ} \times (\mathrm{L\ mol^{-1}}). \tag{7}$$

Thus, the number of moles of complex produced by the ith injection is

$$\begin{aligned}\Delta n_i &= V_0[\mathrm{MX}]_i - (V_0 - v)[\mathrm{MX}]_{i-1} \\ &= V_0([\mathrm{MX}]_i - d[\mathrm{MX}]_{i-1}),\end{aligned} \tag{8}$$

and the associated heat is

$$q_i = \Delta H^{\circ} \Delta n_i. \tag{9}$$

As before, I ignore experimental complications such as the need to estimate heats of dilution for the titrant and the related concentration dependence of q_i (although the latter effect is expected to be negligibly small for many systems at the very dilute concentrations typically employed in such studies).[4] I also use the dimensionless thermodynamic K° interchangeably with K, which is equivalent to assuming that all activity coefficients are unity at all times.

A key parameter in computations with this model is the extent to which the titration is "complete." The parameter used previously and again here to indicate the range of the titration is $R_m = [\mathrm{X}]_{0,m}/[\mathrm{M}]_{0,m}$, the ratio of total titrant to total titrate concentrations after the last (mth) injection.

The model just described has two adjustable parameters, ΔH° and K°, and as many data points as injections. The software general use includes a third parameter, the "site number" n. For well-understood 1:1 complexation, this parameter is simply a concentration correction factor needed to put the concentrations of X and M on a common footing. Its inclusion is important for achieving a good fit of typical ITC data, so it is included here (as before [8]) as a correction factor to $[\mathrm{M}]_0$. As already noted, the model used to analyze the experimental data includes a fourth parameter to correct for the backlash error [18]. This parameter is defined such that it always produces a residual of zero for the first ("throwaway") injection, which accordingly is omitted from the variance function fitting.

The variance functions examined in the GLS fitting are

$$\begin{aligned}
\sigma^2_{\mathrm{A},i} &= \sigma^2[1 + a^2 q_i^2] \\
\sigma^2_{\mathrm{B},i} &= \sigma^2[1 + a q_i]^2 \\
\sigma^2_{\mathrm{C},i} &= \sigma^2[1 + b^2 (q_i/v_i)^2] \\
\sigma^2_{\mathrm{D},i} &= \sigma^2[1 + b q_i/v_i]^2 \\
\sigma^2_{\mathrm{E},i} &= \sigma^2[1 + a^2 q_i^2 + b^2 (q_i/v_i)^2] \\
\sigma^2_{\mathrm{F},i} &= \sigma^2[1 + a q_i + b q_i/v_i]^2.
\end{aligned} \tag{10}$$

In general, models A, C, and E should be closer to physical reality because independent sources of experimental error add as variances rather than as contributions to σ. However, many experimental techniques also have identifiable error sources that give variances proportional to signal (e.g., from counting detectors) [20,21], so models B, D, and F might partially correct for such contributions, albeit in an "ignorant" way.

Results and discussion

Preliminary GLS test computations

The experiments in [10] involved 7 to 14 injections per run, with 10 injections being used in most runs. Preliminary GLS analysis of these data suggested strongly that the data error depended on q, but the dependence on titrate volume v predicted in [8] was less evident. Also, with such a small number of degrees of freedom (average ~6) for each data set, it was questionable how reliably the GLS analysis could hope to recover the true variance function. To address the latter problem, I chose to perform the GLS analysis on the pooled residuals from all 37 data sets under the assumption that all have the same underlying variance function. To check the feasibility of this approach, I did computations on a hypothetical 9-point data set representative of the experimental data (also having $v = 6$, because

[4] It can be shown that for the current system—complexation of Ba^{2+} with crown ether—the standard procedure of subtracting a blank of titrant into solvent does largely correct for the dilution effects. This procedure was followed in the work of [10], so remaining dilution effects are quite small.

the backlash correction was omitted here). The test employed variance function A, with $\sigma = 0.6$ μcal and $a = 0.00133$ μcal^{-1}, as obtained from an early GLS treatment of the data, fitting ln $|\delta_i|$. GLS analysis was performed on a synthetic data set consisting of 30 replicates of the reference set (270 residuals), fitting both ln $|\delta_i|$ and δ_i^2, for both raw and studentized residuals.

Fig. 1 illustrates the typical scale of the heteroscedasticity in this test model and also shows how significantly the fit residuals undershoot the actual data error in some regions. Not surprisingly, the use of studentized versus raw residuals makes a big difference in the outcome of the GLS analysis. The parameter a entirely determines the heteroscedasticity in this model. Fitting ln $|\delta_i|$ underestimated a by nearly a factor of 2 for raw residuals, rising to within 25% for studentized residuals. The latter result was still more than 1 apparent standard error low. (However, as discussed below, the error estimates are optimistic.) Weighted fits of δ_i^2 also fell short for raw residuals (0.0009(1) μcal^{-1}) but agreed well for studentized residuals ($1.42(14) \times 10^{-3}$ μcal^{-1}). From these results, I chose to analyze the actual data by fitting squared studentized residuals. Subsequent more extensive MC computations confirmed that this approach provided nearly bias-free estimates of the variance function parameters, as discussed below.

In a check on the significance of the underestimation of a from the GLS analysis of raw residuals, a full MC analysis (10^5 data sets) was done for data generated with the stated error structure but analyzed using $a = 0.0009$ μcal^{-1} to calculate the weights. Incorrect weighting must always lead to a loss of efficiency (larger variance). However, in this case the effect was modest—a 5% rise in variance for K and half that for ΔH.

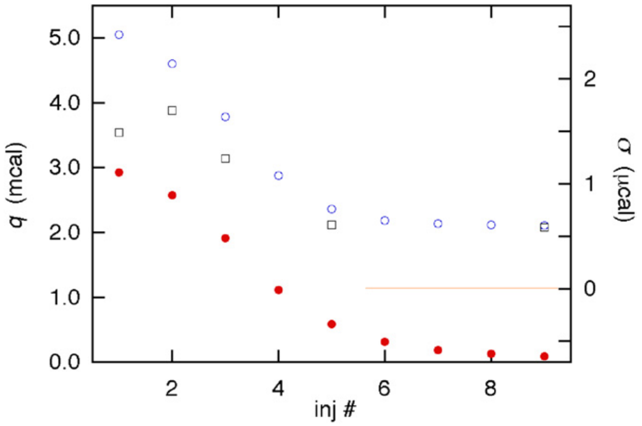

Fig. 1. ITC test model with $K^\circ = 10^4$ and $[M]_0 = 1.0$ mM (giving $K[M]_0 = 10$), $\Delta H^\circ = 7.7$ kcal/mol, and $R_m = 3$. Heats are shown as solid points, and standard deviations are shown as open points (scale to right): data σ, ○; rms δ, □. The latter were determined from statistics of 10^5 MC runs, and all were within 1% of predicted studentized residuals (Eq. (5)). The scale of the data error implies that data weights should span a factor of 10 in a properly weighted fit.

Variance function estimation

Initially, the GLS analysis was carried out for variance models A–D, each of which contains just one heteroscedasticity parameter. These analyses were straightforward and typically yielded convergence in four or five cycles. Results for models A and C are illustrated in Fig. 2. In both of these models, as well as in their counterpart models B and D, all parameters seemed comparably well defined, so the analysis was extended to include dependences in both $|q|$ and $|q|/v$, giving models E and F. Again the calculations converged adequately in a small number of cycles.

An important premise of the GLS analysis of pooled data is that all data sets are characterized by the same variance function. Accordingly, the 37 resulting values of S/v should be distributed roughly as χ_v^2 (roughly, in part due to varying numbers of data points, giving = 3–10). For all variance models, two of the data sets yielded S/v values exceeding 3, which is improbably large for their v values (both 6). When these two sets were removed, the analyses gave significantly larger

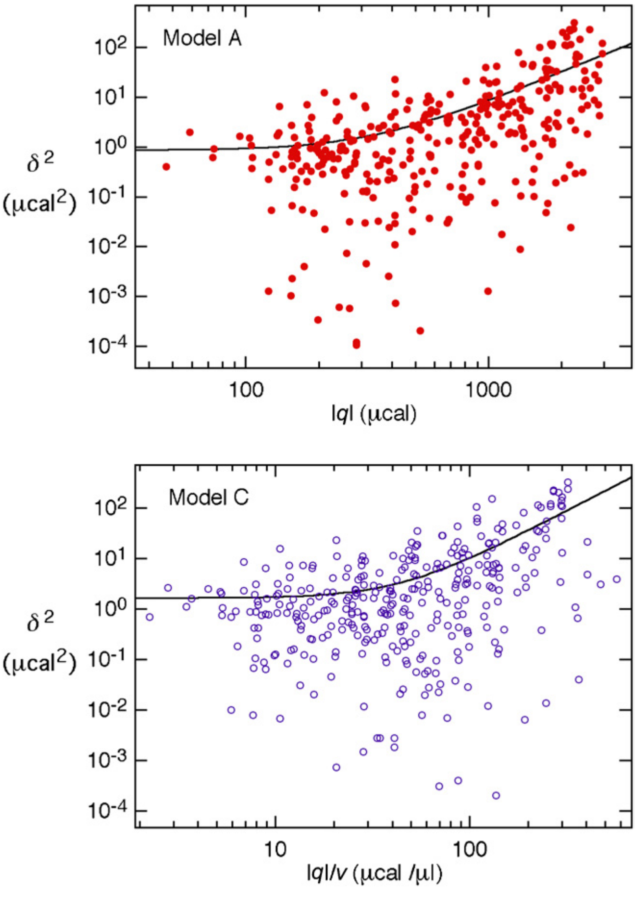

Fig. 2. Display of squared residuals from converged analyses of experimental data by GLS using variance models A (upper) and C (lower). The solid lines represent the results of the weighted LS fits to these variance functions. Note the logarithmic scales on both axes.

Table 1
Variance function parameters as estimated by GLS analysis of pooled data from 35 ITC experiments reported in [10]

Variance model	a(μcal^{-1})	b(μl/μcal)	σ(μcal)
A	0.00405(60)		0.710(84)
B	0.0070(18)		0.382(77)
C		0.0320(40)	1.011(87)
D		0.0349(64)	0.750(87)
E	0.00303(46)[a]	0.0200(44)[a]	0.772(79)[b]
F	0.0041(11)[b]	0.0159(65)[b]	0.458(77)[b]

Note. Variance functions are defined in Eq. (10). Figures in parentheses represent estimated standard errors from variance function fits in terms of final digits.

[a] From MC simulations of 1000 data sets, $a = 0.00309(67)$ and $b = 0.0201(69)$ (same units).

[b] From MC simulations of 1000 data sets, $a = 0.00433(168)$ and $b = 0.0173(97)$.

values for the heteroscedasticity parameters and smaller values for σ, and the S/v values now gave histogrammed distributions adequately consistent with expectations (typical range 0.08–2.6). Accordingly, the results from this subset of the original data (yielding 321 fitted residuals) are taken as the current best estimates of the variance function. These results are summarized in Table 1, where it can be seen that most parameters are determined with apparent percentage standard errors of less than 25%.

However, such parameter error estimates tend to be optimistic, because the residuals fit also depends on the values of the parameters in the data fit, even though the two fits are done separately [12]. It is also not immediately clear how to judge the performances of the different variance functions. For example, comparisons of the S values from the data fit are of no use because the very definition of the procedure is to make the average χ^2_ν from the data fit equal to the theoretical value of unity, no matter which variance function is fitted.[5] Correspondingly, the sum of the pooled δ_i^2/σ_i^2 values always equals the number of residuals (for raw or studentized δ_i, whichever are fitted). We might consider comparing two variance functions point-by-point, under the notion that if one function yields a smaller variance for every data point, it is better than the other. Of course the real situation is not so simple, because two variance functions will perform differently for different data points, making the first function better in some regions and the second better in others. From the statistics of variance ratios, models E and F seemed better than the simpler models. Variance ratios also suggested that model A was somewhat better than model C, and E was better than F. The quantitative significance of such comparisons was pursued further through MC simulations, as discussed below.

Monte Carlo "calibration"

The nominal parameter standard errors in Table 1 indicate that all of the fitted parameters are significant at the two-σ level. However, as just noted, these fit errors are optimistic. For a more realistic assessment, I used MC simulations based on the actual pooled data set. In these calculations, synthetic data were generated for an assumed variance function and then were subjected to GLS analysis using that variance function or a different one. The computations were accomplished using a code with automatic cycling between the data and variance function fits, where the convergence test for the latter was stability in the scale factor σ within one part in 10^4 (which sometimes meant much lower stability in a or b, especially when b tended toward zero).

The MC results from 1000 simulated data sets for models E and F are included as notes to Table 1. As expected, the MC sampling-based standard errors are significantly larger than those from the residuals fit **V** matrices, typically by approximately 50%. Still, all but b for model F remain well defined at the two-σ level. The MC averages of parameters show 6–9% positive bias for model F, meaning that the values for a and b in Table 1 for this model should be reduced by this amount to obtain "truer" estimates. For model E, the bias in a is statistically significant but practically insignificant, whereas the bias in b is not even statistically significant.

In a second experiment designed to check on the significance of the variance term in q/v, data were generated with variance function A but were then analyzed with variance function E. Of 100 synthetic data sets, 48 converged to $b = 0$ and only 4 gave values as large as those actually found for model E (Table 1). Thus, the statistical significance of b in model E is again supported.

As noted, the statistics of variance ratios seemed to favor model E over model F. To test this result, data were simulated with variance function E and submitted to GLS analysis with both models E and F. The variance function ratios (E/F and F/E) were then computed point by point and summed. An MC run of 100 data sets yielded a preference for model E only 60% of the time, indicating that the current data set is not capable of discriminating between these two models.

Implications for ITC experiments

With contributions to the data error strongly indicated for terms in both q and q/v, I next consider the consequences for the design of ITC experiments along the lines of my earlier considerations in [8]. To this end, I consider just model E, in which variances add, consis-

[5] The actual average values were typically 0.93. This means that the variance function fits are returning too-large estimates of the σ scale factors in Eq. (10), suggesting that the studentization procedure overcorrects the raw residuals.

tent with most physical models for experimental random error.

First, consider the term in q/v alone, arising from uncertainty in the titrant volume. Because the relative error in q from this source is equal to the relative error in v, we see that the estimated σ_v is 0.015 μl, which is exactly the estimate given long ago for the progenitor of the instrument used in the current experiments [5]. For the average v in the current data set (15.8 μl), the contribution to the variance is only one-sixth that from the a term. The two become comparable at $v = 6.6$ ml, which is smaller than the values employed in all but 3 of the 35 experiments. From this, it is clear why b is less well defined than a in the GLS analyses. This is also the reason why I have not considered here the other model for volume error, the correlated model.

At approximately 0.7 μcal, the base-level constant error is larger than that estimated previously by Wiseman et al. [5], even though the instrumentation has undergone significant sensitivity enhancement during the interim. The data used in the current analysis were recorded for relatively large q—large enough, in fact, that the instrument had to be set to its least sensitive scale for most experiments. It is possible that the σ estimates in Table 1 are effective parameters for this large q range and that they will drop further for experiments conducted at much lower q values. To some extent, all of the parameters are also subject to remaining uncertainties about the correct fit model for the analysis of the Ba^{2+} complexation with crown ether [10].

The direct term in q (the a term) was not anticipated in the earlier study [8] but is not unreasonable. This term can be seen as an error of scale—an error that remains at all scales of presentation after data "noise" is no longer evident—and it typically dominates experimental error in the strong-signal limit. In the current case, it probably arises from the procedures for estimating the background relative to which the incremental signal is then integrated over time to provide the q_i values. Regardless, this term is the dominant source of heteroscedasticity in the current data. Taken alone, it amounts to approximately 0.2% of q and equals the background constant term when $q = 300$ μcal. Conversely, because many reactions of biochemical interest involve q values much lower than this, heteroscedasticity will not be a problem in such studies, and the conventional unweighted LS analyses will suffice.

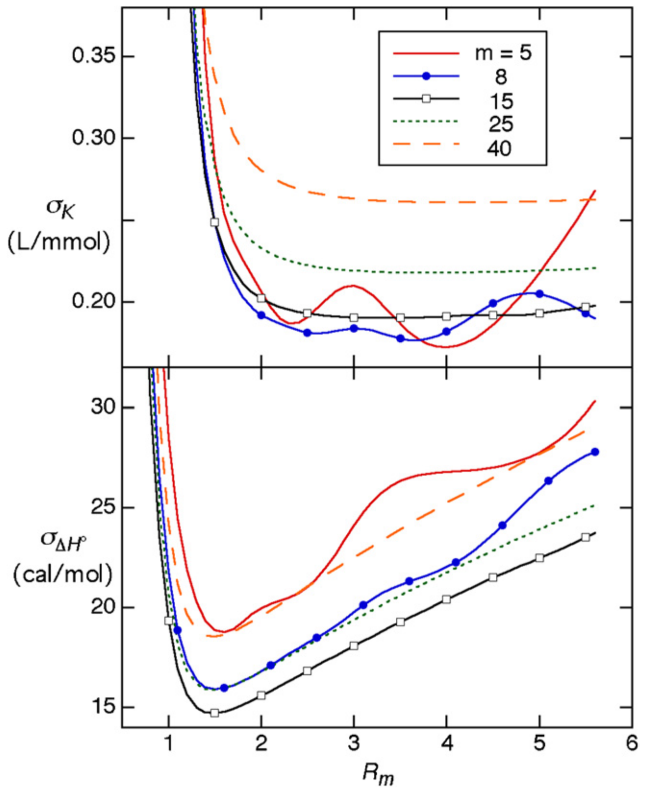

Fig. 3. Calculated standard errors (from $\mathbf{V}_{\text{prior}}$) in K and ΔH° for $\Delta H^\circ = 10$ kcal/mol, $K^\circ = 36{,}000$, and $[M]_0 = 1.0$ mM as functions of the stoichiometry range and the number of titration steps m for variance model E. Other parameters: $V_0 = 1.4$ ml and $v_m = 0.10$ ml. For reference, when $R_m = 3$ and $m = 15$, the heats q_i range from 2600 to 20 μcal and the errors range from 8.4 to 0.77 μcal. Because the computations are "exact," the structure in these curves is real—a consequence of the varying manners in which the titrations sample the structure inherent in the thermogram.

Fig. 3 shows the predicted standard errors in K and ΔH° for variance model E, as functions of the range of titration and number of injections, for typical K and ΔH° values. This is an updated version of Fig. 3 in [8] (which was computed for constant error) and shows some interesting differences. With inclusion of the a term, the precision first increases with increasing number of injections and then decreases with further increases in m, a behavior that is more pronounced for ΔH° than for K. In contrast, the aforementioned Fig. 3 from [8] predicted a monotonic decrease in precision with increasing m over most of the range of titration. This initially surprising result was recognized as a consequence of subdividing a constant supply of reaction enthalpy (the titrate) into smaller increments and was identified previously in a different context [22]. The scale error term is not subject to such limitations, so when this term dominates, more injections give better precision, in keeping with the usual expectations of improvement with more data points. However, increasing m still decreases each q_i, making the constant error term relatively more important. Thus, with inclusion of the scale error for a reaction with strong heat signature, the precision initially will improve with increasing m, giving an optimal m of approximately 8 to 12—with the smaller limit favored for K and the larger limit favored for ΔH°.

These dependences are further clarified in Figs. 4 and 5. In Fig. 4, the computations behind Fig. 3 are repeated for a factor of 10 smaller ΔH° and for a total titrant volume twice as large. Both of these changes make the error

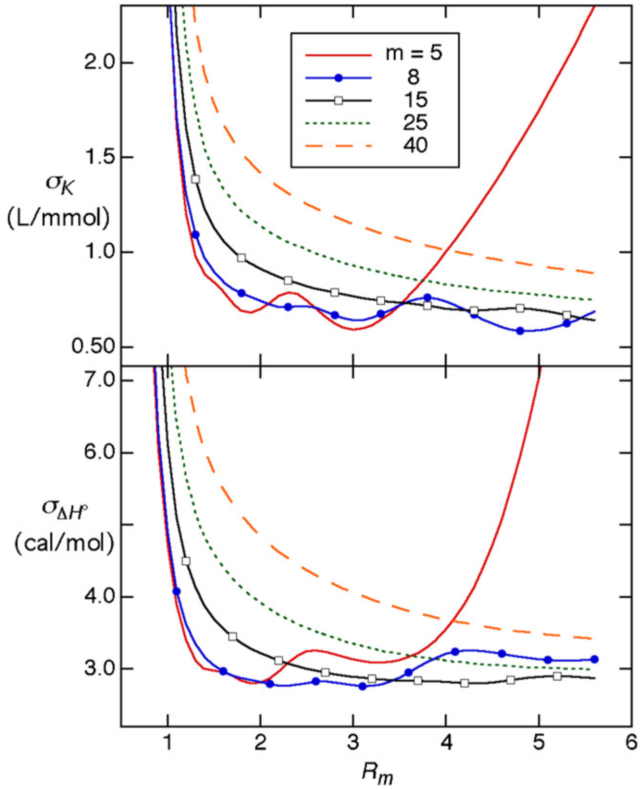

Fig. 4. Same as in Fig. 3, but wit $\Delta H^\circ = 1$ kcal/mol and $v \times m = 0.20$ ml.

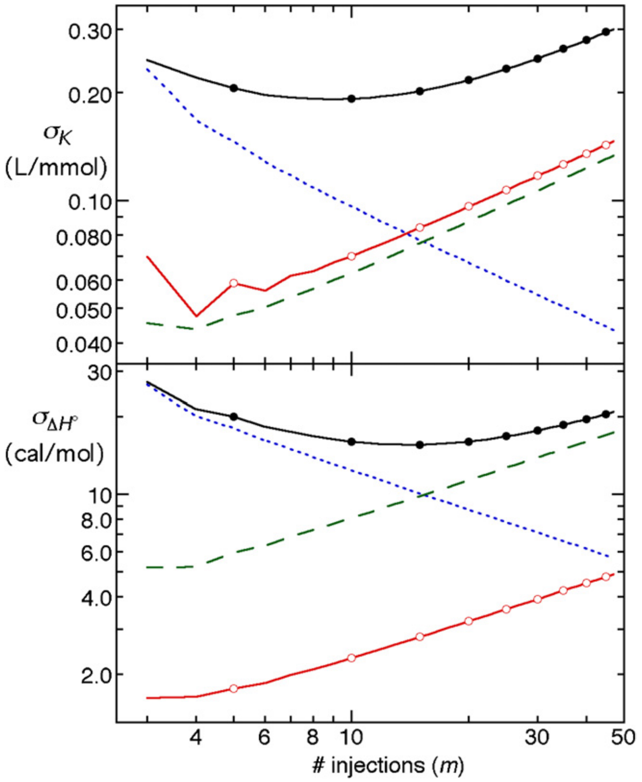

Fig. 5. Dependence of standard errors in K and ΔH° on m for $R_m = 2$ and other conditions as in Fig. 3. Also shown are the computed standard errors for each of the three error terms in isolation: constant error, solid curves with open points; scale error (aq), fine dash; volume error (bq/v), broad dash.

terms in q less significant. With the constant error term more important, the optimal number of injections has shifted down to 5–8 for both K and ΔH°, and the titration range $R_m = 2$–3 is reasonably close to optimal for both. In Fig. 5, the dependence on number of injections is shown for fixed R_m and the conditions of Fig. 3. The occurrence of minima in both σ values can be seen as a consequence of the playoff between increasing precision with increasing m for the scale error term and the opposite behavior for the other two terms.

It is noteworthy that in going from Fig. 3 to Fig. 4, the errors in K have increased, whereas those in ΔH° have decreased. One can show that for purely proportional error (the a and b terms), the relative errors in K and ΔH° are independent of the magnitude of ΔH° (for fixed K). On the other hand, for purely constant error, the relative errors in K and ΔH° are inversely proportional to ΔH°; thus, with constant error, dropping ΔH° from 10 to 1 kcal/mol leaves $\sigma_{\Delta H^\circ}$ unchanged and increases σ_K by a factor of 10. In combination, the proportional and constant errors lead to the noted changes in Fig. 4 from Fig. 3.

It is also worth emphasizing that the dependence on m in Figs. 3–5 is partly a consequence of the choice of fixed total titrant volume for all m. This means that to achieve the dependences illustrated in these figures, the initial titrant concentration $[X]_0$ must be made proportional to R_m in Figs. 3 and 4. For a given $[X]_0$ and fixed injection volume v, additional injections always improve the precision, as illustrated in Fig. 6.

Conclusion

GLS analysis of 35 ITC data sets for the 1:1 complexation of Ba^{2+} with crown ether has yielded the following main conclusions. First, for titrant injection volumes v greater than 10 μl, the dominant contribution to heteroscedasticity comes from a simple scale error rather than the titrant delivery volume error proposed in an earlier study [8]. Second, the proportional error terms are less than the constant error for heats q less than 300 μcal, so for heat-starved reactions (characterized by low ΔH° or very large K, forcing the use of low $[M]_0$), heteroscedasticity will not be important and analysis with the standard unweighted commercial LS codes will suffice. Volume error terms may become important for v less than 7 μl, a region not well sampled by the current data sets. Further work will be needed to characterize the volume error for low v, especially the possible need for the correlated analysis model. Also, it should be emphasized that the specific results obtained here apply only to the

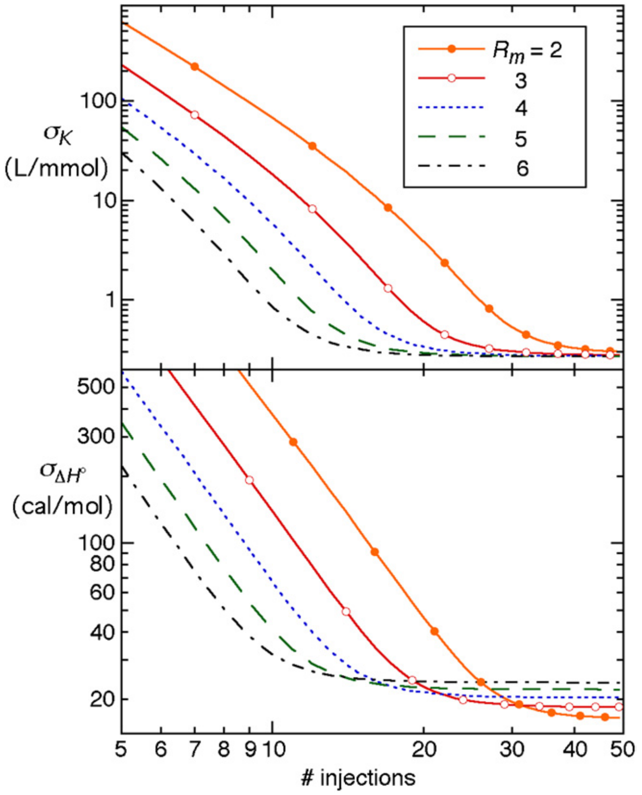

Fig. 6. Standard errors in K and ΔH° as functions of m for fixed titrant volume, $v = 3$ μl. Conditions are as in Fig. 3 except that the initial titrant concentration $[X]_0$ is set to yield the stated R_m after 49 injections and hence gives less complete reaction for smaller m.

MicroCal VPITC instrument used in the experimental studies [10]. However, one can expect the results to apply in a general way to other ITC instrumentation.

The estimated constant error from these studies is on the order 1 μcal, which is significantly larger than expected for today's instruments. However, the experimental data fall mostly in the range of $q = 100$–2000 μcal, so this is still only approximately 1% of the minimum q. As noted, it is likely that the estimated variance parameters in Table 1 are effective constants for the range of the data covered in the experiment and that data sampling much smaller q (e.g., 1–100 μcal) might yield different results. It seems unlikely that adding more parameters to the variance functions could help to answer this question without additional data, so no such efforts have been made.

Even with questions about the absolute validity of current estimates of the constant error, it does seem likely that this term will predominate in many practical biochemical studies, where ΔH° may be less than 10 kcal/mol and K° may be much greater than the 3.6×10^4 employed in Figs. 3–6. Thus, the important finding of the earlier study is substantiated: In most cases, the use of 10 or fewer injections not only suffices but also actually improves precision. This result is "win–win," because the use of fewer injections also permits ITC experiments to be completed in less time.

The computations behind Figs. 3–6 provide information about optimal conditions for the ITC study of 1:1 binding processes for a fixed K and a fixed starting titrate concentration $[M]_0$. However, the key operational quantities for a user designing an experiment are actually the products $K[M]_0$ (now often called c [5,23]) and $\Delta H^\circ[M]_0$. If the only data error is constant error, the precisions in K and ΔH° increase practically without limit as $[M]_0$ is increased. However, with increasing $[M]_0$, the experimental q_i increase, and eventually the proportional error dominates and limits the precisions. In addition, one may exceed instrumental limits at high q_i. The role of these interdependencies in the optimization of ITC experimental parameters is under continuing study, as is the use of varying titrant volumes v_i within a given experiment.

For relatively large-q reactions, such as the Ba^{2+}/crown ether complexation at the heart of the current study, one should use weights in LS analyses to properly account for the varying data error and thereby to extract the optimal estimates of the key parameters. To this end, it is useful to note that the standard ITC analysis package in the Origin program provided with ITC instruments from MicroCal does have a weighting option. This option is accessed by opening the Control box under the Options menu and then selecting one of the weighting modes. Several of these modes can be used to accomplish the weighted fit. For example, in the Instrumental mode, one enters values for σ_i for each q_i value in a new column of the data sheet, and the program computes the weights as $w_i = \sigma_i^{-2}$. In the spirit of the current work, these σ_i values would be calculated using the appropriate function (Eq. (10)). Alternatively, to investigate the effects of assuming proportional weight, one might enter value 1% of $|q_i|$ in this column (or 2 or 5%). Because these values are "known" in only a relative sense, one should then check the box labeled "Scale errors by square root of reduced χ^2." This changes the calculation of the parameter variances from $\mathbf{V}_{\text{prior}}$ (Eq. (2)) to $\mathbf{V}_{\text{post}}$ (Eq. (3)). (It also produces the same parameter standard errors for any choice of the percentage error.)

The current GLS analysis was enabled only through the use of pooled data sets under the assumption that all of the data from the various runs were characterized by the same variance function. It would not be feasible to use this approach for the analysis of a typical single ITC data set of 10–25 injections, because the number of LS residuals is too small. However, the results of the current study are at least consistent with the notion that day-to-day variations in experimental procedures and parameters can be accommodated by a common variance function for the method, so that pooled data sets might be similarly analyzed for other ITC problems and for other experimental techniques where procedures are repeated in routine fashion. This also supports the notion of "global" assessment of experimental statistical error over the "every day a new day" approach that is standard in analytical work. A big advantage of the global approach is the ability to use $\mathbf{V}_{\text{prior}}$ to assess the standard errors and thereby to avoid the large statistical error that is inherent in the estimation of χ^2 (and hence $\mathbf{V}_{\text{post}}$) from small data sets [24].[6]

Finally, I should emphasize that the current study has been entirely about random statistical error, whereas in practice systematic errors, whether recognized or unrecognized, often dominate the uncertainties in the derived results. In ITC work, one such concern is the proper treatment of heats of dilution as the relatively concentrated titrant is added to the more dilute mixture in the cell. Such concerns were, in fact, raised in the earlier study of the Ba^{2+}/ether complexation as a possible explanation for the persistent deviations between direct and van't Hoff estimates of the reaction ΔH as a function of temperature [10]. More careful examination of this case has shown that here, where both titrant (Ba^{2+}) and product (the complex) have the same charge and where the solutions are relatively dilute, the standard procedure of subtracting a blank obtained by injecting titrant into pure solvent actually does largely correct for such effects. Because solutions are typically quite dilute in ITC experiments, one can hope that dilution errors will be similarly small in other cases. However, even small molar concentrations of macromolecules can be effectively large for dilution effects, so such problems warrant vigilance and further attention.

References

[1] J.J. Christensen, R.M. Izatt, L.D. Hansen, New precision thermometric titration calorimeter, Rev. Sci. Instrum. 36 (1965) 779–783.

[2] N.V. Beaudette, N. Langerman, Improved method for obtaining thermal titration curves using micromolar quantities of protein, Anal. Biochem. 90 (1978) 693–704.

[3] R.B. Spokane, S.J. Gill, Titration microcalorimeter using nanomolar quantities of reactants, Rev. Sci. Instrum. 52 (1981) 1728–1733.

[4] G. Ramsay, R. Prabhu, E. Freire, Direct measurement of the energetics of association between myelin basic-protein and phosphatidylserine vesicles, Biochemistry 25 (1986) 2265–2270.

[5] T. Wiseman, S. Williston, J.F. Brandts, L.-N. Lin, Rapid measurement of binding constants and heats of binding using a new titration calorimeter, Anal. Biochem. 179 (1989) 131–137.

[6] M. El Harrous, S.J. Gill, A. Parody-Morreale, Description of a new Gill titration calorimeter for the study of biochemical reactions: I. Assembly and basic response of the instrument, Meas. Sci. Technol. 5 (1994) 1065–1070.

[7] M. El Harrous, O.L. Mayorga, A. Parody-Morreale, Description of a new Gill titration calorimeter for the study of biochemical reactions: II. Operational characterization of the instrument, Meas. Sci. Technol. 5 (1994) 1071–1077.

[8] J. Tellinghuisen, A study of statistical error in isothermal titration calorimetry, Anal. Biochem. 321 (2003) 79–88.

[9] J. Tellinghuisen, Statistical error in isothermal titration calorimetry, Methods Enzymol. 383 (2004) 245–282.

[10] L.S. Mizoue, J. Tellinghuisen, Calorimetric vs. van't Hoff binding enthalpies from isothermal titration calorimetry: Ba^{2+}—crown ether complexation, Biophys. Chem. 110 (2004) 15–24.

[11] M. Davidian, R.J. Carroll, Variance function estimation, J. Am. Stat. Assoc. 82 (1987) 1079–1091.

[12] R.J. Carroll, D. Ruppert, Transformation and Weighting in Regression, Chapman & Hall, New York, 1988.

[13] M. Davidian, P.D. Haaland, Regression and calibration with nonconstant error variance, Chemometr. Intel. Lab. Sys. 9 (1990) 231–248.

[14] P.R. Bevington, Data Reduction and Error Analysis for the Physical Sciences, McGraw–Hill, New York, 1969.

[15] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, Numerical Recipes, Cambridge University Press, Cambridge, UK, 1986.

[16] J. Tellinghuisen, A Monte Carlo study of precision, bias, inconsistency, and non-Gaussian distributions in nonlinear least squares, J. Phys. Chem. A 104 (2000) 2834–2844.

[17] J. Tellinghuisen, Bias and inconsistency in linear regression, J. Phys. Chem. A 104 (2000) 11829–11835.

[18] L.S. Mizoue, J. Tellinghuisen, The role of backlash in the "first injection anomaly" in isothermal titration calorimetry, Anal. Biochem. 326 (2004) 125–127.

[19] J. Tellinghuisen, Volume errors in isothermal titration calorimetry, Anal. Biochem. 333 (2004) 405–406.

[20] L.D. Rothman, S.R. Crouch, J.D. Ingle Jr., Theoretical and experimental investigation of factors affecting precision in molecular absorption spectrophotometry, Anal. Chem. 47 (1975) 1226–1233.

[21] J. Tellinghuisen, Statistical error calibration in UV–visible spectrophotometry, Appl. Spectrosc. 54 (2000) 431–437.

[22] M.L. Doyle, J.H. Simmons, S.J. Gill, Analysis of parameter resolution from derivatives of binding isotherms, Biopolymers 29 (1990) 1129–1135.

[23] W.B. Turnbull, A.H. Daranas, On the value of *c*: can low affinity systems be studied by isothermal titration calorimetry? J. Am. Chem. Soc. 125 (2003) 14859–14866.

[24] J. Tellinghuisen, Simple algorithms for nonlinear calibration by the classical and standard addition methods, Analyst 130 (2005) 370–378.

[6] The resulting relative standard deviation in the estimated parameter errors is $(2\nu)^{-1/2}$. This means, for example, that in replications of an 11-injection ITC experiment, one can expect the $\mathbf{V}_{\text{post}}$-based parameter errors to vary by approximately 25% from run to run.