Eur Biophys J (2010) 39:1261–1275
DOI 10.1007/s00249-009-0545-7

BIOPHYSICS LETTER

On computational approaches for size-and-shape distributions from sedimentation velocity analytical ultracentrifugation

Peter Schuck

Received: 3 August 2009 / Revised: 8 September 2009 / Accepted: 14 September 2009 / Published online: 6 October 2009

Abstract Sedimentation velocity analytical ultracentrifugation has become a very popular technique to study size distributions and interactions of macromolecules. Recently, a method termed two-dimensional spectrum analysis (2DSA) for the determination of size-and-shape distributions was described by Demeler and colleagues (Eur Biophys J 2009). It is based on novel ideas conceived for fitting the integral equations of the size-and-shape distribution to experimental data, illustrated with an example but provided without proof of the principle of the algorithm. In the present work, we examine the 2DSA algorithm by comparison with the mathematical reference frame and simple well-known numerical concepts for solving Fredholm integral equations, and test the key assumptions underlying the 2DSA method in an example application. While the 2DSA appears computationally excessively wasteful, key elements also appear to be in conflict with mathematical results. This raises doubts about the correctness of the results from 2DSA analysis.

Keywords Analytical ultracentrifugation · Fredholm integral equations · 2DSA · Sedimentation velocity · Lamm equation · Size distribution

P. Schuck (✉)
Dynamics of Macromolecular Assembly,
Laboratory of Bioengineering and Physical Science,
National Institutes of Biomedical Imaging and Bioengineering,
National Institutes of Health, Bethesda, MD, USA
e-mail: schuckp@mail.nih.gov

Introduction

The use of sedimentation velocity analytical ultracentrifugation (SV) has significantly expanded in the last decade (Howlett et al. 2006; Scott and Schuck 2006; Cole et al. 2008), and new computational methods for SV analysis are being actively developed by several groups (Balbo et al. 2005; Philo 2006; Brown et al. 2007, 2009; Behlke and Ristau 2009; Brookes et al. 2009; Correia and Stafford 2009). In particular, diffusion-deconvoluted sedimentation coefficient distributions calculated from direct boundary modeling of experimental data (Schuck 2000; Schuck et al. 2002) have proven to be very useful tools in many biophysical applications (for a list of references see Schuck 2007). They can achieve relatively high hydrodynamic resolution of pauci- and polydisperse macromolecular mixtures, exhibit exquisite sensitivity for trace components (Berkowitz 2006; Liu et al. 2006; Brown et al. 2008a, b; Gabrielson et al. 2009), and can be related to sedimentation coefficient isotherms and Gilbert–Jenkins theory for the analysis of slowly or rapidly interacting systems (Dam and Schuck 2005; Dam et al. 2005). The extension of sedimentation coefficient distributions to two-dimensional size-and-shape distributions was introduced (Schuck 2002; Brown and Schuck 2006) and applied in numerous studies (Markossian et al. 2006; Chang et al. 2007; Deng et al. 2007; Race et al. 2007; Broomell et al. 2008; Brown et al. 2008; Chebotareva et al. 2008; Iseli et al. 2008; Moncrieffe et al. 2008; Paz et al. 2008; Sivakolundu et al. 2008; Wang et al. 2008; Eronina et al. 2009; Mortuza et al. 2009). More recently, the Demeler laboratory has described the concept of a novel algorithm ("2DSA") for determining size-and-shape distributions, as implemented in the software ULTRASCAN (Brookes et al. 2006, 2009; Demeler et al. 2009). In the present work, we critically

compare the background of the different algorithms and assess their performance.

Methods

The SV experiment was carried out with a Beckman-Coulter XL-I analytical ultracentrifuge, following standard protocols as described by Brown et al. (2008a, b). A monoclonal immunoglobulin G (IgG) preparation in phosphate-buffered saline (PBS) buffer was inserted in 12-mm Epon centerpieces, temperature equilibrated at 18°C, and then accelerated to 45,000 rpm and scanned with absorbance optics at 280 nm. Data analysis was performed with SEDFIT 11.8 using $c(s)$ models as described by Schuck et al. (2002), the two-dimensional size-and-shape model $c(s, f_r)$ as described by Brown and Schuck (2006), and applying Bayesian prior knowledge as described in detail by Brown et al. (2007). The computer used for these analyses was a Dell Precision T5400 workstation, with dual 32-bit quadcore 3.16-MHz processors and Windows operating system.[1]

Outline of the algorithms

For clarity of the analysis of the algorithms, we first provide a mathematical outline of the problem. This is followed by a more detailed discussion of appropriate discretization parameters, and from this we derive the demands on the computational platforms. Then we discuss algorithmic aspects for calculating Lamm equation solutions and for computing a size-and-shape distribution from the experimental data, and finally comment on methods for estimating their true information content.

Mathematical description of the problem

The size-and-shape distribution problem is a Fredholm integral equation of the form

$$a(r,\,t) = \int_{s_{\min}}^{s_{\max}} \int_{f_{r,\,min}}^{f_{r,\,max}} c(s, f_r)\,\chi(s, f_r,\, r,\, t)\mathrm{d}s\mathrm{d}f_r, \qquad (1)$$

where the data $a(r,\,t)$ are the measured evolution of the radial signal profiles, and $c(s, f_r)$ is a differential size-and-shape distribution, expressed most conveniently for the modeling of SV data in coordinates of sedimentation coefficient s and frictional ratio f_r (Brown and Schuck 2006). $\chi(s, f_r,\, r,\, t)$ are normalized solutions of the Lamm equation (Lamm 1929)

$$\frac{\partial \chi}{\partial t} = \frac{1}{r}\frac{\partial}{\partial r}\left[Dr\frac{\partial \chi}{\partial r} - s\omega^2 r^2 \chi\right], \qquad (2)$$

which predicts the evolution of the concentration profiles of an ideally sedimenting species with sedimentation coefficient s and diffusion coefficient $D(s, f_r)$ that is initially uniformly distributed between the meniscus and bottom of the solution column at loading concentration of 1.

Equation (1) can be discretized on a rectangular grid with $(S \times F)$ size-and-shape values $(s_i, f_{r,j})$ comprising all combinations of S equidistant sedimentation coefficient values from $s_1 = s_{\min}$ to $s_S = s_{\max}$ (with constant mesh size $\Delta s = (s_S - s_1)/(S-1) = s_{i+1} - s_i$), and F frictional ratio values from $f_{r,1} = f_{r,\min}$ to $f_{r,F} = f_{r,\max}$ (with constant mesh size $\Delta f_r = (f_{r,F} - f_{r,1})/(F-1) = f_{r,j+1} - f_{r,j}$). With the data being $(N \times M)$ discrete signal values at radius r_n and time t_m, abbreviated as a_{nm}, (1) leads to the linear least-squares problem

$$\underset{c_{i,j}\geq 0}{\mathrm{Min}} \sum_{n,m} \left(a_{nm} - \sum_{i=1}^{S}\sum_{j=1}^{F} c_{i,j}\chi(s_i, f_{r,j}, r_n, t_m) - b(r_n) - \beta(t_m) \right)^2. \qquad (3)$$

The $c_{i,j}$ provide an estimate of the size-and-shape distribution with $c(s,f_r) \approx c_{i,j}/(\Delta s \Delta f_r)$. Signal offsets from systematic time-invariant $[b(r_n)]$ and radial-invariant $[\beta(t_m)]$ noise contributions are indicated in Eq. (3), but their simultaneous optimization with the method of separation of linear and nonlinear parameters (Ruhe and Wedin 1980) poses no significant further complications (Schuck and Demeler 1999) and therefore they will be dropped from further consideration in order to make the notation more transparent in the following.[2]

We can introduce a new index l that lexicographically orders all data points (a total of $L = N \times M$), and a single index k that enumerates all size-and-shape grid points $(s_i, f_{r,j})$ from 1 to $K = S \times F$, which allows us to write (3) as a simple sum

[1] We also analyzed the data with ULTRASCAN II version 9.9 to confirm our results as far as possible. Unfortunately, the current lack of a manual section for the use of the 2DSA analysis and the excessive computational times involved prevented us from a direct comparative analysis of the same data with the full 2DSA model as described by Brookes et al. (2009). Further, a detailed comparison does not seem possible due to seemingly unavoidable data truncation steps when loading data in ULTRASCAN II, and due to our inability to write the entire calculated distribution into a text file.

[2] They cannot, however, be calculated in a first analysis and then be subtracted from the experimental data, as described by Demeler and colleagues (Brookes et al. 2009). Since systematic noise components are part of the model, and since their estimates can correlate with the description of the macromolecular sedimentation distribution, they need to be simultaneously optimized (Schuck and Demeler 1999; Dam and Schuck 2004).

$$\underset{c_k \geq 0}{\text{Min}} \sum_l \left(a_l - \sum_{k=1}^{K} c_k \chi_{kl} \right)^2 . \quad (4)$$

This highlights the nature of the problem being a standard nonnegative linear least-squares problem. The unconstrained problem can be solved with the method of normal equations (Lawson and Hanson 1974; Golub and VanLoan 1989)

$$\mathbf{P}\vec{\mathbf{c}} = \vec{\mathbf{d}}, \quad (5)$$

with the $K \times K$ matrix $\mathbf{P}$ (sometimes referred to as the Gram matrix) with elements $P_{\kappa\lambda} = \sum_l \chi_{\kappa l}\chi_{\lambda l}$, the $K \times 1$ vector $\vec{\mathbf{d}}$ with elements $d_k = \sum_l a_l \chi_{kl}$, and the $K \times 1$ vector $\vec{\mathbf{c}}$ representing the unknown distribution.

The unique best-fit solution with the nonnegativity constraint $c_k \geq 0$ can be found unambiguously with the algebraic algorithm NNLS, which was introduced and proven by Lawson and Hanson (1974). We first used the NNLS algorithm in the context of SV distribution analysis, in a form where we expressed all requisite quantities with elements of the normal equations (Schuck 2000). NNLS is an active set algorithm that divides the unknowns into sets with active ($c_k = 0$) and inactive ($c_k > 0$) inequalities, and iteratively establishes the active set producing the best-fit solution. For the inactive set, the problem takes the same form as (5), but with all matrix and vector elements from components with active constraints deleted (Gill et al. 1986).

Frequently the problem of fitting distributions of the form (1) is ill posed, meaning that many different solutions will fit the data statistically indistinguishably well (Louis 1989; Hansen 1998; Engl et al. 2000). For example, Provencher (1982) has illustrated this point via the Lemma of Riemann–Lebesgue, showing that one should expect a large set of very different solutions to fit the data equally well within the experimental error. In practice, noise of the data can amplify to determine even the overall features of the best-fit solution $\vec{\mathbf{c}}$, and often the strictly best-fit solution consists of a series of spikes whose number, location, and height may not reflect the presence of such species in the physical experiment, but are governed by the details of the noise and other imperfections in the data.

It is therefore desirable to suppress, among all possible solutions, those that contain a potentially misleading amount of detail arising from noise amplification. Towards this goal, regularization is a standard approach that determines the most parsimonious solution of all that fit the data statistically indistinguishably well. It minimizes a measure of the information content of the solution while optimizing the quality of fit. A well-known and widely applied strategy to suppress artificial spikes is Tikhonov–Phillips regularization (Phillips 1962; Provencher 1982; Louis 1989; Hansen 1992; Press et al. 1992), which uses, for example, the square of the second-derivative matrix ($H_{k\kappa}$) to stabilize the solution of (4):

$$\underset{c_k \geq 0}{\text{Min}} \left[\sum_l \left(a_l - \sum_{k=1}^{K} c_k \chi_{kl} \right)^2 + \alpha \sum_{k,\kappa} c_k H_{k\kappa} c_\kappa \right] \quad (6)$$

or, formulated with normal equations,

$$(\mathbf{P} + \alpha\mathbf{H})\vec{\mathbf{c}} = \vec{\mathbf{d}}, \quad (7)$$

where α is a parameter that scales the regularization constraint (Louis 1989; Press et al. 1992). Again, (7) has an unambiguous best-fit solution that can be determined algebraically with NNLS for any value of α, and the latter can be adjusted in a simple one-dimensional search such that the least-squares fit remains at a statistically indistinguishable quality compared with the initial best fit in the absence of regularization (Bevington and Robinson 1992). A Bayesian variation of this approach is possible that modulates the regularization matrix to enhance the information content of the solution in view of existing (or hypothesized) prior knowledge (Sivia 1996; Brown et al. 2007; Patel et al. 2008).

We will refer to this approach as the "standard algorithm," because it is firmly rooted in textbook linear algebra and basic linear least-squares optimization, and utilized in many applications throughout the biophysical literature and physical sciences. We have introduced this approach previously into the SV analysis, and it underlies all size-distribution analyses in SEDFIT and SEDPHAT. If used without regularization, it provides exact solutions (within numerical precision) to the least-squares problem (3), and when used with regularization, the algorithms ensure that fits with statistically indistinguishable quality are obtained.

The 2DSA method by Demeler and colleagues aims to solve the same Eqs. (1), (3), and (4), respectively. This is described by Brookes et al. (2009), and with less mathematical detail by Demeler et al. (2009). The Demeler approach deviates in key aspects from the strategies described above. Apparently in order to circumvent perceived computational limitations, a novel multigrid scheme is conceived that would allow a sequence of fits with low-resolution 10×10 ($S \times F$) grids to approximate the solution of (1) and (3) with high-resolution $S \gg 10$ and $F \gg 10$. For achieving parsimonious results Monte Carlo iterations are applied (Brookes et al. 2009). Some of the key ideas will be discussed in the following.

Appropriate mesh sizes for the two-dimensional problem

First, in order to assess the size of the problem and computational requirements, we need to clarify how fine the

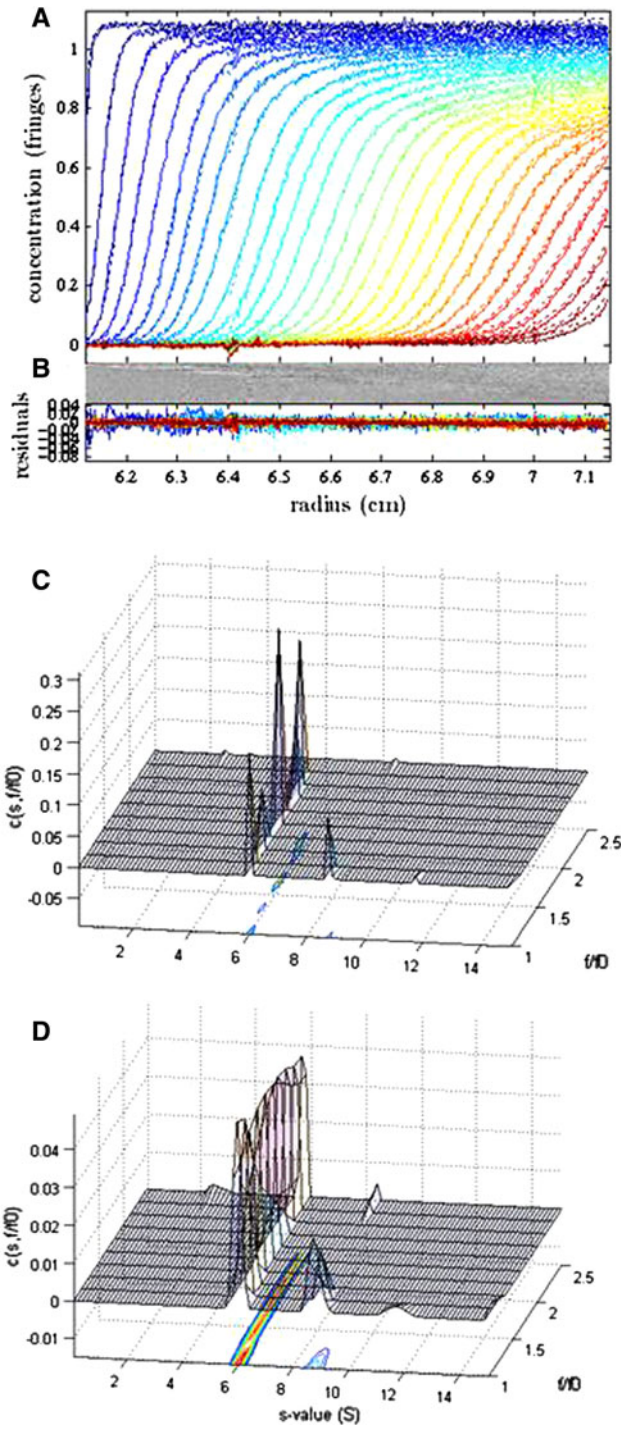


Fig. 1 Illustration of the standard algorithm for size-and-shape distributions applied to the experimental data of an immunoglobulin G sample, sedimenting at 45,000 rpm. **a** Experimental data acquired with the absorbance optical system (*solid lines*). The color temperature indicates the temporal order of the scans, with blue for the early and red for the late scans. The dotted lines, virtually overlapping the experimental data, are the best-fit distribution from modeling with Eq. (3) for a grid of $K = 10 \times 100$ (f_r, s)-values ranging from f_r-values of 1.0 to 2.5 in 10 equidistant steps, and from s-values of 0.1 to 15.0 S in 100 equidistant steps. **b** Residuals of the fit, presented as a bitmap (Dam and Schuck 2004) and as an overlay plot for all traces. The root-mean-square deviation is 0.00672 OD. **c** Raw size-and-shape distribution without regularization. As in (Brown and Schuck 2006), the 2D grid of (f_r, s)-values is indicated by *solid lines*, combined with a color temperature contour map in the *plane below*. The solution is a series of spikes in f_r-dimension, with a comparatively well-defined s-value of $\sim$5.8 S for the main species. An observation familiar in the study of IgG (and many other protein) samples is the low-level population of dimeric species at $\sim$9 S, as well as trimeric traces at $\sim$11–12 S. **d** Tikhonov–Phillips regularization applied to produce the most parsimonious size-and-shape distribution of all that fit the data statistically indistinguishably well at a $P = 0.95$ confidence level (i.e., that produce a rmsd value of 0.00677 OD or better)

grid of s-values and f_r-values needs to be in order to fully extract all information from a typical set of sedimentation velocity data. Let us consider as an example the experimental data from a preparation of IgG molecules sedimenting at 45,000 rpm, as shown in Fig. 1a. It is useful to start the analysis with a one-dimensional sedimentation coefficient distribution analysis $c(s)$, since the sedimentation coefficients are the experimentally best determined quantities. $c(s)$ eliminates the shape dimension by using hydrodynamic scaling laws such as the traditional $s \sim M^{2/3}$ law for globular particles (Schuck 2000), theoretical models for wormlike chains (Yamakawa and Fujii 1973) or any user-defined exponential scaling laws for polymers (Pavlov et al. 2009). For the given data we can determine from the $c(s)$ analysis (not shown) that s-values from 0.1 to 15 S will be sufficient to describe all sedimenting species. Equidistant discretizations with $S = 100$ or $S = 200$ lead to statistically indistinguishable quality of fit, as measured by F-statistics (Bevington and Robinson 1992; Straume and Johnson 1992), and therefore we preliminarily conclude that $S = 100$ will be a reasonable choice.

Typically, the resolution in the frictional ratio dimension cannot be expected to be very high, even in combination with data from SV experiments at a range of rotor speeds (Schuck 2002). Therefore, a discretization providing $F = 10$ values between 1.0 and 2.5 (ranging from extremely compact to very extended protein structures) should be a sufficiently flexible basis to describe the actual frictional ratio for each species (knowing that we have inserted folded proteins into the sample solution, and keeping in mind the average frictional ratio of 1.68 estimated from the $c(s)$ analysis). The resulting 10×100 grid with a total of $K = 1{,}000$ species was fitted with the standard algorithm to the data in Fig. 1a, leading to virtually random distribution of residuals (1b), with a root-mean-square deviation (rmsd) of 0.00672, consistent with the noise in the data acquisition. The resulting distribution is shown with and without regularization in Fig. 1d and c, respectively. As expected, while the s-values of the species are well defined, the shape dimension is highly underdetermined, resulting in the

typical series of peaks in 1c, and in a smooth relatively uniform distribution after regularization in 1d. (This justifies, in retrospect, the choice of $F = 10$ values as a sufficiently detailed discretization of the frictional ratio dimension.)

We can compare the rmsd achieved with this 10 × 100 grid (0.00672) with a fit under otherwise identical conditions but different grids: a coarser 10 × 50 grid leads to an rmsd of 0.00678, which is barely statistically worse (on a one standard deviation confidence level), and a finer grid with 20 × 200 grid leads to an rmsd of 0.00670, which is statistically indistinguishable. Thus, a 10 × 100 grid is of sufficiently high resolution to extract the entire information content of the experimental data.

Memory requirements and computational platforms

After outlining the structure of the problem and the discretization parameters typically required for a size-and-shape analysis of SV data, it is possible to discuss the computational requirements. Demeler's 2DSA method was implemented with the goal of accessing a high-performance computing environment (TeraGrid) in order to avoid prohibitive memory limitations that Demeler and colleagues perceive to occur when using ubiquitous ordinary laboratory workstations. Specifically, the authors (Brookes et al. 2009) estimate the memory needs for modeling a set of $M = 50$–100 sedimentation velocity scans with typically $N = 500$–800 data points each by only a low-resolution 10 × 10 ($S \times F$) grid. They conclude that "Performing just a 10 × 10 grid search on such an array would require close to half a gigabyte of memory just for data storage of a single experiment." (Brookes et al. 2009). We will examine this estimate in more detail.

In practice, when using the absorbance optics with the recommended and widely applied setting of 0.003 cm (Brown et al. 2008a, b) for the radial intervals, in order to diminish errors from sample migration during the scan (Brown et al. 2009), we obtain only on the order of ∼200–250 points per scan in a long-column SV experiment. In typical high-speed SV experiments with eight-hole rotors, we can acquire usually only 50 scans or fewer before depletion occurs and/or migration and backdiffusion approach steady state, even with small solutes. This is sufficient for a highly detailed analysis of multicomponent systems, as discussed by Balbo et al. (2005). Predicted values $\chi(s_i, f_{r,\,j}, r_n, t_m)$ need to be calculated for each species $(s_i, f_{r,\,j})$ with arrays of the same size as the data $a(r_n, t_m)$. Since the experimental data have a precision not better than four decimal places, their representation as a standard 32-bit floating-point data type with eight significant figures is already wasteful. Nevertheless, calculating conservatively with 32-bit floats we arrive at a memory requirement of only ∼4.8 MB for storage of model data, rather than 0.5 GB [250 × 50 × 10 × 10 × 4 bytes × (1,048,576 bytes/MB)$^{-1}$ = 4.76 MB]. With interference optical (IF) data, the native radial density of points is higher (∼1,500 per scan). Since the radial density of points of interference scans is not exploited experimentally, it could be safely reduced to the level of absorbance data by preaveraging, which reduces the statistical noise approximately by a factor of 2. However, again calculating conservatively and using the native resolution of IF data, this would lead to ∼28 MB storage space, or ∼57 MB if 100 scans were used to represent the evolution in a SV experiment.

We find that the ∼5–50 MB actually required for calculating size-and-shape distributions with 10 × 10 grids is compatible with the available memory on many different platforms, ranging from >200,000 MB available on TeraGrid systems, to ∼2,000–3,000 MB typically available on 32-bit Windows, and even the ∼50–90 MB available on current smartphones.

Consistent with this result, we and others (Markossian et al. 2006; Chang et al. 2007; Deng et al. 2007; Race et al. 2007; Broomell et al. 2008; Brown et al. 2008; Chebotareva et al. 2008; Iseli et al. 2008; Moncrieffe et al. 2008; Paz et al. 2008; Sivakolundu et al. 2008; Wang et al. 2008; Eronina et al. 2009; Mortuza et al. 2009) have regularly used full high-resolution grids (such as 10 × 50, 10 × 100, or higher) in SEDFIT on ordinary personal computers or laptops, an exercise that is a regular part of the Workshop on Hydrodynamic and Thermodynamic Analysis of Macromolecules with SEDFIT and SEDPHAT at the National Institutes of Health (Schuck 2009). This is possible due to the fact that the memory requirement for the high-resolution grid would be 48–286 MB to store the model data (assuming 50 scans for data absorbance or native interference data modeled with a 10 × 100 grid). It is readily verified that, even for the complete high-resolution grid and when globally analyzing many experimental data sets, this is well below the memory limit of currently common 32-bit Windows operating systems.

Further, all computations can be condensed to the normal Eq. (5), requiring essentially only a matrix $\mathbf{P}$ of 1,000 × 1,000 numbers to be operated on, which even as double-precision data type requires less than 8 MB, trivial by current standards on any platform. Once condensed to the form of Eq. (5), our SV problem is far smaller (often several orders of magnitude) than common problems of analogous mathematical structure, for example, in astronomical image analysis (Narayan and Nityananda 1986). For the data shown in Fig. 1, in the implementation in SEDFIT (which does not optimize memory allocation), ∼20 MB of RAM are used.

The necessary computational power will strongly depend on the implementation of the algorithms, of course. Parallelization can be readily achieved in the standard

algorithm, which can in many places decrease the computation time by a factor virtually proportional to the number of threads. This is true, for example, for solving the Lamm equations, and for the most time-consuming step of building the normal equations matrix. The time for a complete calculation with a full high-resolution grid (10 × 100) for the data shown in Fig. 1, on a current dual-processor quadcore 3.16-MHz PC (Dell Precision T5400), is only 42 s.[3] The time required for a 10 × 50 grid, which we have already seen leads to an adequate fit within the noise of data acquisition, is 10 s. Finally, it is ∼15 min for a 20 × 200 grid. In the standard algorithm a Monte Carlo statistical analysis may be desired, for example, in order to examine the statistical accuracy of a particular species population as determined from the integration of the distribution in a certain range. In the standard algorithm, each iteration requires only updating the vector $\vec{\mathbf{d}}$ of the normal Eq. (5) and solving these equations. For the data shown in Fig. 1, one iteration takes ∼3 s on a single thread on our PC.

We conclude that ordinary current workstations do not pose a limitation for rigorously determining the size-and-shape distributions, neither with regard to available memory, nor with regard to processor capabilities.

Lamm equation solutions

Modeling a distribution of species with different size and shape to the data depends critically on the accuracy of the Lamm equation solutions (2) that predict the sedimentation profiles for all species. For calculating Lamm equation solutions, Demeler and colleagues apply the ASTFEM algorithm that was recently introduced by Cao and Demeler (2005). In that work, the authors report two criteria for the performance of their ASTFEM algorithm in comparison with the reference (true) solution: (1) the overall rmsd (referred to by Cao and Demeler as "L^2 error," in a nonstandard definition), and (2) the maximum error in the evolution of concentration profiles.

That the rmsd is small (compared with the noise of data acquisition) is a necessary but not sufficient condition for the algorithm to be useful in modeling experimental data. In fact, the majority of points of the predicted concentration profiles typically fall into the plateau regions, which are trivial to predict (those in the solvent plateau are constant zero) but have limited or no information about the sedimentation process. These plateau points can keep the overall rmsd error of the solution below the statistical errors of the data acquisition, even though the maximum errors in the sedimentation boundaries may be much larger.

[3] Scaling this to the processor clock rate of a G1 smartphone, we expect this calculation should take less than 1 h, which would still compare well with the experimental time of several hours.

The accuracy of the description of the shape of the sedimentation boundary (rather than the plateaus) is critical for modeling the size-and-shape distributions. Therefore, a sufficient condition is that the maximum error is smaller than the noise of the data acquisition. For example, in order to model experimental data with signal-to-noise ratio of up to ∼1000:1, the maximum errors of the Lamm equation solutions at unit concentration should be less than 0.001.

For numerically solving the Lamm equation, an overriding question is the discretization of the radial coordinate. Solutions with fine radial mesh are generally more accurate but computationally more expensive, and conversely, coarsely discretized Lamm equation solutions are quicker to calculate. Even though it has not been explicitly mentioned in the SV literature until recently (Brown and Schuck 2008), it is easy to see that a fundamental limitation for any finite-element algorithm with linear elements is the obligate error that occurs when approximating a smooth, curved function with piecewise linear segments. This is illustrated in Fig. 2 for a system chosen by Cao and Demeler (2005) as a benchmark in the introduction of their ASTFEM algorithm. Figure 2 shows the deviations of the curved, accurate solution from a series of linear segments with a total of only 100 (red) or 200 (blue) radius values from meniscus to bottom.

For the determination of suitable radial mesh sizes for calculating the Lamm equation solution, Cao and Demeler applied the L^2 error criterion. This led to the recommendation of very coarse grids with ∼100 points, and indeed the main benefit of the ASTFEM algorithm perceived by Cao and Demeler (2005) is numerical stability even for such very coarse radial grids.

Unfortunately, large maximum errors in the approximation of the sedimentation boundaries are an unavoidable consequence of coarse radial discretization. In fact, the errors in the sedimentation boundaries shown in Fig. 2 are similar in magnitude to those of Figs. 8b and 9b in Cao and Demeler (2005). Remarkably, none of the examples provided by Cao and Demeler (2005) led to maximum errors below 0.001, and in most cases it was a factor of 10 or more above this mark. Such errors can be expected to significantly impact the result of the size-and-shape distribution analysis.

We have recently derived a new finite-element algorithm (Brown and Schuck 2008) based on the recognition that the approximation of the concentration profiles as linear segments does not only generate an obligate error (independent of the algorithm), but that this also represents the dominant source of error in the finite-element approach as described by Claverie et al. (1975). Accordingly, we generate a set of nonequidistant radial grid points with optimal spacing to achieve Lamm equation solutions with constant, predetermined accuracy (as measured by the maximum error for the radial data range to be analyzed).

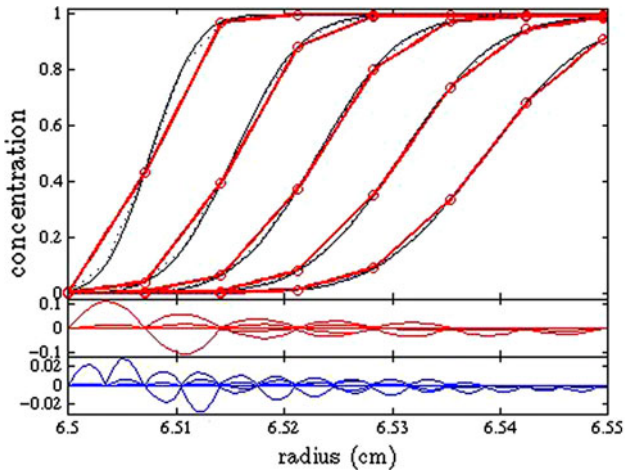


Fig. 2 Accuracy of the solution of the Lamm equation. Whenever using linear elements for the finite-element solution, an obligatory error is the approximation of the true boundary shape by piecewise linear segments. This is illustrated here for a system chosen as model system by Cao and Demeler (2005, compare Fig. 8b), with $s = 10\ S$ and $D = 2 \times 10^{-7}$ cm^2/s, for which very accurate Lamm equation solutions were calculated with a very fine discretization (*black thin line*). If the radial range from meniscus to bottom is divided evenly in a set of only 100 radial points and the boundary shape is approximated by piecewise linear segments (*red line*, residuals shown in enhanced scale in the *graph below*), very large deviations occur, even if at these points the correct Lamm equation solutions were calculated. For an even division with 200 radial points (*blue*) the obligatory errors are smaller but still approximately ten times the experimental noise. Grids with 100 radial points were proposed by Cao and Demeler (2005), leading for samples at unit concentration to maximum errors far exceeding the experimental noise. As shown by Brown and Schuck (2008), the minimum number of radial points that for this system allow for this obligate error to be <0.001 is ~300, based on an optimized nonequidistant spacing of radial points (using high density where boundaries are steep)

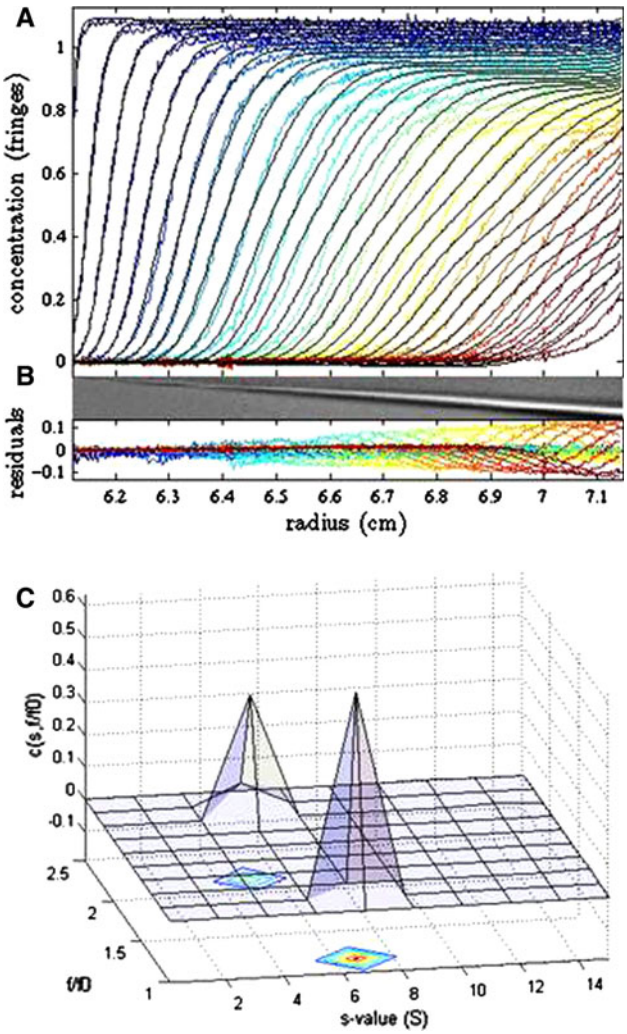


Fig. 3 Analysis of the same data shown in Fig. 1, using a coarse grid of only $K = 10 \times 10$ (f_r, s)-values as introduced by Brookes et al. **a** Experimental data, with the color temperature blue to red indicating the temporal evolution of the sedimentation, as in Fig. 1. Shown as *black* lines are the best-fit distributions with the 10×10 grid distribution model, ranging from f_r-values of 1.0 to 2.5 in ten equidistant steps, and from s-values of 0.1 to 15.0 S in ten equidistant steps. **b** Residuals bitmap and overlay. The rmsd of the fit is 0.03088 OD. **c** Best-fit size-and-shape distribution with the 10×10 model, in the same presentation as the 10×100 model in Fig. 1c

To optimize the efficiency, all points in the solvent and solution plateaus are calculated with the trivial analytical expressions (Brown and Schuck 2008). We note that, for the 10×100 grid shown in Fig. 1, the calculation of the Lamm equation solutions for all 1,000 species with an accuracy of better than 0.001 (maximum error) requires a total of less than 2 s on our PC. Thus, computational expense for achieving high-accuracy Lamm equations should not be limiting the size-and-shape distribution analysis of SV data.

The 2DSA "divide-and-conquer" algorithm by Brookes et al.

The 2DSA algorithm applied by the Demeler laboratory consists of a large number of repeated applications of Eq. (4) with K ≈ 10×10 and similarly low-resolution grids. Figure 3 shows the results of fitting the same data as shown in Fig. 1 with a 10×10 grid under otherwise identical conditions. The deviations are ±10% of the maximum signal, and clearly this model does not even qualitatively describe the data well. As a consequence, we cannot assume that the distribution obtained from this model reflects in any way the species present in the experiment. (It is grossly different, for example, from the distribution shown in Fig. 1c, d.)

Brookes et al. (2009) recognize that such a fit is insufficient and consistently attribute the idea of using 10×10 grids to Brown and Schuck. For example, the authors state "…a 10×10 grid as proposed by Brown and Schuck

(2006) is insufficient to reliably describe the experimental parameter space. If the actual solute is not aligned with a grid point, false positives are produced," and even declare the second major point in their results as "A 10 × 10 grid suggested by Brown and Schuck (2006) is clearly insufficient…," and state in the summary "We have shown that low resolution grids as proposed by Brown and Schuck (2006) are insufficient to obtain reliable information." This attribution is not based on reality. Unmistakably, we have used in the referenced work (Brown and Schuck 2006) exclusively high-resolution grids (11 × 100 in Fig. 1 and 2, 12 × 60 in Fig. 3, 15 × 50 in Fig. 4, 13 × 100 in Fig. 5, 11 × 100 in Fig. 6, and finally 13 × 50 in Fig. 7), all of which are shown to describe the data well to within the noise of data acquisition (Brown and Schuck 2006), and similar is true for other published applications of the method by other laboratories and by us. Thus, the idea of using 10 × 10 grids is entirely a product of the Demeler laboratory, and, to our knowledge, first described in the Brookes et al. (2009) paper.

Despite the failure of overly coarse grids, remarkably, Demeler's 2DSA algorithm consists exclusively of repeat applications of such coarse grids: They are considered "subgrids" of a hypothetical grid with much higher resolution, which is never actually completely fitted to the data, but nevertheless suggested to reflect the final resolution of the distribution. The details are not entirely clear, but there are two key ideas: (I) The coarse grids are translated relative to each other multiple times by increments $\Delta_2 s$ and $\Delta_2 f_r$, and their results are joined. (II) The joined set of grid points with inactive nonnegativity constraints from (I) is used to form a new, second-stage irregular grid of similarly low number of grid points as the initial grid.[4] The Demeler scheme of repeat application of different coarse subgrids, storage, and combination of their results, is termed a "divide-and-conquer" strategy. Divide-and-conquer algorithms are well-known tools in numerical mathematics that facilitate the use of parallel computation to solve problems, such as singular value decomposition (Arbenz and Golub 1988; Gu and Eisenstat 1995; Xu and Qiao 2008). Generally, such algorithms are established by proof of their correctness. This criterion has not been attempted for the 2DSA algorithm. Concerns arise from the following arguments:

[4] As described, for example, by Demeler et al. (2009): "Typically, we apply 100–300 grid movings of a 10 × 10 grid to obtain a resolution that is commensurate with the resolution of the analytical ultracentrifuge." and "Solutes with positive amplitudes from different grids are then unioned with each other to form new grids with a maximum number of solutes equivalent to that of a single initial grid (generally less than 100 solutes)."

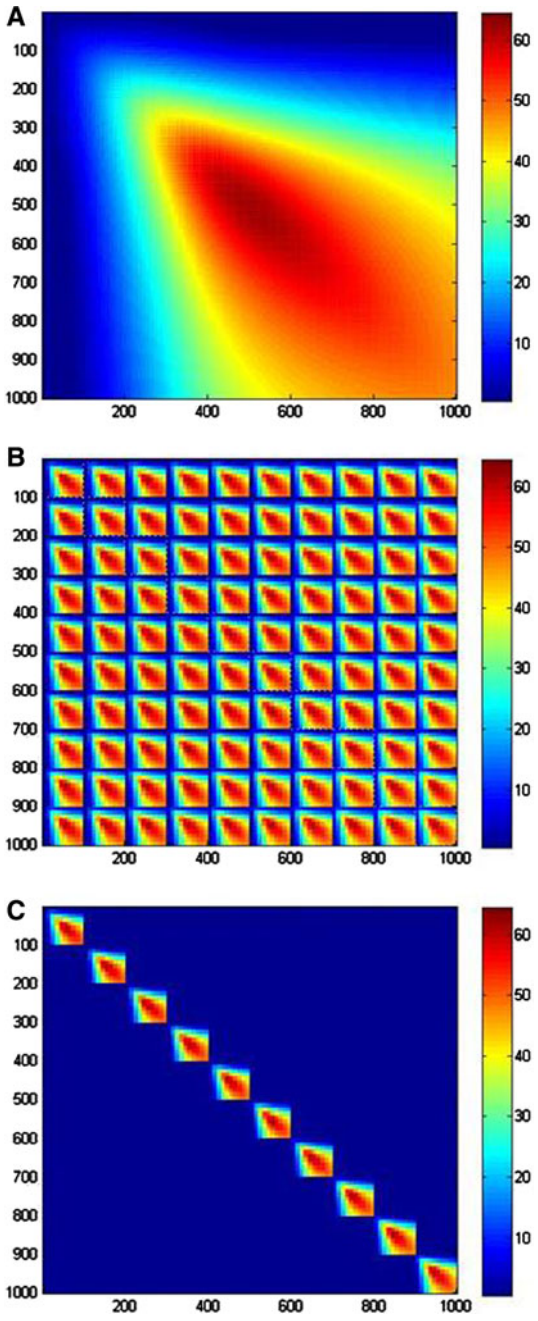
(I) Combination of subgrids

The premise underlying (I) is that the results from independent application of different grids can be meaningfully combined. Following the idea of the Demeler laboratory that low-resolution subgrids can be "refined into a grid of any desired resolution" through their combination scheme, let us consider that putative final regular high-resolution grid, which would have mesh size $\Delta s = \Delta_2 s$ and $\Delta f_r = \Delta_2 f_r$. As shown above, one can actually solve the size-and-shape distribution problem directly using the standard algorithm with this full-sized high-resolution grid with even mesh size, via the normal equation (5) with the $K \times K$ matrix $\mathbf{P}$ and $K \times 1$ vector $\vec{\mathbf{d}}$, where K is the total number of species of the two-dimensional grid. In our example of Fig. 1, $K = 1{,}000$ for the 10 × 100 grid that is of sufficient resolution to describe all aspects of the experiment. Now going backwards, one may consider our high-resolution grid to be represented by a total of Γ different equal-sized subgrids, each referenced with index γ (e.g., ten grids of 10 × 10 resolution), such that merging all grid points of the subgrids produces the high-resolution grid. For each subgrid, one can solve the distribution with the normal matrix method, and it is easy to show that the relevant matrix equations are $\mathbf{P}^{(\gamma)}\vec{\mathbf{c}}^{(\gamma)} = \vec{\mathbf{d}}^{(\gamma)}$, where $\mathbf{P}^{(\gamma)}$ are square submatrices of $\mathbf{P}$ of size $(K/\Gamma) \times (K/\Gamma)$ and $\vec{\mathbf{d}}^{(\gamma)}$ are subvectors of $\vec{\mathbf{d}}$ of size $1 \times (K/\Gamma)$. One can use a nomenclature for the elements of the high-resolution grid such that the points are ordered in sequence of the low-resolution subgrids.

In general, it is not true that the individual results $\vec{\mathbf{c}}^{(\gamma)}$ from the individual problems $\mathbf{P}^{(\gamma)}\vec{\mathbf{c}}^{(\gamma)} = \vec{\mathbf{d}}^{(\gamma)}$ can be combined to a concatenated vector $\left(\vec{\mathbf{c}}^{(1)}, \ldots, \vec{\mathbf{c}}^{(\Gamma)}\right)$ that would represent the result $\vec{\mathbf{c}}$ of the full solution (with or without nonnegativity). This would require the cross-correlation between points from the different grids to vanish, and the high-resolution $K \times K$ matrix $\mathbf{P}$ to have a structure

$$\mathbf{P} = \begin{pmatrix} \mathbf{P}^{(1)} & & & 0 \\ & \mathbf{P}^{(2)} & & \\ & & \ddots & \\ 0 & & & \mathbf{P}^{(\Gamma)} \end{pmatrix}. \tag{8}$$

This is not the case, as illustrated in Fig. 4 for our example data. As can be discerned clearly, the structure of $\mathbf{P}$ when sorted along subgrids (Fig. 4b) is different from merging the submatrices $\mathbf{P}^{(\gamma)}$ (Fig. 4c), which neglects very substantial features of the model. If we ignore this problem and calculate the distribution with the matrix of Fig. 4c (or, equivalently, if we simply merge all solutions from consecutively fitting the distribution data with all ten 10 × 10 grids and plot them at their appropriate points in the high-



◀ **Fig. 4** Magnitude of the elements of the normal matrix **P** calculated for the 10 × 100 model shown in Fig. 1. **P** is symmetrical and has 1,000 × 1,000 values, plotted here by row and column numbers as indicated in the abscissa and ordinate of the picture, and the values $|\mathbf{P}_{kl}|$ are plotted using the color scale. In principle, the nomenclature indexing the 10 × 100 grid points for the $f_r \times s$ grid to form the vector of 1,000 parameters is arbitrary. **a** Here, all grid points are sorted by increasing s-value, i.e., $(s_1, f_{r,1})$, $(s_1, f_{r,2})$,…$(s_1, f_{r,10})$, $(s_2, f_{r,1})$,…$(s_2, f_{r,10})$,…$(s_{100}, f_{r,10})$. As can be discerned from the smooth appearance, the matrix elements are not strongly dependent on the f_r-value. **b** The same matrix can be reordered to reflect subdivision along ten regular subgrids γ, each of the form $(s_{10(\gamma-1)+1}, f_{r,1})$, $(s_{10(\gamma-1)+1}, f_{r,2})$,…$(s_{10(\gamma-1)+1}, f_{r,10})$, $(s_{10(\gamma-1)+2}, f_{r,1})$, $(s_{10(\gamma-1)+2}, f_{r,2})$,…$(s_{10(\gamma-1)+2}, f_{r,10})$,…$(s_{10(\gamma-1)+10}, f_{r,1})$, $(s_{10(\gamma-1)+10}, f_{r,2})$,… $(s_{10(\gamma-1)+10}, f_{r,10})$ with $\gamma = 1\ldots10$. Each of the subgrids represents an evenly spaced 10 × 10 grid with origin shifted by $\Delta_2 s = 0.1505\ S$. **c** The idea that one could determine a high-resolution size-and-shape distribution from merging the results obtained separately in fits with subgrids corresponds to the assumption that there be no correlation between points from the different grids, i.e., that **P** can be subdivided into the ten submatrices $\mathbf{P}^{(\gamma)}$. For the present data, this corresponds to the solution of the problem with a normal matrix as shown in **c**. Clearly, this is very different from the true matrix shown in **b**

resolution grid), we arrive at a size-and-shape distribution as shown in Fig. 5b. This is very different from the known exact solution shown in Fig. 1, which is reproduced for convenience in Fig. 5a.

(II) Formation of new, irregular coarse subgrids

Apparently to address this problem, the 2DSA method takes from the concatenated solution of the subgrids only the pattern of active/inactive nonnegativity constraints. Demeler and colleagues construct from the points with nonzero concentration values in the concatenated partial solutions (or a subset thereof) a new grid, conceived to be equal in size to the original low-resolution grids, but now with uneven spacing of the grid points.

Again, we can analyze this approach best by comparison with the full, high-resolution grid with the full matrix **P**, where the unambiguous best-fit nonnegative solution is found exactly with the proven NNLS algorithm (Lawson and Hanson 1974). The ad hoc exclusion of grid points that did not produce positive concentration values in any of the subgrids is in direct conflict with NNLS. Nothing guarantees that the (s, f_r)-values populated in the exact solution will be correctly recognized as populated species (be assigned nonzero values) in the fit with the low-resolution grid of which they are a part. The points populated in the exact solution may therefore simply not be part of the second-stage grid.

Illustrating this problem, the crosses in Fig. 5d represent all the grid points that made positive contributions in any of the preliminary sequence of low-resolution fits (which covers all grid points of the high-resolution grid, as described above). All the dots (red and blue) are the positive solution components of the exact high-resolution solution. They are colored blue if they coincide with a cross, i.e., have been correctly identified in the first stage as being part of the solution, and they are colored red if they were never part of any low-resolution fit and were therefore excluded from the second-stage grid. If the analysis proceeds with the second-stage grid (i.e., preconstraining the analysis to the values indicated by crosses in Fig. 5d), we arrive at the solution shown in Fig. 5c. This is very different from the true high-resolution solution shown in Fig. 5a. Thus, the second stage cannot correct for the errors that occur from a naïve subdivision of grids in (I).

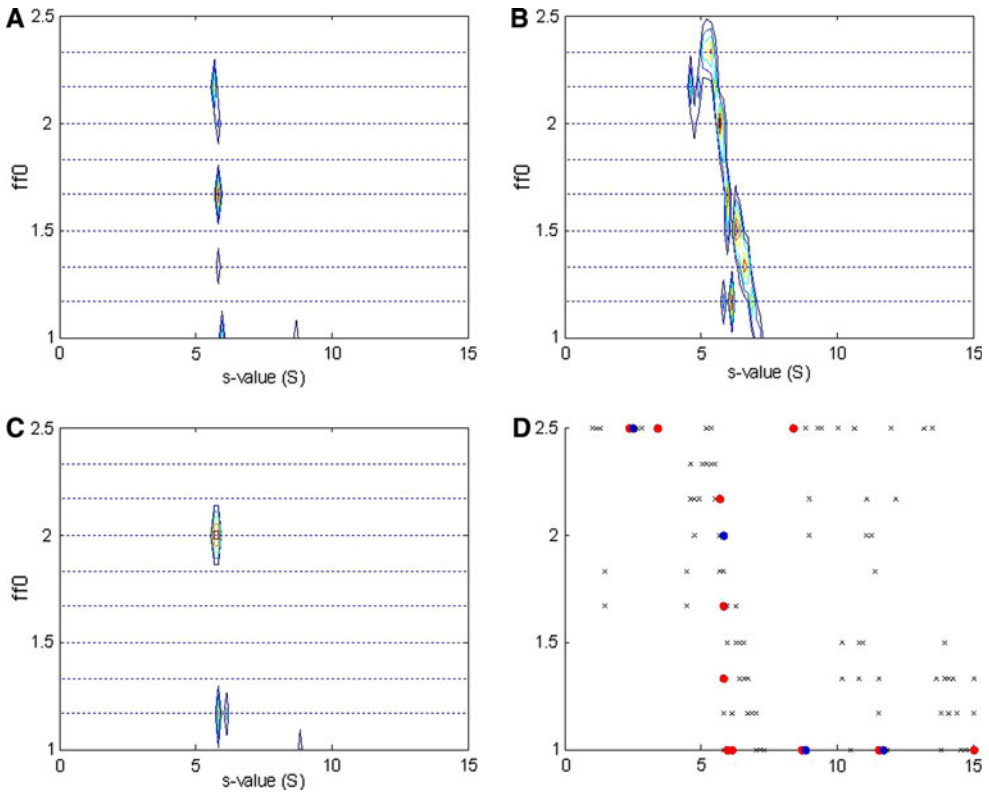


Fig. 5 Contour plots of the size-and-shape distributions calculated with different models for the IgG data shown in Fig. 1. The color temperature (*blue* to *red*) indicates increasing height of the peaks. (**a**) For comparison, this panel shows the same distribution as shown in Fig. 1c, calculated with the high-resolution grid of 10 × 100 (f_r, s)-values. **b** Distribution obtained in stage (I) by merging the distributions calculated sequentially and independently with different low-resolution grid of 10 × 10 (f_r, s)-values, each translated by $\Delta_2 s = 0.1505\ S$. One example for the low-resolution grid analysis is shown in Fig. 3. All low-resolution grids are chosen such that they are evenly spaced subgrids of the high-resolution grid, and such that, by joining the grid points of (f_r, s)-values of all the low-resolution grids, the high-resolution grid of **a** is obtained. **c** A new grid is defined in stage (II) by joining all grid points from the entire sequence of low-resolution grids that yielded positive contributions to the fit. This is the set of grid points for which b displays nonzero populations of the distribution. In a secondary analysis, a fit to this irregular grid is performed, and the results are shown as a contour plot. Although the smallest differences Δs and Δf_r in this secondary grid are the same as those of the high-resolution grid, it considers only a small subset of the points from the high-resolution grid. This causes the deviations from the exact results in **a** and those in **c**. **d** Illustration of the grid points used in **c**, showing as *black crosses* all points that yielded positive contributions in any of the first-stage low-resolution fits. For comparison, *solid circles* are the grid points that make positive contributions in the exact direct high-resolution analysis of **a**. *Blue circles* indicate those that coincide with grid points in the Demeler scheme, and *red circles* indicate those that are populated in the exact solution but not found in the grid of the Demeler scheme

Brookes mentions multiple stages of the sequence (I) and (II) with different mesh sizes $\Delta_2 s$ and $\Delta_2 f_r$, and an "iterative refinement" of the procedure that utilizes in stage (I) the coarse starting grids that have been extended with populated points from the results of stage (II) of the previous iteration (Brookes et al. 2009). The same fundamental concerns apply to this iteration. To the extent that the results from (II) may not contain the grid points of the exact solution, it is unclear how the inclusion of these additional grid points would aid in the recognition of the correct solution. Even if the added grid points in (I) do represent part of the correct solution, it is not certain that they would be correctly maintained as part of the solution after (II). Empirically, the Demeler laboratory reports convergence of this iteration series in the absence, but not in the presence, of systematic noise corrections to the data (Brookes et al. 2006). Even if the iteration does converge, it is unclear whether it is convergent to the correct solution.

Parsimony: suppressing artificial detail

Since the 2DSA algorithm never actually applies a model with a full regularly spaced high-resolution grid, the traditional regularization methods, such as Tikhonov or maximum-entropy regularization described above, do not seem to be easily applicable. In fact, Brookes et al. (2009) express the view that the fit of $c(s, f_r)$ with a high-resolution grid in conjunction with regularization suffers from "lack of resolution," and "produces unnecessarily broad molecular weight distributions." We believe that, if prior

knowledge about the sharpness of the expected $c(s, f_r)$ peaks is available, this can be inserted with a Bayesian refinement of the Tikhonov regularization as we have reported for SV analysis (Brown et al. 2007) and implemented in SEDFIT.

In the absence of such prior knowledge, however, the resolution of the regularized solution is limited not by the analysis (assuming reasonable discretization), but rather by the information content of the experiment. It is important to recognize the nature of this limit, in order not to overinterpret the data. Of course, it also would be trivial, although usually misguided, to perform a distribution analysis simply not applying this regularization step at all, and to rely on the exact solution of the fit with the high-resolution model, which usually produces artifactual detail that is the result of noise amplification due to the ill-conditioned nature of the basic Eq. (3).

In our example, these aspects can be discerned when comparing the most parsimonious solution in Fig. 1d from Tikhonov regularization with the spiky exact solution in Fig. 1c, or with the incorrect solution in Fig. 5c from one iteration adapted from the Demeler scheme. Even though the spiky solutions suggest very few and discrete species to be in solution, the smooth Tikhonov solution fits the data indistinguishably well from the exact best-fit solution. Its nearly featureless appearance in the f_r-dimension highlights simply the lack of sufficient information in the raw data in order to determine the f_r-values well.

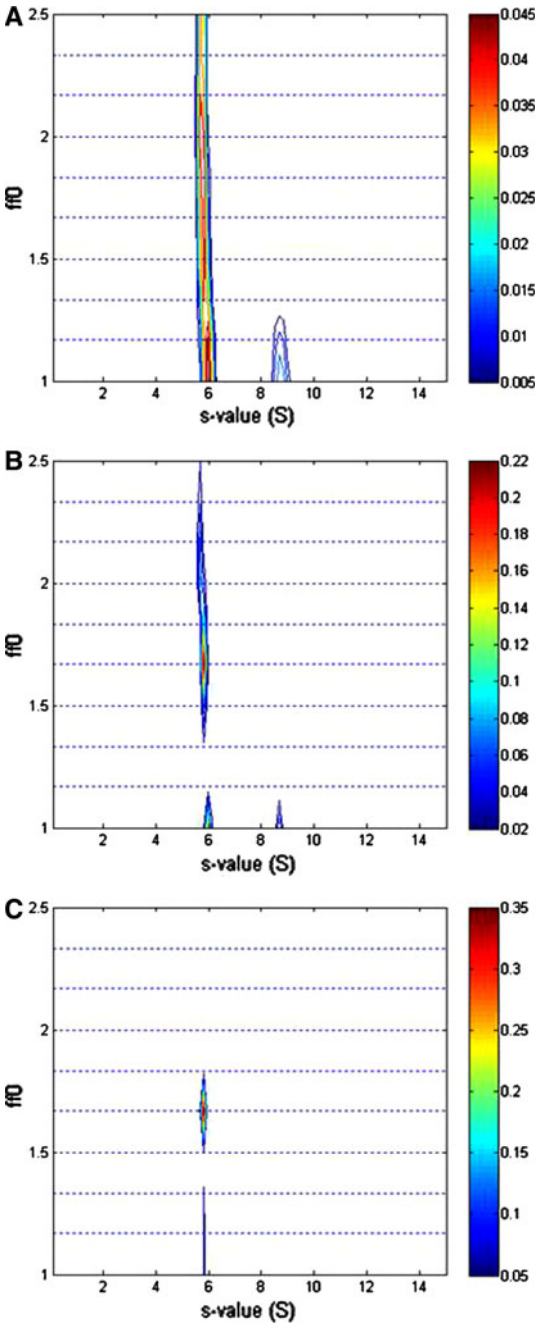
In order to address the impact of noise and error amplification on the results of the 2DSA algorithm, it was combined by Brookes et al. (2009) with a Monte Carlo analysis. Fifty iterations were performed by the Demeler laboratory in order to determine 95% confidence intervals. This seems to be an unusually low number of iterations, in particular since the high confidence limits require estimating the quantiles of rare events, in this case the 2.5 and 97.5 percentiles. With 50 iterations, they are determined by the extreme 1.25 occurrences of parameter values, which makes these estimates of the confidence intervals quite variable statistical quantities themselves. As is well known, usually the number of Monte Carlo iterations required to produce meaningful results is typically on the order of 1,000–10,000. However, it seems this would lead to excessive computational effort, several orders of magnitude more costly than the Tikhonov regularization in the standard approach with the full high-resolution grid, which requires for our standard example only a few seconds on our PC.

The authors report confidence intervals for molecular parameters of the identified solutes, but it is not clear whether these were determined (1) by statistical analysis of the results obtained in each Monte Carlo iteration after the integration of putative solute peaks, or (2) if these confidence intervals reflect the uncertainties of the putative solute peaks in a distribution that, as a whole, has gained error bars at each grid point from the statistics of the Monte Carlo iterations. For method (1), the problem arises of how to identify the group peaks representing a putative solute species. For method (2), the question arises of whether the Monte Carlo approach is effective in providing parsimonious "average" distributions.

Generally, Monte Carlo simulations are not part of the diverse set of regularization methods explored in the standard literature (Louis 1989; Hansen 1992, 1998; Kress 1999; Engl et al. 2000), although Monte Carlo methods have been used for estimating the regularization parameters of standard regularization functionals (Ramani et al. 2008). The concept of a statistical distribution of parameter values should be confused neither with the real population distribution of coexisting species in the sample mixture nor the estimate of the latter in the form of a calculated size-and-shape distribution. Nevertheless, one could ask to what extent one can rely on the statistical nature of the noise in the data, in combination with noise amplification, to produce a parsimonious two-dimensional histogram of (s, f_r)-species populations. As an example, we compare in Fig. 6 the standard analysis of our model data with the 10 × 100 grid and Tikhonov regularization (6a) with the histogram of all distributions from 50 Monte Carlo iterations (each based on an exact standard fit with the 10 × 100 grid; 6b). After 50 iterations the histogram clearly shows multimodal and spiky behavior suggesting the presence of multiple species, in contrast to the single broad peak representing the smoothest solution of all that fit the originally measured data. Thus, the 50 Monte Carlo iterations do not provide an effective means to correctly identify the information content of the data. If, on the other hand, we are independently knowledgeable about the monodisperse nature of the sample, we can use the Bayesian approach (Brown et al. 2007) to calculate the size-and-shape distribution that is closest to a single peak, and these results are shown, for comparison, in Fig. 6c.

Summary and conclusions

In the present letter, we have examined the different algorithmic elements that were conceived and applied in the recently suggested "2DSA" size-and-shape distribution by Brookes et al. (2009). We have compared this with the standard approach that is well established for solving ill-posed integral equations problems in many fields, which rests on well-established linear algebra and related numerical tools of linear least-squares analysis. Contrary to the assertion of Brookes, Cao, and Demeler, the application of the standard approach to the size-and-shape distribution



◀ **Fig. 6** Comparison of strategies to compute parsimonious distributions that display the information content of the IgG data shown in Fig. 1. **a** Contour plot of the size-and-shape distribution obtained with the high-resolution grid of 10 × 100 (f_r, s)-values, after application of Tikhonov regularization, as shown in Fig. 1d. **b** The sum of 50 size-and-shape distributions calculated with the exact standard method using the same high-resolution grid, but each based on synthetic data sets generated from the best-fit distribution of Fig. 1 with added normally distributed noise at the same magnitude as exhibited by the experimental data. **c** Integration of the main 6 S peak of the size-and-shape distribution as calculated in Fig. 1 allows to determine the weighted-average s-value and f_r-value, which can be used in the Bayesian framework to calculate the size-and-shape distribution $c^{\delta}(s, f_r)$ (Brown et al. 2007) that is closest to that of a single species, within the limits imposed by the requirement to produce a fit of statistically indistinguishable quality to that shown in **a**. As can be discerned from the secondary peak at ∼6 S with low frictional ratio, a strictly monodisperse interpretation of the main peak is contradicted by the experimental data. (Note the different scales on the color bar)

problem is quite feasible on ordinary laboratory computers within only a few minutes of computation time, even when using high-resolution grids suitable to fully extract the experimental information content. As implemented in SEDFIT, this approach is being applied in many laboratories (Markossian et al. 2006; Chang et al. 2007; Deng et al. 2007; Race et al. 2007; Broomell et al. 2008; Brown et al. 2008; Chebotareva et al. 2008; Iseli et al. 2008; Moncrieffe et al. 2008; Paz et al. 2008; Sivakolundu et al. 2008; Wang et al. 2008; Eronina et al. 2009; Mortuza et al. 2009). Since this can supply exact (up to numerical precision) best-fit solutions, we have applied it to a data analysis example to serve as a reference solution in a study of the performance of different computational strategies on which 2DSA relies. This illustrates the consequences of the deviations from the established mathematical reference frame that should be expected to arise in Demeler's 2DSA approach.

First, there are concerns about the accuracy of the evaluated Lamm equation solutions serving as kernel to the size-and-shape distribution integral. This could likely be addressed by deviating from the discretization parameters advocated by Cao and Demeler (2005).

Second, a more fundamental problem is the use of grids with extremely small number of points, far below the resolution required to describe the data. If, as illustrated in Fig. 3, the predicted concentration profiles from these coarse models do not even qualitatively follow the experimental data, we question whether there are any meaningful conclusions that can be drawn from these results. Brookes et al. (2009) distract from this problem by incorrectly stating that such grids were the basis of the implementation of $c(s, f_r)$ models in SEDFIT, which is well described in the literature to achieve excellent fits of the data to within their statistical noise. To the best of our knowledge the attempt to utilize coarse grids is without precedent prior to the Brookes et al. paper.

Despite the inability of these grids to describe the data, Demeler and colleagues suggest that the combination of results from the application of a large number of different, but similarly coarse, grids (all with 10 × 10 or lower resolution; Demeler et al. 2009) can be used in some way to achieve an analysis equivalent to that of a high-resolution grid. In the simplest form, this argument would be incompatible with basic matrix algebra, because it neglects cross-correlation between points from different grids. Discarding the magnitude of species' populations in this concatenated distribution, and using only the pattern of

nonnegativity constraints from such an analysis, is in conflict with the established Lawson and Hanson algorithm NNLS. The effect of the empirical extension to multiple stages is uncertain, and may not converge. Although one could construe cases where it will certainly work (such as distributions consisting of a single species), the Demeler scheme for generating nonequidistant small grids in multiple stages appears fundamentally flawed for the general case.

The strategy of sequentially applying different, equally coarse, grids is in contrast to established multigrid methods for integral equations, which provide successfully finer parameter discretization (Kress 1999). The division of the full problem into separate subproblems to be solved in parallel, followed by merging their partial solutions, has been used successfully in some image restoration problems (Bevilacqua and Piccolomini 2000) where the image regions are known to be uncorrelated with each other due to a localized point-spread function. However, this condition is not fulfilled in the present case. In SV analysis, the cross-correlation of signals from different species can be very large. This is reflected by the fact that (1) is ill posed, and illustrated by the fact that the matrices in Fig. 4b and c are different. For a correct solution of the SV problem, the regular high-resolution grid should be considered fully and unbiased by any scheme of preselection of excluded parameter regions. The latter is quite feasible with standard algorithms and commonly available computational resources, and we note that the problem is fairly small compared with many image analysis problems of similar structure.

Finally, the application of the Monte Carlo approach to achieve greater parsimony of the results (i.e., simplicity of the distribution in the sense of suppressing artificial detail) is equally novel, but not very successful when we applied this idea to our example data analysis. An example of the lack of regularization in the 2DSA method resulting in artificial detail can be found in the data shown by Planken et al. (2008), where a standard $c(s)$ analysis of SV data with maximum-entropy regularization exhibits only a single broad skewed distribution [Fig. 3c in Planken et al. (2008)], consistent with the expected continuous size distribution of the material studied, yet the 2DSA analysis of the same data suggests the presence of more than 14 discrete peaks (at different s-values and all at similar frictional ratio) [Fig. 4 in Planken et al. (2008)]. The Monte Carlo approach is certainly an extremely computationally costly step, in particular if one would carry it out with statistically meaningful iteration numbers. In contrast, application of the standard Tikhonov regularization to the full high-resolution problem, with or without Bayesian modulation, takes a small fraction of the computational effort of the original problem, i.e., on the order of seconds on a PC.

In conclusion, we would regard the computational effort to be a secondary problem, and the choice of computational platform rather inconsequential, relative to the main concern arising from simple mathematical arguments that Demeler's algorithm may not give correct results. The authors do qualify their algorithm to be empirical, and that "the results are not generally in exact correspondence with the original problem" (Brookes et al. 2006). They argue that "[the results] can be made sufficiently close through careful use of the given heuristics" (Brookes et al. 2006). We are uncertain of the process referred to here, and how closeness to the exact solution would possibly be assessed without explicitly calculating the exact best-fit solution. So far Demeler and colleagues have not brought forward any proof that the distributions returned by the 2DSA method are at least close in the major attributes to the correct solution. We believe that the question of correctness of the algorithm is critical, especially since the authors invite the general application of this method, as implemented in the ULTRASCAN software, to address data analysis problems in novel biophysical and biochemical studies, rather than simple model problems with known solutions.

Acknowledgments This work was supported by the Intramural Research Program of the National Institute of Bioimaging and Bioengineering, NIH.

References

Arbenz P, Golub GH (1988) On the spectral decomposition of Hermitian matrices modified by low rank pertubationswith aplications. SIAM J Matrix Anal Appl 9:40–58

Balbo A, Minor KH et al (2005) Studying multi-protein complexes by multi-signal sedimentation velocity analytical ultracentrifugation. Proc Natl Acad Sci USA 102:81–86

Behlke J, Ristau O (2009) Enhanced resolution of sedimetnation coefficient distribution profiles by extrapolation to infinite time. Eur Biophys J. doi 10.1007/s00249-009-0425-1

Berkowitz SA (2006) Role of analytical ultracentrifugation in assessing the aggregation of protein biopharmaceuticals. AAPS J 8(3):E590–E605

Bevilacqua A, Piccolomini EL (2000) Parallel image restoration on parallel and distributed computers. Parallel Comput 26:495–506

Bevington PR, Robinson DK (1992) Data reduction and error analysis for the physical sciences. New York, Mc-Graw-Hill

Brookes E, Boppana RV et al (2006) Computing large sparse multivariate optimization problems with an application in biophysics. In: Proceedings of the 206 ACM/IEEE conference on supercomputing, Tampa, Florida

Brookes E, Cao W et al (2009) A two-dimensional spectrum analysis for sedimentation velocity experiments of mixtures with heterogeneity in molecular weight and shape. Eur Biophys J. doi: 10.1007/s00249-009-0413-5

Broomell CC, Chase SF et al (2008) Cutting edge structural protein from the jaws of Nereis virens. Biomacromolecules 9(6):1669–1677

Brown PH, Balbo A et al (2008b) Characterizing protein–protein interactions by sedimentation velocity analytical ultracentrifugation. Curr Protoc Immunol, Chap 18: Unit 18 15

Brown PH, Schuck P (2006) Macromolecular Size-And-Shape Distributions by Sedimentation Velocity Analytical Ultracentrifugation. Biophys J 90:4651–4661

Brown PH, Schuck P (2008) A new adaptive grid-size algorithm for the simulation of sedimentation velocity profiles in analytical ultracentrifugation. Comput Phys Commun 178(2):105–120

Brown P, Balbo A et al (2007) Using prior knowledge in the determination of macromolecular size-distributions by analytical ultracentrifugation. Biomacromolecules 8:2011–2024

Brown J, Delaine C et al (2008a) Structure and functional analysis of the IGF-II/IGF2R interaction. EMBO J 27(1):265–276

Brown PH, Balbo A et al (2008b) A bayesian approach for quantifying trace amounts of antibody aggregates by sedimentation velocity analytical ultracentrifugation. Aaps J 10(3):481–493

Brown PH, Balbo A et al (2009) On the analysis of sedimentation velocity in the study of protein complexes. Eur Biophys J 38:1079–1099

Cao W, Demeler B (2005) Modeling analytical ultracentrifugation experiments with an adaptive space–time finite element solution of the lamm equation. Biophys J 89(3):1589–1602

Chang HP, Chou CY et al (2007) Reversible unfolding of the severe acute respiratory syndrome coronavirus main protease in guanidinium chloride. Biophys J 92(4):1374–1383

Chebotareva NA, Meremyanin AV et al (2008) Cooperative self-association of phosphorylase kinase from rabbit skeletal muscle. Biophys Chem 133(1–3):45–53

Claverie J-M, Dreux H et al (1975) Sedimentation of generalized systems of interacting particles. I. Solution of systems of complete Lamm equations. Biopolymers 14:1685–1700

Cole JL, Lary JW et al (2008) Analytical ultracentrifugation: sedimentation velocity and sedimentation equilibrium. Methods Cell Biol 84:143–179

Correia JJ, Stafford WF (2009) Extracting equilibrium constants from kinetically limited reacting systems. Methods Enzymol 455:419–446

Dam J, Schuck P (2004) Calculating sedimentation coefficient distributions by direct modeling of sedimentation velocity profiles. Methods Enzymol 384:185–212

Dam J, Schuck P (2005) Sedimentation velocity analysis of protein–protein interactions: Sedimentation coefficient distributions c(s) and asymptotic boundary profiles from Gilbert-Jenkins theory. Biophys J 89:651–666

Dam J, Velikovsky CA et al (2005) Sedimentation velocity analysis of protein–protein interactions: Lamm equation modeling and sedimentation coefficient distributions c(s). Biophys J 89:619–634

Demeler B, Brookes E et al (2009) Analysis of heterogeneity in molecular weight and shape by analytical ultracentrifugation using parallel distributed computing. Methods Enzymol 454:87–113

Deng L, Langley RJ et al (2007) Structural basis for the recognition of mutant self by a tumor-specific, MHC class II-restricted T cell receptor. Nat Immunol 8(4):398–408

Engl HW, Hanke M et al (2000) Regularization of inverse problems. Kluwer, Dordrecht

Eronina TB, Chebotareva NA et al (2009) Effect of proline on thermal inactivation, denaturation and aggregation of glycogen phosphorylase b from rabbit skeletal muscle. Biophys Chem 141(1):66–74

Gabrielson JP, Arthur KK et al (2009) Common excipients impair detection of protein aggregates during sedimentation velocity analytical ultracentrifugation. J Pharm Sci 98(1):50–62

Gill PE, Murray W et al (1986) Practical optimization. Academic Press, Amsterdam

Golub GH, VanLoan CF (1989) Matrix computations. The Johns Hopkins University Press, Baltimore, MD

Gu M, Eisenstat SC (1995) A divide-and-conquer algorithm for the bidiagonal SVD. SIAM J Matrix Anal Appl 16(1):79–92

Hansen PC (1992) Numerical tools for analysis and solution of Fredholm integral equations of the first kind. Inverse Probl 8:849–872

Hansen PC (1998) Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion. SIAM, Philadelphia

Howlett GJ, Minton AP et al (2006) Analytical ultracentrifugation for the study of protein association and assembly. Curr Opin Chem Biol 10(5):430–436

Iseli TJ, Oakhill JS et al (2008) AMP-activated protein kinase subunit interactions: beta1:gamma1 association requires beta1 Thr-263 and Tyr-267. J Biol Chem 283(8):4799–4807

Kress R (1999) Linear integral equations. Springer, New York

Lamm O (1929) Die Differentialgleichung der Ultrazentrifugierung. Ark Mat Astr Fys 21B(2):1–4

Lawson CL, Hanson RJ (1974) Solving least squares problems. Prentice-Hall, Englewood Cliffs

Liu J, Andya JD et al (2006) A critical review of analytical ultracentrifugation and field flow fractionation methods for measuring protein aggregation. Aaps J 8(3):E580–E589

Louis AK (1989) Inverse und schlecht gestellte Probleme. Teubner, Stuttgart

Markossian KA, Khanova HA et al (2006) Mechanism of thermal aggregation of rabbit muscle glyceraldehyde-3-phosphate dehydrogenase. Biochemistry 45(44):13375–13384

Moncrieffe MC, Grossmann JG et al (2008) Assembly of oligomeric death domain complexes during Toll receptor signaling. J Biol Chem 283(48):33447–33454

Mortuza GB, Goldstone DC et al (2009) Structure of the capsid amino-terminal domain from the betaretrovirus, Jaagsiekte sheep retrovirus. J Mol Biol 386(4):1179–1192

Narayan R, Nityananda R (1986) Maximum entropy image restoration in astronomy. Ann Rev Astron Astrophys 24:127–170

Patel A, Vought VE et al (2008) A conserved arginine-containing motif crucial for the assembly and enzymatic activity of the mixed lineage leukemia protein-1 core complex. J Biol Chem 283(47):32162–32175

Pavlov GM, Korneeva EV et al (2009) Hydrodynamic properties of cyclodextrin molecules in dilute solutions. Eur Biophys J. doi: 10.1007/s00249-008-0394-9

Paz A, Zeev-Ben-Mordehai T et al (2008) Biophysical characterization of the unstructured cytoplasmic domain of the human neuronal adhesion protein neuroligin 3. Biophys J 95(4):1928–1944

Phillips DL (1962) A technique for the numerical solution of certain integral equations of the first kind. Assoc Comput Mach 9:84–97

Philo JS (2006) Improved methods for fitting sedimentation coefficient distributions derived by time-derivative techniques. Anal Biochem 354(2):238–246

Planken KL, Kuipers BW et al (2008) Model independent determination of colloidal silica size distributions via analytical ultracentrifugation. Anal Chem 80(23):8871–8879

Press WH, Teukolsky SA et al (1992) Numerical recipes in C. University Press, Cambridge

Provencher SW (1982) A constrained regularization method for inverting data represented by linear algebraic or integral equations. Comp Phys Comm 27:213–227

Race PR, Solovyova AS et al (2007) Conformation of the EPEC Tir protein in solution: investigating the impact of serine phosphorylation at positions 434/463. Biophys J 93(2):586–596

Ramani S, Blu T et al (2008) Monte-Carlo Sure: A black-box optimization of regularization parameters for general denoising algorithms. IEEE Trans Image Process 17:1540–1554

Ruhe A, Wedin PÅ (1980) Algorithms for separable nonlinear least squares problems. SIAM Review 22:318–337

Schuck P (2000) Size distribution analysis of macromolecules by sedimentation velocity ultracentrifugation and Lamm equation modeling. Biophys J 78:1606–1619

Schuck P (2002) Measuring size-and-shape distributions of protein complexes in solution by sedimentation and dynamic light scattering. Autrans, France, Euroconference "Advances in Analytical Ultracentrifugation and Hydrodynamics"

Schuck P (2007) http://www.analyticalultracentrifugation.com/references.htm

Schuck P (2009) https://sedfitsedphat.nibib.nih.gov/workshop/default.aspx

Schuck P, Demeler B (1999) Direct sedimentation analysis of interference optical data in analytical ultracentrifugation. Biophys J 76:2288–2296

Schuck P, Perugini MA et al (2002) Size-distribution analysis of proteins by analytical ultracentrifugation: strategies and application to model systems. Biophys J 82(2):1096–1111

Scott DJ, Schuck P (2006) A brief introduction to the analytical ultracentrifugation of proteins for beginners. In: Scott DJ, Harding SE, Rowe AJ (eds) Modern analytical ultracentrifugation: techniques and methods. The Royal Society of Chemistry, Cambridge, pp 1–25

Sivakolundu SG, Nourse A et al (2008) Intrinsically unstructured domains of Arf and Hdm2 form bimolecular oligomeric structures in vitro and in vivo. J Mol Biol 384(1):240–254

Sivia DS (1996) Data analysis. A bayesian tutorial. Oxford University Press, Oxford

Straume M, Johnson ML (1992) Analysis of residuals: criteria for determining goodness-of-fit. Methods Enzymol 210:87–105

Wang L, Gilbert RJ et al (2008) Peptidoglycan recognition protein-SD provides versatility of receptor formation in Drosophila immunity. Proc Natl Acad Sci U S A 105(33):11881–11886

Xu W, Qiao S (2008) A divide-and-conquer method for the Takagi factorization. SIAM J Matrix Anal Appl 30(1):142–153

Yamakawa H, Fujii M (1973) Translational friction coefficient of wormlike chains. Macromolecules 6:407–415