Analysis of small-angle scattering data using model fitting and Bayesian regularization

Larsen, Andreas Haahr; Arleth, Lise; Hansen, Steen

Published in:
Journal of Applied Crystallography

DOI:
10.1107/S1600576718008956

Publication date:
2018

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Larsen, A. H., Arleth, L., & Hansen, S. (2018). Analysis of small-angle scattering data using model fitting and Bayesian regularization. *Journal of Applied Crystallography*, *51*(Part: 4), 1151-1161. https://doi.org/10.1107/S1600576718008956

JOURNAL OF
APPLIED
CRYSTALLOGRAPHY

ISSN 1600-5767

Analysis of small-angle scattering data using model fitting and Bayesian regularization

Andreas Haahr Larsen,* Lise Arleth and Steen Hansen

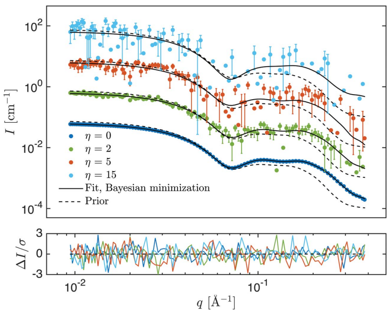
Niels Bohr Institute, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen, Denmark. *Correspondence e-mail: andreas.larsen@nbi.ku.dk

Received 7 February 2018
Accepted 19 June 2018

Edited by D. I. Svergun, European Molecular Biology Laboratory, Hamburg, Germany

Keywords: small-angle scattering; Bayesian regularization; information content; molecular constraints.

The structure of macromolecules can be studied by small-angle scattering (SAS), but as this is an ill-posed problem, prior knowledge about the sample must be included in the analysis. Regularization methods are used for this purpose, as already implemented in indirect Fourier transformation and bead-modeling-based analysis of SAS data, but not yet in the analysis of SAS data with analytical form factors. To fill this gap, a Bayesian regularization method was implemented, where the prior information was quantified as probability distributions for the model parameters and included *via* a functional S. The quantity $Q = \chi^2 + \alpha S$ was then minimized and the value of the regularization parameter α determined by probability maximization. The method was tested on small-angle X-ray scattering data from a sample of nanodiscs and a sample of micelles. The parameters refined with the Bayesian regularization method were closer to the prior values as compared with conventional χ^2 minimization. Moreover, the errors on the refined parameters were generally smaller, owing to the inclusion of prior information. The Bayesian method stabilized the refined values of the fitted model upon addition of noise and can thus be used to retrieve information from data with low signal-to-noise ratio without risk of overfitting. Finally, the method provides a measure for the information content in data, N_g, which represents the effective number of retrievable parameters, taking into account the imposed prior knowledge as well as the noise level in data.



1. Introduction

Small-angle scattering (SAS) is widely used for investigating the low-resolution structure of macromolecules (Svergun & Koch, 2003; Svergun *et al.*, 2013). Physical quantities such as the radius of gyration and molecular weight can be obtained directly from the data, and the overall structure of the macromolecules can be probed indirectly by modeling.

Deducing a structure exclusively from SAS data is an ill-posed problem, meaning that several structures can explain the data. In SAS modeling with analytical form factors, a geometrical model that describes the scattering intensity in terms of a set of model parameters is tested against data (see *e.g.* Pedersen, 1997). Typical parameters include particle dimensions, excess scattering length densities, concentration *etc*. These parameters are then refined to obtain the values that provide the best fit to data. In order to circumvent the ill-posed nature of the problem and minimize the number of free parameters, Hayter & Penfold (1981) introduced molecular constraints in an early small-angle neutron scattering (SANS) study of SDS micelles. This allowed for explicit use of the information available about the SDS chemical structure, the partial specific molecular volumes and the sample concentration, such that the model could be reparametrized into a

https://doi.org/10.1107/S1600576718008956

minimal number of free parameters. The core–shell micelle model and associated interparticle structure factor were reparametrized into a particularly simple model with only two free parameters: the charge and aggregation number of the micelles. The approach of using molecular constraints has been generalized to various later and more complicated applications in SAS (*e.g.* by Cabane *et al.*, 1985; Arleth *et al.*, 1997; Kučerka *et al.*, 2004; Skar-Gislinge & Arleth, 2011). However, the approach may lead to an over-constrained fit where the experimental data cannot be fitted. This will often be the case if one or more of the fixed parameters are slightly wrong. At the same time, all information about the fixed parameters in the new data is ignored. To circumvent these problems, model parameters that, according to Hayter & Penfold (1981), should ideally be well known and kept fixed are instead taken as free parameters. This may, on the other hand, create a situation where the most optimal fit has unrealistic values for central parameters; for example, the fitted concentration could be incompatible with an independent concentration assessment, the shape of the particle unrealistic, or the fitted internal scattering length densities too far from the expected values. If the overall model is trusted, this creates a situation where the scientist has to make a choice: either the inconsistent parameters are fixed, thereby ignoring any information about those parameters in the new data and possibly having to accept a poor fit, or alternatively, the new refined values are trusted, thus effectively ignoring the prior knowledge. Clearly, none of these solutions are optimal and an improved framework for inclusion of the prior knowledge is required.

As will be shown in the following, regularized expressions provide such a framework and can be utilized to include prior knowledge directly in the data analysis. Regularization methods are already used extensively in the analysis of SAS data, for example in indirect Fourier transformation (Glatter, 1977; Svergun, 1992), where a smoothness constraint is imposed on the pair distance distribution function, in *ab initio* modeling (Svergun, 1999), where a compactness constraint is applied to the refined models, and in rigid-body modeling (Petoukhov & Svergun, 2005), where regularization terms prevent overlap of the rigid bodies and ensure that the solution does not diverge significantly from known residue distances. However, to the best of our knowledge they have not been used in the analysis of SAS data modeled with analytical form factors, as proposed in the present work.

In this paper, a regularization method that allows for inclusion of prior knowledge and avoids fixing parameter values is presented. The prior knowledge is quantified as probability distributions, so-called priors. The approach exploits Bayesian statistics, which provides an ideal framework for inclusion of priors in analysis of experimental data. Bayesian methods have been used for decades in the field of image processing (see *e.g.* Gull, 1989; Schultz & Stevenson, 1994) and more recently in the processing of electron microscopy images, as implemented, for example, in the program *RELION* (Scheres, 2012). Moreover, Bayesian statistics is used in the effort of effectively combining experimental data with molecular dynamics simulations, as presented for instance in the recent paper by Shevchuk & Hub (2017).

The second issue treated in the present paper is the quantification of information in data. It is of fundamental interest to assess the information in experimental data and thus be able to optimize the information content under different experimental conditions that may be varied, such as concentration, exposure time and neutron contrast situation (Pedersen *et al.*, 2014), and it will be argued that the 'number of good parameters' N_g constitutes a suitable measure for that purpose. N_g, as introduced by Gull (1989), has been discussed in relation to indirect Fourier transform of SAS data by Müller *et al.* (1996) and by Vestergaard & Hansen (2006), and in the present paper we show how it applies in the context of SAS data analysis using analytical form factors.

2. Theory

In conventional analysis of SAS data with analytical form factors, a mathematical model is hypothesized, which describes the theoretical intensity and can be tested against data (see *e.g.* Pedersen, 1997). The model is expressed in terms of a set of model parameters, for example the particle dimension, the contrast situation, the concentration or the polydispersity of the sample. These parameters are refined by minimizing the likelihood function, χ^2, defined in terms of the theoretical intensities $I^{\rm th}$ and the experimentally measured intensities $I^{\rm exp}$ as

$$\chi^2(\mathbf{p}) = \sum_{i=1}^{N} \frac{\left[I_i^{\rm exp} - I_i^{\rm th}(\mathbf{p})\right]^2}{\sigma_i^2}. \qquad (1)$$

Here, N is the number of data points and σ_i is the experimental standard deviation of data point i. $I_i^{\rm th}(\mathbf{p})$ is assumed to be a function of K model parameters $\mathbf{p} = (p_1, \ldots, p_K)$. Both experimental and theoretical intensities are functions of the momentum transfer, q, given in terms of the wavelength of the incoming beam λ and the scattering angle 2θ, $q = 4\pi\sin(\theta)/\lambda$. The detector image is azimuthally averaged and binned into discrete q values such that the intensity is also discretized, *i.e.* $I_i = I(q_i)$. The reduced χ^2 is used to assess the goodness of fit and is defined as $\chi^2_{\rm r} = \chi^2/f$, where f is the number of degrees of freedom, conventionally found as $f = N - K$. Residual plots are used to evaluate the goodness of fit visually and give the difference in intensity in units of σ, *i.e.* $(\Delta I/\sigma)_i = (I_i^{\rm exp} - I_i^{\rm th})/\sigma_i$.

In the Bayesian approach, the prior knowledge is directly incorporated in the minimization process through a functional, $S(\mathbf{p})$, that gives a penalty to solutions with parameter values far from the prior values. We will assume normally distributed priors with mean values $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$ and standard deviations $\boldsymbol{\delta}\mathbf{p} = (\delta p_1, \ldots, \delta p_K)$. Then $S(\mathbf{p})$ takes the form

$$S(\mathbf{p}) = \sum_{k=1}^{K} \frac{(p_k - \mu_k)^2}{\delta p_k^2}. \qquad (2)$$

μ_k and δp_k reflect the prior knowledge about the kth parameters. If this comes from a measurement, or a previous experiment, a mean and a standard deviation is usually available. If the prior, on the other hand, is based on general biophysical knowledge about the system, this knowledge must be expressed in terms of μ_k and δp_k. If almost no knowledge is available, a mostly non-informative prior should be used, for example a uniform prior or a very wide normal distribution. The determination of priors is exemplified and explained for the two experimental examples in §3. $\chi^2(\mathbf{p})$ is then replaced in the minimization routine by the expression

$$Q(\mathbf{p}) = \chi^2(\mathbf{p}) + \alpha S(\mathbf{p}), \qquad (3)$$

where α is a regularization parameter, balancing the influence of the prior knowledge (S) and the data (χ^2).

2.1. Determining α and introducing the Bayesian Occam term

The Bayesian method provides a consistent way of determining α, the regularization parameter. α is a so-called hyperparameter and must be determined by other means than the model parameters (MacKay, 1999; Hansen, 2000), namely by maximizing the probability for α and the data D given the hypothesized model H. Using standard probability rules, we can express this probability as a product,

$$P(D, \alpha \mid H) = P(D \mid \alpha, H)\, P(\alpha), \qquad (4)$$

where $P(D \mid \alpha, H)$ is the evidence, describing the probability for the data set given both α and the model. For a more elaborate introduction to the evidence and Bayesian probability theory see, for example, Bolstad (2007). $P(\alpha)$ is the prior for α. As α is a so-called scale parameter, Jeffreys' prior, $P(\alpha) = 1/\alpha$ (Jeffreys, 1946), is used in the following. Also, it is exploited that minimizing $-2\log[P(D, \alpha \mid H)]$ is analogous to maximizing $P(D, \alpha \mid H)$. Denoting by $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ the curvature matrices $\mathbf{A} = \nabla\nabla\alpha S$, $\mathbf{B} = \nabla\nabla\chi^2$ and $\mathbf{C} = \nabla\nabla Q$, and denoting by Γ the fraction $\Gamma = \det(\mathbf{C})/\det(\mathbf{A})$, it can be shown that (Hansen, 2000)

$$-2\log[P(D, \alpha \mid H)] = Q(\mathbf{p}) + \log(\Gamma) + 2\log(\alpha), \qquad (5)$$

where $Q(\mathbf{p})$ is defined in equation (3) and the third term is the Jeffreys prior for α. Γ plays a significant role in the analysis: the determinant $\det(\mathbf{A})$ is given as $\alpha(\prod_{j=1}^{K} \delta p_j)^{-2}$, *i.e.* it is inversely proportional to the squared product of the standard deviations of the priors for the model parameters. This product spans the volume in the parameter space where the solution is expected to exist *a priori*. The determinant $\det(\mathbf{C})$ can be written as $\det(\nabla\nabla\chi^2 + \alpha\nabla\nabla S)$, where the curvature matrix $\nabla\nabla\chi^2$ depends on the analytical model and must be found numerically. So the expression cannot be simplified any further in the general case. However, $\det(\mathbf{C})$ is generally inversely proportional to the *a posteriori* solution volume. In summary, $\Gamma \propto$ (*a priori* volume)/(*a posteriori* volume).

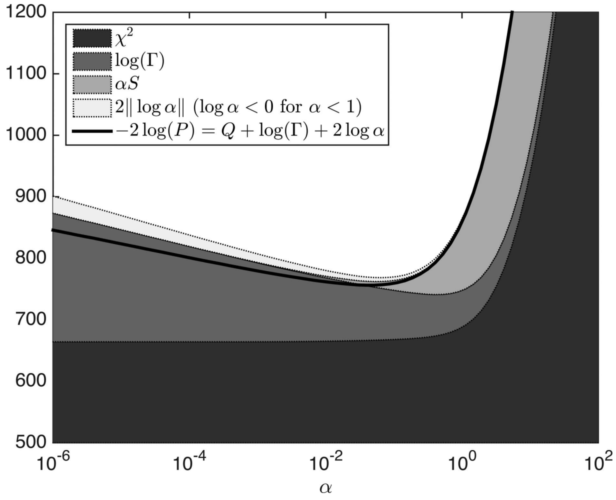


Figure 1
Graphical representation of equation (5) for the nanodisc example. The optimal value of α is found at the minimum ($\alpha = 0.24$). Note that the lower y limit is 500, *i.e.* χ^2 constitutes the major contribution.

In the simplest possible solution where the data contain no new information about the parameters ($\nabla\nabla\chi^2 = 0$), the two volumes are identical, *i.e.* the prior knowledge is not altered, and $\log(\Gamma)$ is zero. Otherwise, the term will be positive, since the *a priori* volume is generally larger than the *a posteriori* volume. Hence, the term favors simple solutions and will be denoted the Occam term (MacKay, 1992). The contributions of all terms of equation (5) are shown graphically for the nanodisc example in Fig. 1, and it is clearly seen how the Occam term 'pushes' the solution towards higher α values, *i.e.* towards simpler solutions closer to the prior.

2.2. Quantifying the information content in data

Following the argumentation in previous work (Gull, 1989; Müller *et al.*, 1996; Vestergaard & Hansen, 2006), the information content can be quantified as the number of good parameters $N_{\rm g}$, describing the effective number of free parameters retrievable by the data. It is defined in terms of α and the eigenvalues η_i and γ_i of the diagonalized curvature matrices $\mathbf{B}$ and $\mathbf{C}$, respectively. By change of units $C_{ij} \rightarrow C_{ij}\delta p_i \delta p_j$, the eigenvalues of $\mathbf{C}$ can be written as $\gamma_i = \alpha + \eta_i$, and $N_{\rm g}$ can then be expressed simply in terms of α and η_i as

$$N_{\rm g} = \sum_{i=1}^{K} \frac{\eta_i}{\gamma_i} = \sum_{i=1}^{K} \frac{\eta_i}{\alpha + \eta_i}, \qquad (6)$$

where K is the number of parameters in the model. The measure is similar in methodology to single value decomposition, *i.e.* the model is, so to say, redescribed in a new basis. The good parameters do not therefore correspond directly to parameters in the investigated model, but $N_{\rm g}$ is the minimum number of independent effective parameters retrievable from the data. The magnitude of η_i (eigenvalue i of $\mathbf{B} = \nabla\nabla\chi^2$) expresses the significance of the ith effective parameter. All eigenvalues are positive, but some are very small compared with α. If an eigenvalue is very large, $\eta_i \gg \alpha$, it will contribute 1 to $N_{\rm g}$, and if $\eta_i \ll \alpha$, then η_i will not contribute to the sum at all. Thus $N_{\rm g}$ is between 0 and K. The information may be distributed evenly among the physical model parameters, but the data may also contain much information about some parameters and very limited information about others. This will be reflected in the difference between the prior and the posterior distribution for each parameter.

3. Methods

3.1. Experimental examples

To test the method, we analyzed the experimental small-angle X-ray scattering (SAXS) data from two different macromolecular samples.

The first sample contained nanodiscs of 1,2-dilauroyl-*sn*-*glycero*-3-phoshocholine (DLPC) and the membrane scaffolding protein MSP1D1, measured at 293 K. The data set was previously obtained and analyzed by Skar-Gislinge *et al.* (2010). The nanodisc is a composite particle consisting of a phospholipid bilayer surrounded by two amphipathic and α-helical scaffolding proteins that form a stabilizing belt around the hydrophobic edge of the bilayer (Fig. 2). Each belt protein has a protruding His tag with a tobacco etch virus (TEV) cleavage site, and these were modeled as random Gaussian coils. The nanodisc itself was modeled by combining analytical form factor amplitudes, as described and illustrated by Skar-Gislinge & Arleth (2011). In brief, the bilayer was described as stacked elliptical cylinders with different scattering length densities, and the two scaffolding proteins were collectively described as a homogeneous hollow cylinder with elliptical cross section. For the purpose of the present work, the model was parametrized to have 12 physically relevant parameters, as listed in Table 1. The parameters were background B, concentration c, molecular volume of the lipids $V_{\rm l}$, molecular volume of the lipid tailgroups $V_{\rm t}$, volume of the protein $V_{\rm p}$, number of lipids per nanodisc N, number of water molecules per lipid headgroup $n_{\rm w}$, thickness of the protein belt T, surface roughness $\sigma_{\rm R}$ [implemented as in the work of Skar-Gislinge *et al.* (2010)], area per lipid A, ellipticity of the disk ε and radius of gyration of the random Gaussian coils $R_{\rm g}$. $V_{\rm l}$ was determined by densitometry with an estimated 2% uncertainty and $V_{\rm t}$ was given by Tanford's formula (Tanford, 1972), also with an estimated uncertainty of 2%, and from these, the volume of the lipid headgroups could be calculated as $V_{\rm h} = V_{\rm l} - V_{\rm t}$. $V_{\rm p}$ was calculated by summing the atomic van der Waals volumes (Svergun *et al.*, 1995), assuming a relative error of 4%. Excess scattering length densities, $\Delta\rho$, were calculated from the molecular volumes and scattering lengths, with the latter calculated from the chemical composition of the relevant molecules. T was known approximately from the α-helical structure of the protein belt, and the priors for A and $n_{\rm w}$ were estimated in accordance with the work of Kučerka *et al.* (2005). SAS experiments on similar systems (Midtgaard *et al.*, 2015; Kynde *et al.*, 2014) were used to estimate the prior for ε. Finally, the prior for $R_{\rm g}$ was estimated from molecular dynamics simulations of proteins with random coil structure by Fitzkee & Rose (2004).

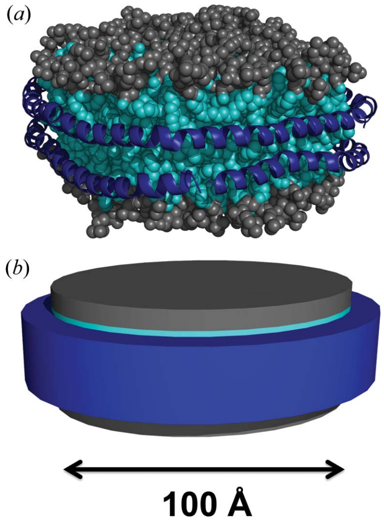


Figure 2
Illustration of a nanodisc. (*a*) All-atom structure from Shih *et al.* (2007), with a hydrophobic core of lipid tails (turquoise), caps of hydrophilic lipid headgroups (gray), and a surrounding 'belt' of two amphipathic and α-helical proteins (blue). (*b*) Analytical nanodisc model with dimensions corresponding to the prior values in Table 1. The His tags with TEV sites were not included in the illustrations.

Table 1
Refined parameter values from the analysis of the nanodisc data set, comparing the Bayesian regularization method with conventional χ^2 minimization.

One standard deviation is given as error (in parentheses). The prior values are listed in the middle column in terms of the mean (and standard deviation) of the respective prior normal distributions. The goodness of the fits were evaluated with the reduced χ^2 and the Cmap test (Franke *et al.*, 2015).

Model parameter	χ^2 minimization	Prior	Bayesian minimization
N	103 (22)	152.0 (10.0)	119 (7)
ε	1.3 (4.5)	1.40 (0.15)	1.33 (0.03)
A (Å^2)	76 (19)	61 (5)	70 (4)
$n_{\rm w}$	18 (12)	8 (2)	10 (3)
$V_{\rm l}$ (Å^3)	996 (19)	985 (30)	1001 (3)
$V_{\rm t}$ (Å^3)	702 (111)	666 (20)	684 (22)
$V_{\rm p}$ ($\times10^4$ Å^3)	5.3 (0.7)	5.7 (0.2)	5.4 (0.2)
T (Å)	11.4 (2.1)	10.0 (0.3)	10.2 (0.6)
$R_{\rm g}$ (Å)	13.8 (1.2)	12.5 (1.0)	14.1 (0.7)
$\sigma_{\rm R}$ (Å)	3.2 (1.0)	6.0 (1.0)	3.3 (0.5)
c (μM)	22 (19)	22.6 (3.0)	23.3 (4.9)
B ($\times10^{-4}$ cm^{-1})	-0.3 (2.1)	1.0 (10.0)	0.1 (1.1)
Goodness of fit, $\chi^2_{\rm r}$	6.26	–	6.30
Goodness of fit, Cmap	$C = 10$, $N = 106$ $P(C \geq 10 \mid N) = 9.2\%$	– –	$C = 10$, $N = 106$ $P(C \geq 10 \mid N) = 9.2\%$

The second example was a sample of self-assembled N-dodecyl-β-maltoside (DDM) micelles, measured at room temperature. The micelles were modeled as core–shell ellipsoids (Pedersen, 1997), using seven parameters, as listed in Table 2. The seven parameters were constant background B, concentration c, scattering contrast of the detergent headgroups in the shell $\Delta\rho_{\rm h}$ and of the detergent tailgroups in the core $\Delta\rho_{\rm t}$, number of detergents per micelle N, ellipticity ε of the micelle, and surface roughness $\sigma_{\rm R}$. The form factor and parametrization are as described by Arleth *et al.* (1997), with a roughness term added, as in the nanodisc model. The partial specific molecular volumes used to determine the scattering

Table 2
Refined parameter values for the micelle data set.

Notation as in Table 1, and b_e is the electron scattering length (2.82 fm).

Model parameter	χ^2 minimization	Prior	Bayesian minimization
N	125.0 (0.3)	130 (15)	125.0 (0.3)
ε	0.5398 (0.0007)	1.00 (0.30)	0.5398 (0.0007)
$\Delta\rho_h$ (b_e Å^{-3})	0.183 (0.033)	0.184 (0.013)	0.184 (0.006)
$\Delta\rho_t$ (b_e Å^{-3})	−0.055 (0.010)	−0.056 (0.006)	−0.056 (0.002)
σ_R (Å)	5.41 (0.03)	6.0 (1.0)	5.41 (0.03)
c (mM)	30.3 (11.0)	30.0 (3.0)	29.8 (1.9)
B ($\times 10^{-3}$ cm^{-1})	0.89 (0.01)	1.0 (10.0)	0.89 (0.01)
Goodness of fit, χ^2_r	170	–	170
Goodness of fit, Cmap	$C = 36$, $N = 90$ $P(C \geq 10 \mid N) \simeq 0\%$	– –	$C = 36$, $N = 90$ $P(C \geq 10 \mid N) \simeq 0\%$

contrasts, $\Delta\rho_h$ and $\Delta\rho_t$, were found with densitometry and the volumes were assumed to have a relative uncertainty of 2% (supporting information of Midtgaard *et al.*, 2018). The priors for N were estimated according to Oliver *et al.* (2013), and the detergent concentration was determined by weighing the added detergent in the stock solution before making the samples, with an estimated uncertainty of 10%.

3.2. Implementation of the Bayesian optimization routine

The Bayesian fitting algorithm was implemented in Fortran 77 and the source code is freely available online (https://github.com/Niels-Bohr-Institute-XNS-StructBiophys/BayesFit). A Levenberg–Marchardt algorithm (Levenberg, 1944; Marquardt, 1963) was used to minimize $Q(\mathbf{p})$. It was implemented with minor modifications of the algorithm from *Numerical Recipes* (Press *et al.*, 1992) and with the parameters constrained to a range defined by the prior mean μ_i and standard deviation δp_i such that $\mu_i - 5\delta p_i < p_i < \mu_i + 5\delta p_i$. A golden section search was used to determine the most probable α, assuming that $-10 < \log(\alpha) < 10$. The CPU time for the refinement of the nanodisc model is about 20 min on a typical PC, searching 17 α values to determine the optimal α. The CPU time for conventional χ^2 minimization is thus 17 times faster, *i.e.* approximately 1 min. The CPU time for the the micelle model is only about 2 min with 19 steps in α (*i.e.* less than 10 s for a χ^2 minimization). Parallelization has not been included in the present implementation but is in principal easy to implement, since the calculations for each q value are independent. With other α-optimization algorithms, the α calculations would also be independent and thus parallelizable, for example with grid search or random search (Bergstra & Bengio, 2012).

4. Results

4.1. Nanodiscs

The Bayesian approach was compared with conventional χ^2 minimization. As seen in Fig. 3(a), both methods found a solution that fitted the data well. The conventional method varied the 12 parameters freely to minimize χ^2, with the mean of the prior values used as the starting point for the fitting routine. In the Bayesian approach, the most probable α was determined, and the parameters were refined as described in §§2 and 3. The optimal α was found at 0.24. Moreover, to monitor the effect of α, a minimization of Q [equation (3)] was performed for a range of logarithmically spaced values of α from 10^{-10} to 10^{10}, and $-2\log[P(D, \alpha \mid H))$ [equation (5)] was calculated at each step.

The refined values of the fitting parameters obtained with both the Bayesian and the χ^2-minimization methods are listed in Table 1. The parameters refined by the Bayesian approach are generally closer to the prior and have smaller uncertainties, as a consequence of including the regularization term. Notice, for example, that the area per lipid headgroup, A, was refined to 70 ± 4 with the Bayesian method (prior value 61 ± 5) as compared to 76 ± 19 with χ^2 minimization, and N was

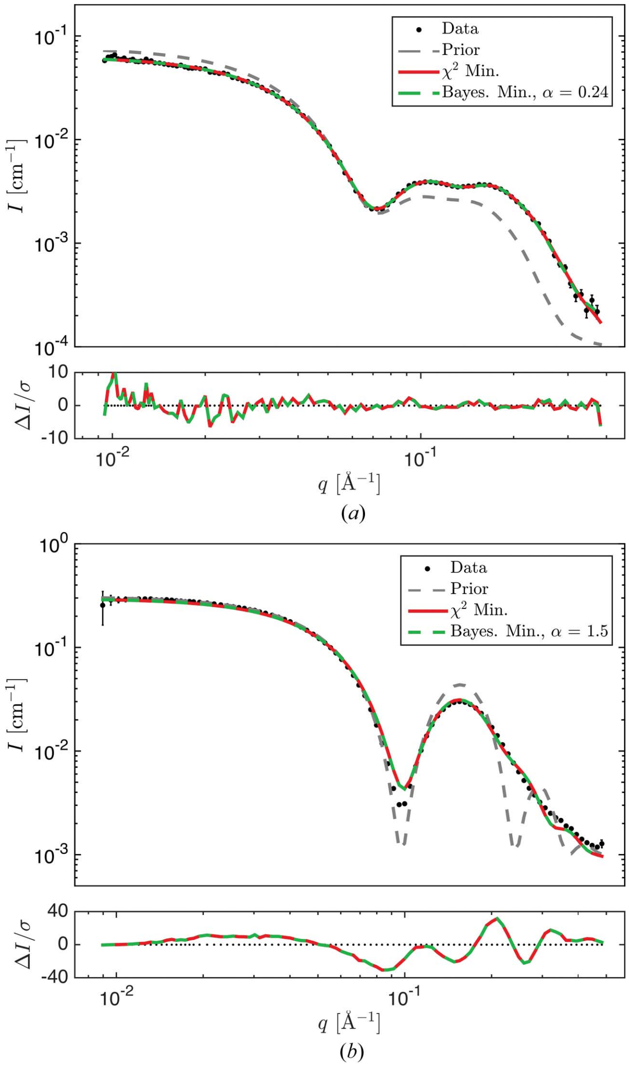


Figure 3
Analyzed examples of SAXS data sets for (a) a nanodisc sample and (b) a sample of detergent micelles. The data sets (black points with error bars) were fitted using conventional χ^2 minimization (red solid line) and Bayesian minimization (green dashed line). The gray dashed line is the prior. Residual plots are shown below, where $\Delta I = I_{\rm exp} - I_{\rm fit}$ and σ is the experimental standard deviation.

refined to, respectively, 119 ± 7 and 103 ± 22 with the Bayesian and the conventional methods (prior value 152 ± 10). These two parameters have been plotted for a range of α values in Figs. 4(*a*) and 4(*b*), and they clearly approach the prior value as α increases. The refined values were thus influenced concurrently by the SAXS data and the prior. In Fig. 5 the prior, likelihood and posterior distributions for N are plotted, clearly showing how the refined value for N using the Bayesian method (posterior distribution) is affected both by the prior and by the likelihood. Figs. 4(*c*) and 4(*d*) show the values of ε and V_1, which were not affected significantly by the prior at the optimal α. Generally, parameters are mostly effected by the prior if, firstly, there is a large discrepancy between the prior mean value and the likelihood value (see Fig 5), secondly, δp (the prior width) is narrow, and, thirdly, the parameters have little effect on χ^2.

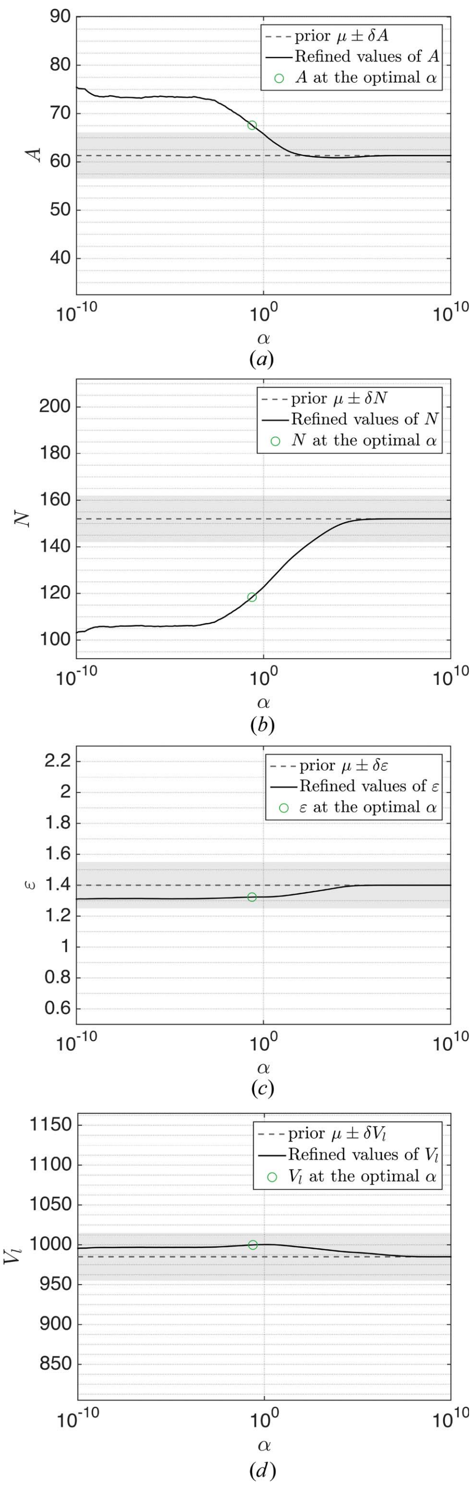


Figure 4
The refined value of four different parameters for the nanodisc model, as a function of α. The refined parameter values for the optimal α are marked by a green ring. The gray dashed line and the gray shaded area show, respectively, the prior mean and the prior standard deviation. Some parameters were significantly altered by the prior, *e.g.* A (*a*) and N (*b*), whereas other parameters were virtually unaffected, *e.g.* ε (*c*) and V_1 (*d*).

4.2. Detergent micelles

In the micelle example, both the χ^2 minimization and the Bayesian minimization found a solution that fitted the data relatively well as judged by visual inspection (Fig. 3*b*), and the regularization parameter, α, was optimized to 1.5. The residual plot reveals some systematic discrepancies. This is verified by a correlation map (Cmap) test (Franke *et al.*, 2015), from which it can be concluded that the data are significantly different from the model [significance level 1%, C = 36, $P(C \geq 36 \mid N = 90) \simeq 0\%$]. The monodisperse prolate ellipsoidal model is thus not a perfect description of the physical micelles, but constitutes an approximate model. In the micelle example the prior had only a minor effect on the fitted results, as seen from Table 2. This means that the global minimum for χ^2 in the parameter space is physically meaningful and consistent with the prior. While the prior hardly affects the model parameters, it does lead to more reasonable errors (Table 2). Note that the concentration had a prior value of 30.0 ± 3.0 m*M*. The error should decrease after taking the

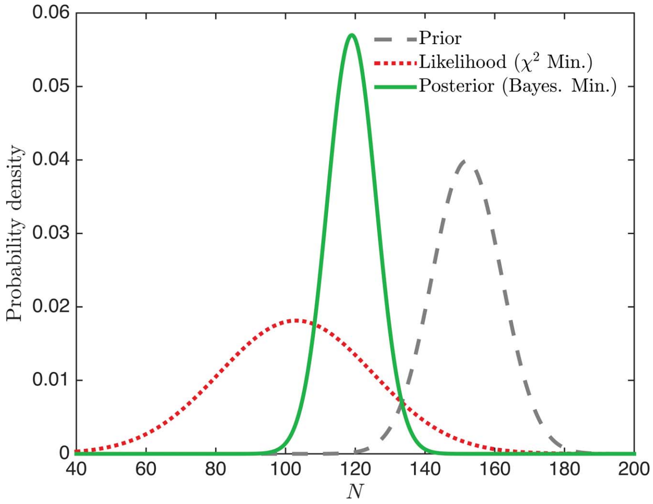


Figure 5
Probability distributions for N in a nanodisc sample. N was refined with χ^2 minimization to obtain the likelihood distribution (red dotted line) and with Bayesian minimization to obtain the posterior distribution (green solid line), which was regularized by the prior distribution (gray dashed line).

SAS data into account, since these data refined the concentration to a value very close to the prior value (30.3 and 29.8 for the conventional and Bayesian methods, respectively). Thus, the error of ±1.9 found with the Bayesian approach is more sensible than the error of ±11.0 found with conventional χ^2 minimization. The same applies for the refined values of $\Delta\rho_{\rm h}$ and $\Delta\rho_{\rm t}$.

4.3. The regularization stabilizes the solution upon addition of noise

Noisy data were simulated with different noise levels to examine the influence of the Bayesian regularization on noisy data. The best fits for the nanodisc and the micelle data sets were used to generate respective simulated data sets. Standard deviations (error bars) were assigned to each point in q by $\sigma(q) = \eta[I_{\rm fit}(q)]^{1/2} + B$, where $I_{\rm fit}(q)$ is the refined fit value found by the Bayesian approach, η is a relative noise parameter and B is a constant noise level, set to $B = 10^{-5}$. The simulated intensities were randomly sampled from a normal distribution with mean $\mu = I_{\rm fit}(q)$ and standard deviation σ. The simulated data and corresponding fits for selected noise levels can be seen in Fig. 6. As in the experimental situation, the prior differs slightly from the simulated data, and it is also plotted in Fig. 6.

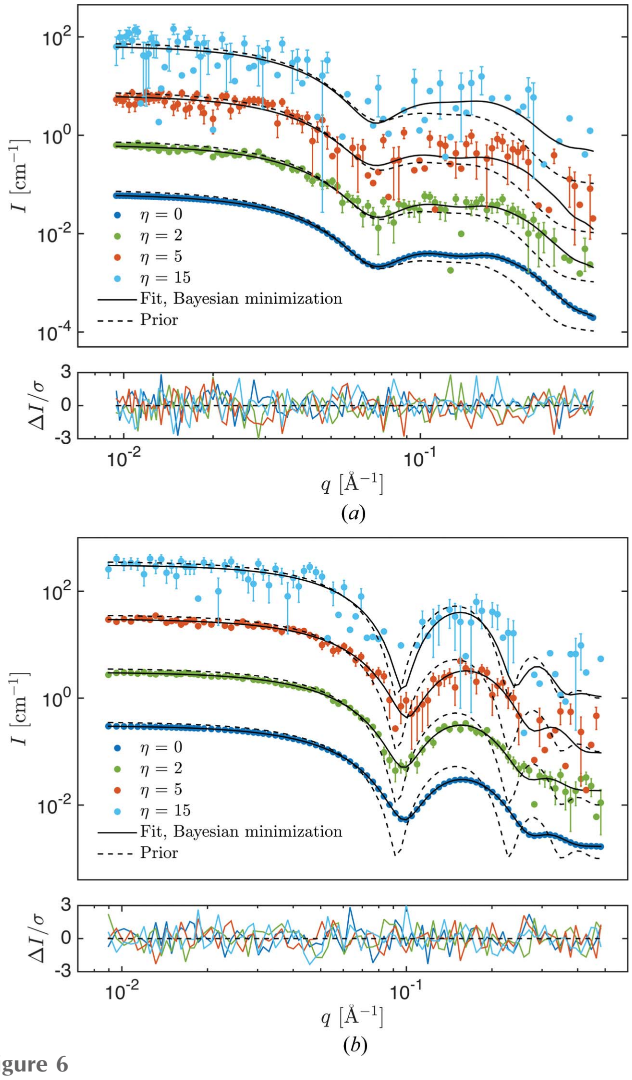


Figure 6
Simulated data with increasing relative noise, $\eta = 0$ (blue), $\eta = 2$ (green), $\eta = 5$ (red) and $\eta = 15$ (cyan). Fit with Bayesian minimization (solid line) and regularized with the prior (dashed line). (a) Simulated nanodisc data and (b) simulated micelle data.

For each noise level, several data sets were generated by random sampling from the normal distribution and fitted with the model, so the variation in the refined parameter values could be evaluated. This is shown for A in Fig. 7, where each point is the mean value of five runs simulated with the same noise level and the error bars are standard deviations. The final refined value of A was stabilized considerably in the Bayesian method as compared to the conventional method, expressed by a nearly constant mean value for all noise levels and small standard deviations.

4.4. The information content in data

The information content for the nanodisc SAXS data, according to equation (6) and given the prior, was $N_{\rm g} = 9.1$, while the number of fitted parameters was 12: that is, 12 parameters were refined, but the information coming from the SAXS data corresponded to nine parameters. The rest of the information came from the prior. For the micelle data set, the information content from the SAXS data was $N_{\rm g} = 6.0$, while the model had seven fitting parameters. Therefore, in both cases, the parameters were refined mainly from the SAXS data and to a lesser degree from the prior. However, when analyzing the simulated data with added noise, the prior played a greater role. In Fig. 8(b), $N_{\rm g}$ is plotted for an increasing value of the relative noise parameter η. $N_{\rm g}$ decreases from around 10 (nanodisc example) and 7 (micelle example) at $\eta = 0$ to $N_{\rm g} < 3$ (both cases) at $\eta = 40$: that is, for noisy data sets, the refined parameters are mainly determined

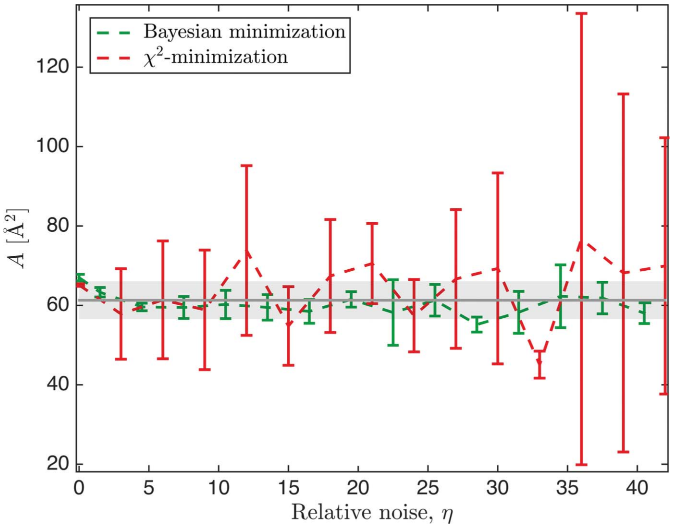


Figure 7
Refined value for the area per headgroup A, found by χ^2 minimization (red) and by Bayesian minimization (green), for increasing relative noise η. The prior, used in the Bayesian minimization, is shown with a gray line for the mean and a gray area for the standard deviation.

by the prior. In accordance with our intuition, this shows that less information can be obtained from noisy data, but intriguingly, it also implies that, since the risk of fitting the noise in data is circumvented by the prior, some information can still be extracted with the Bayesian regularization method, even from very noisy data. This would not be possible with the conventional approach, owing to the large fluctuations of the refined parameter values, as exemplified in Fig. 7. The information content depends on the value of α, *i.e.* on how the prior information is weighted with respect to the new data set. In Fig. 8(a), it is shown how $N_{\rm g}$ decreases as α increases, from $N_{\rm g} \simeq K$ at $\alpha = 10^{-10}$ ($K = 12$ for the nanodisc example and 7 for the micelle example) to $N_{\rm g} \simeq 0$ for $\alpha = 10^{10}$. Large α values give weight to the prior, resulting in a low estimated information content of the new data set.

After having introduced $N_{\rm g}$, it is worth returning to the Occam term from equation (5). This term pushes the algorithm towards solutions with higher α values and closer to the prior parameter values (Fig. 1). Higher α values also imply a smaller $N_{\rm g}$ (Fig. 8a), that is, fewer parameters can be retrieved from the data. Hence, the Occam term favors simpler solutions with fewer effective parameters.

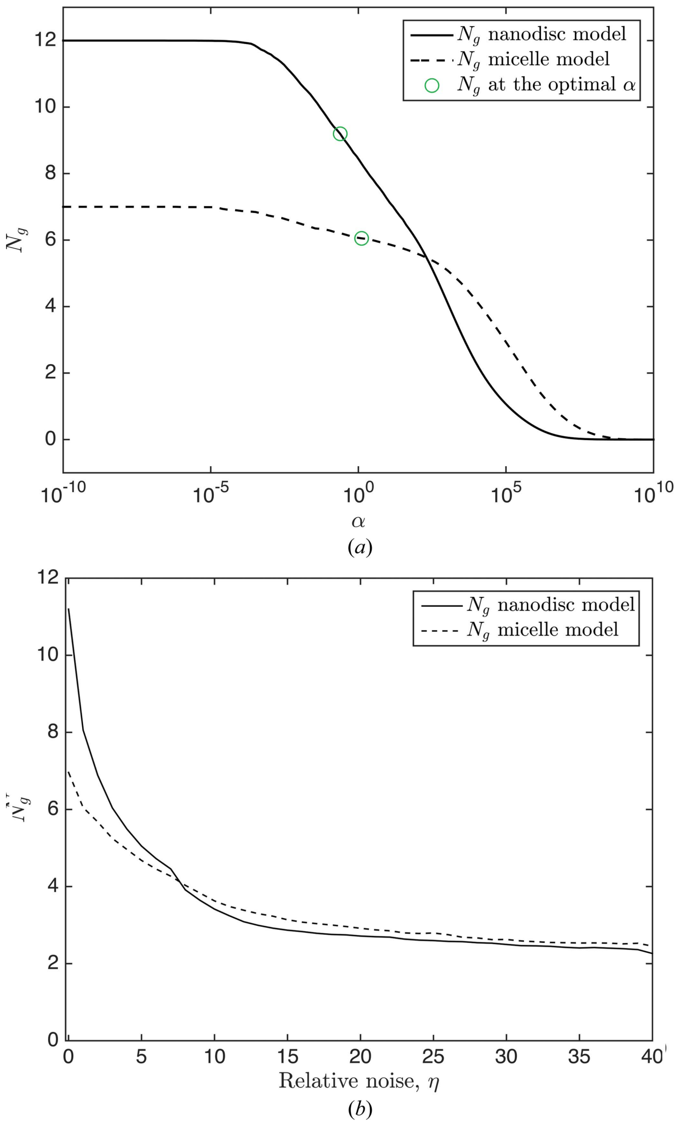


Figure 8
(a) $N_{\rm g}$ as a function of α, with the value for the optimal value of α marked in green. (b) $N_{\rm g}$ for varying noise levels. Each point was a mean for a small range of subsequent values of α (a) or η (b).

5. Discussion

In SAS data analysis with analytical form factors, the prior knowledge can be included *via* molecular constraints as implemented in the parametrization of the hypothesized model. The remaining model parameters are then, in principle, free and can take any value. In practice, however, many parameter values cannot be accepted, owing to inconsistency with the prior knowledge about these parameters, for example from other experiments. This is often accounted for by fixing certain parameters or by setting up limits for the parameter values, *i.e.* not allowing the parameters to exceed a certain range. This is implemented in several commonly used programs for SAS data analysis with analytical form factors, for example *SasView* (http://www.sasview.org), *SASfit* (Breßler *et al.*, 2015), *Scatter* (Förster *et al.*, 2010) and *WillItFit* (Pedersen *et al.*, 2013). It can be argued that this practice corresponds to a Bayesian approach using uniform priors with a finite probability in a given interval and zero probability outside this interval. In the present paper we improve this conventional method by allowing for normally distributed priors that better represent the prior knowledge than uniform priors.

The Bayesian approach is similar to other optimization methods using regularized expressions, but the regularization parameter is here determined automatically and in a statistically sound way, such that a subjective choice of α is avoided.

In a wider perspective, the presented method is a solution to a multi-objective problem (for details see *e.g.* Miettinen, 1998). The objectives are here quantified in terms of the likelihood and the prior functions (χ^2 and S), and the wanted solution is a set of model parameters. The objective functions may be minimized by different sets of model parameters, and the goal is to find the most probable solution taking into account both functions. The χ^2 *versus* S solution space can be divided into two regions, as shown for the nanodisc example in Fig. 9. One region is unreachable since no set of parameters results in these combinations of χ^2 and S values. The other region is reachable, but most solutions here are non-optimal since there exists another set of parameters which is superior with respect to one of the objective functions without being inferior with respect to the others. The border between the regions is denoted the Pareto frontier (Miettinen, 1998). It contains all sets of model parameters that constitute an optimal solution for a given weight between the two objective functions (Pareto optimal sets). A scan over α corresponds to a walk along the Pareto frontier, as indicated in Fig. 9. At $\alpha = 0$, χ^2 is minimized and S takes a a relatively high value. As α increases, S converges towards 0 and χ^2 towards the χ^2 value for the prior solution. Intriguingly, the Pareto frontier is convex for the nanodisc example, meaning that a small perturbation of χ^2 allows a large improvement of S, and *vice versa*. The Pareto frontier for the micelle example is almost single valued, since the same set of parameters minimizes both χ^2 and S. The present method is a so-called scalarization, transforming the multi-objective problem into a single-objective problem with only one solution, namely that for the most probable α.

We have chosen to use Gaussian priors for all parameters, despite the fact that non-Gaussian priors may better represent the knowledge about some of the model parameters. Gaussian priors are, however, computationally economical and simpler to comprehend. The computational speed is relevant, because the Bayesian algorithm needs to refine the model for several values of α to find the most probable solution, thus being 10–20 times slower than conventional χ^2 minimization (depending on the effectiveness of the α-optimization algorithm). For a complex model with two (or more) numerical integrals, such as the nanodisc model, the CPU time can thus extend to 20 min on a standard PC (single core). Considerable speedup can, however, be obtained by parallelization in q.

An inherent problem of the presented method is that it relies on the principle that priors and experimental errors are correctly estimated. Priors may be wrongly estimated, for example because of an erroneous concentration measurement, or errors on refined parameters from previous experiments may be underestimated. A prior for a certain parameter can either be too wide, be too narrow or have a wrong mean value. If the prior is too wide, its effect on the refined value will be underestimated and the errors overestimated. If, on the other hand, a prior is too narrow, it will over-restrict the refined parameter, and the refined error will be underestimated. In the case of a wrong prior mean value, the data will pull the solution far away from this value. Large deviations are thus apparent when comparing the prior with the refined result, so the method constitutes an evaluation of prior assumptions. Generally, a wrongly estimated prior for a given parameter will affect the solution the most if the new data contain relatively little information about that parameter, but will only have a minor effect if the parameter is well determined by the new data. Wrongly estimated priors should, of course, be avoided since inaccurate input will inevitably lead to inaccurate output.

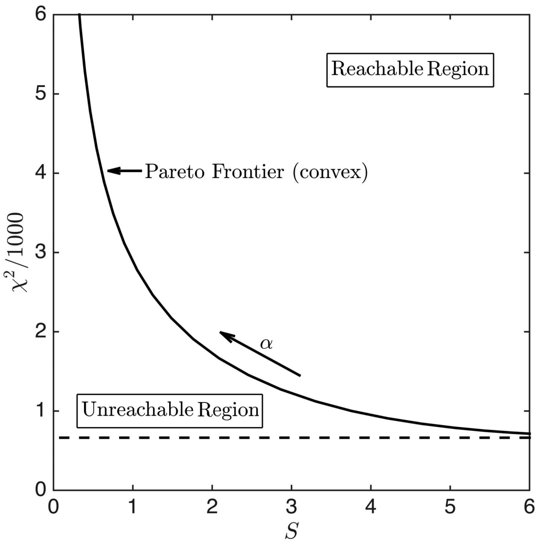


Figure 9
The χ^2 *versus* S space for the nanodisc example. The Pareto frontier (black line) separates the unreachable region and the reachable region. The minimum χ^2 value (dashed line; $\alpha = 0$) and the direction of increasing α are shown. The most probable solution was found at $\alpha = 0.24$, $S = 14$ and $\chi^2 = 668$ (point not included).

The errors on SAS data may likewise be wrongly estimated, as discussed for example by Franke *et al.* (2015) and Rambo & Tainer (2013). In the nanodisc example the fit is good, as judged by visual inspection. However, the residuals (Fig. 3*a*) are expected to be within $\pm 3\sigma$ for a good fit, but in this case reach up to $\pm 10\sigma$. In the same way, $\chi^2_{\rm r}$ is expected to be in the range [0.67, 1.43] (95% confidence interval), but a value of 6.26 was obtained. The size of the experimental errors can be evaluated by indirect Fourier transformation, since data are here fitted with a generic function that should result in a $\chi^2_{\rm r}$ value close to unity. However, a $\chi^2_{\rm r}$ value of 6.6 was obtained in the Bayesian indirect Fourier transformation, thereby indicating that the experimental errors are underestimated. With the Cmap test, the fit could be evaluated independently of the experimental errors. The Cmap test confirmed that the similarity of model and data could not be rejected [significance level of 1%, $C = 10$, $P(C \geq 10 \mid N = 106) = 9.2\%$] and hence confirmed that the experimentally determined error bars were underestimated.

Underestimation of the experimental errors will give too much weight to data (and too little to the prior), since the weight given to data is inversely proportional to the square of the experimental errors [equation (1)]. For a data set with severely over- or underestimated errors, an error correction could therefore be included either separately before the analysis or as an implicit part of the analysis to avoid the effect of erroneously determined experimental errors. We have not included that in the present work because we believe it deserves a more thorough discussion, and it is not a question related specifically to the Bayesian method presented here but affects all methods based on χ^2.

The stabilization of the refined solution upon addition of noise, as exemplified in Fig. 7, shows that the Bayesian regularization method is especially relevant for data with a low signal-to-noise ratio: that is, when sample concentration is limited, for example for protein samples with low-yield expression and samples that are only stable at low concentrations, when exposure time is limited, for exmaple in time-resolved studies, or when flux is limited, for example in SANS and in SAXS at home-source instruments.

The number of degrees of freedom in a SAS data set with q range $q_{\rm max}$–$q_{\rm min}$ and maximum intraparticle distance $D_{\rm max}$ has been described in terms of the number of Shannon channels (Shannon, 1949; Moore, 1980) as $N_{\rm S} = D_{\rm max}(q_{\rm max} - q_{\rm min})/\pi$, provided that $q_{\rm min} < \pi/D_{\rm max}$. $N_{\rm S}$ is widely used to assess the information content in data (*e.g.* Grant *et al.*, 2015). As a measure for the information content, however, $N_{\rm S}$ has the obvious shortcoming that it does not take into account the noise level of data. A solution was proposed by Konarev & Svergun (2015), who introduced an effective number of Shannon channels M_S by truncation of data at high q values with poor signal-to-noise ratio, thus taking into account the noise level of data.

As shown here, and by Pedersen *et al.* (2014), the noise is also effectively taken into account by N_g. Moreover, N_g takes into account the included prior knowledge. Pedersen *et al.* (2014) and Vestergaard & Hansen (2006) used a generic prior, namely that $p(r)$ is a smooth function. In fact, this is the same general information used to estimate M_S. We will in the following denote the number of good parameters obtained with the smoothness constraint by N_g^S (not to be confused with N_S). N_g^S can be calculated with the indirect Fourier transform algorithm in *BayesApp* (http://www.bayesapp.org; Hansen, 2012). The N_g introduced in the present paper uses Gaussian priors for each parameter and will therefore be denoted N_g^G. For the micelle data set $N_g^S = 8.8$ and $N_g^G = 6.0$, and for the nanodisc data set $N_g^S = 7.3$ and $N_g^G = 9.1$: that is, the estimated information content varies with the prior. In the same way, if the Gaussian prior is altered, then N_g^G will change accordingly. To show this, the priors (Tables 1 and 2) were altered by rescaling the prior width with a scale factor ν, *i.e.* $\delta p \rightarrow \nu\,\delta p$, corresponding to a change in the certainty about the priors. N_g^G increases asymptotically as the prior width increases (Fig. 10), *i.e.* when the *a priori* certainty about the parameters decreases. The dependence on prior knowledge is especially evident for repetition series. Here, the first measurement has a relatively high information content, but since that measurement will be included in the updated prior knowledge, the second measurement will contain less information, the third repetition even less, *etc*. At some point, no more measurements need to be taken, since the information content of succeeding measurements would effectively be zero. The prior knowledge has no effect on N_S, which is nevertheless widely used as a measure for the information in data. Therefore, we propose to use N_g^S or M_S instead of N_S to assess the information content in a single SAS data set or a repetition series prior to modeling. After modeling, N_g^G can be used to evaluate the information obtained when SAS is combined with other experimental results and/or other available prior knowledge, as shown in the two examples.

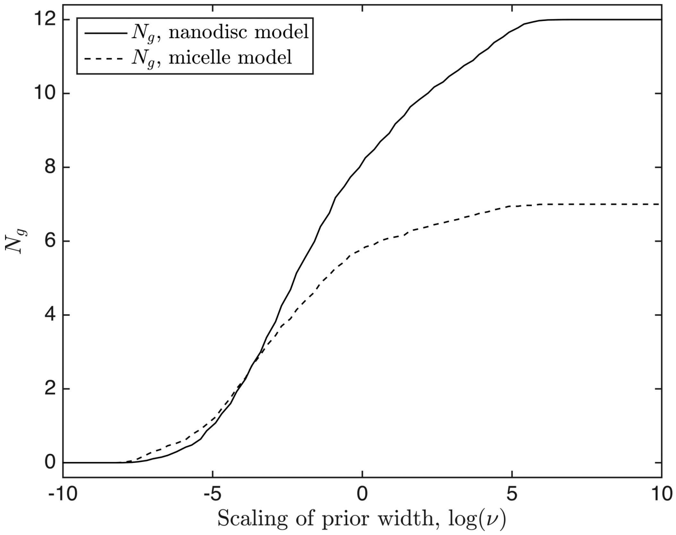


Figure 10
N_g *versus* prior width. The prior width, $\delta\mathbf{p} = (\delta p_1, \ldots, \delta p_K)$, where K is the number of model parameters, was scaled by $\delta\mathbf{p} \rightarrow \nu\,\delta\mathbf{p}$. Each point was a mean for five subsequent values of log(ν).

6. Conclusion

A Bayesian regularization method for SAS data analysis was developed and tested on two data sets: a sample of nanodiscs described by a model with 12 parameters and a sample of detergent micelles described by a model with seven parameters. In both cases, the Bayesian regularization method found a set of model parameters that were physically meaningful without compromising the goodness of fit. The regularization method, furthermore, stabilized the solution when tested against simulated data with increasing noise, thereby preventing overfitting of random noise. This had the important advantage that information could be retrieved even from very noisy data. The method is founded upon probability theory and provides an automatic procedure for weighing the likelihood function χ^2 and the prior function S with respect to each other, by optimizing the regularization parameter α. Moreover, the Bayesian method provides a measure for the information content in data, the number of good parameters N_g, which takes into account both the noise level of the data and the prior knowledge about each model parameter.

Bayesian regularization is generally applicable to inverse problems and is indeed widely applied in many other fields, as mentioned in §1. But, owing to the relatively low information content in SAS data combined with the use of models with multiple parameters, the Bayesian regularization method is of clear relevance for this field.

Acknowledgements

The authors would like to thank Per Hedegård and Martin Cramer Pedersen for fruitful discussions. Thanks to Pie Huda, Søren Roi Midtgaard and Nicholas Skar-Gislinge for providing the example data.

Funding information

Thanks to CoNeXT and University of Copenhagen for co-funding the project.

References

Arleth, L., Posselt, D., Gazeau, D., Larpent, C., Zemb, T., Mortensen, K. & Pedersen, J. S. (1997). *Langmuir*, **13**, 1887–1896.
Bergstra, J. & Bengio, Y. (2012). *J. Mach. Learn. Res.* **13**, 281–305.
Bolstad, W. M. (2007). *Introduction to Bayesian Statistics*, pp. 121–330. Hoboken: John Wiley and Son.
Breßler, I., Kohlbrecher, J. & Thünemann, A. F. (2015). *J. Appl. Cryst.* **48**, 1587–1598.
Cabane, B., Duplessix, R. & Zemb, T. (1985). *J. Phys. Fr.* **46**, 2161–2178.

Fitzkee, N. C. & Rose, G. D. (2004). *Proc. Natl Acad. Sci. USA*, **101**, 12497–12502.

Förster, S., Apostol, L. & Bras, W. (2010). *J. Appl. Cryst.* **43**, 639–646.

Franke, D., Jeffries, C. M. & Svergun, D. I. (2015). *Nat. Methods*, **12**, 419–422.

Glatter, O. (1977). *J. Appl. Cryst.* **10**, 415–421.

Grant, T. D., Luft, J. R., Carter, L. G., Matsui, T., Weiss, T. M., Martel, A. & Snell, E. H. (2015). *Acta Cryst.* D**71**, 45–56.

Gull, S. F. (1989). *Maximum Entropy and Bayesian Methods*, edited by J. Skilling, pp. 53–71. Dordrecht: Springer.

Hansen, S. (2000). *J. Appl. Cryst.* **33**, 1415–1421.

Hansen, S. (2012). *J. Appl. Cryst.* **45**, 566–567.

Hayter, J. B. & Penfold, J. (1981). *J. Chem. Soc. Faraday Trans. 1*, **77**, 1851–1863.

Jeffreys, H. (1946). *Proc. R. Soc. London Ser. A*, **186**, 453–461.

Konarev, P. V. & Svergun, D. I. (2015). *IUCrJ*, **2**, 352–360.

Kučerka, N., Liu, Y., Chu, N., Petrache, H. I., Tristram-Nagle, S. & Nagle, J. F. (2005). *Biophys. J.* **88**, 2626–2637.

Kučerka, N., Nagle, J. F., Feller, S. E. & Balgavý, P. (2004). *Phys. Rev. E*, **69**, 051903.

Kynde, S. A. R., Skar-Gislinge, N., Pedersen, M. C., Midtgaard, S. R., Simonsen, J. B., Schweins, R., Mortensen, K. & Arleth, L. (2014). *Acta Cryst.* D**70**, 371–383.

Levenberg, K. (1944). *Q. Appl. Math.* **2**, 164–168.

MacKay, D. J. C. (1992). *Adv. Neural Inf. Process. Syst.* **4**, 839–846.

MacKay, D. J. C. (1999). *Neural Comput.* **11**, 1035–1068.

Marquardt, D. W. (1963). *J. Soc. Ind. Appl. Math.* **11**, 431–441.

Midtgaard, S. R. *et al.* (2018). *FEBS J.* **285**, 357–371.

Midtgaard, S. R., Pedersen, M. C. & Arleth, L. (2015). *Biophys. J.* **109**, 308–318.

Miettinen, K. (1998). *Nonlinear Multiobjective Optimization.* Boston: Kluwer Academic Publishers.

Moore, P. B. (1980). *J. Appl. Cryst.* **13**, 168–175.

Müller, J. J., Hansen, S. & Pürschel, H.-V. (1996). *J. Appl. Cryst.* **29**, 547–554.

Oliver, R. C., Lipfert, J., Fox, D. A., Lo, R. H., Doniach, S. & Columbus, L. (2013). *PLoS One*, **8**, e62488.

Pedersen, J. S. (1997). *Adv. Colloid Interface Sci.* **70**, 171–210.

Pedersen, M. C., Arleth, L. & Mortensen, K. (2013). *J. Appl. Cryst.* **46**, 1894–1898.

Pedersen, M. C., Hansen, S. L., Markussen, B., Arleth, L. & Mortensen, K. (2014). *J. Appl. Cryst.* **47**, 2000–2010.

Petoukhov, M. V. & Svergun, D. I. (2005). *Biophys. J.* **89**, 1237–1250.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1992). *Numerical Recipes*, pp. 4–93. Cambridge University Press.

Rambo, R. & Tainer, J. A. (2013). *Nature*, **496**, 477–481.

Scheres, S. H. W. (2012). *J. Mol. Biol.* **415**, 406–418.

Schultz, R. R. & Stevenson, R. L. (1994). *IEEE Trans. Image Process.* **3**, 233–242.

Shannon, C. E. (1949). *Proc. IRE*, **37**, 10–21.

Shevchuk, R. & Hub, J. S. (2017). *PLOS Comput. Biol.* **13**, e1005800.

Shih, A. Y., Freddolino, P. L., Sligar, S. G. & Schulten, K. (2007). *Nano Lett.* **7**, 1692–1696.

Skar-Gislinge, N. & Arleth, L. (2011). *Phys. Chem. Chem. Phys.* **13**, 3161–3170.

Skar-Gislinge, N., Simonsen, J. B., Mortensen, K., Feidenhans'l, R., Sligar, S. G., Lindberg Møller, B., Bjørnholm, T. & Arleth, L. (2010). *J. Am. Chem. Soc.* **132**, 13713–13722.

Svergun, D. I. (1992). *J. Appl. Cryst.* **25**, 495–503.

Svergun, D. I. (1999). *Biophys. J.* **76**, 2879–2886.

Svergun, D., Barberato, C. & Koch, M. H. J. (1995). *J. Appl. Cryst.* **28**, 768–773.

Svergun, D. I. & Koch, M. H. J. (2003). *Rep. Prog. Phys.* **66**, 1735–1782.

Svergun, D. I., Koch, M. H. J., Timmins, P. A. & May, R. P. (2013). *Small Angle X-ray and Neutron Scattering from Solutions of Biological Macromolecules.* Oxford University Press.

Tanford, C. (1972). *J. Phys. Chem.* **76**, 3020–3024.

Vestergaard, B. & Hansen, S. (2006). *J. Appl. Cryst.* **39**, 797–804.