4-1-2013

Bayesian Reconstruction of P(r) Directly From Two-Dimensional Detector Images Via a Markov Chain Monte Carlo Method

Sudeshna Paul
Purdue University

Alan M. Friedman
Purdue University

Chris Bailey-Kellogg
Dartmouth College

Bruce Craig
Purdue University

Dartmouth Digital Commons Citation

Paul, Sudeshna; Friedman, Alan M.; Bailey-Kellogg, Chris; and Craig, Bruce, "Bayesian Reconstruction of P(r) Directly From Two-Dimensional Detector Images Via a Markov Chain Monte Carlo Method" (2013). *Dartmouth Scholarship*. 1024.
https://digitalcommons.dartmouth.edu/facoa/1024

Journal of
Applied
Crystallography

ISSN 0021-8898

Received 26 June 2012
Accepted 11 January 2013

Bayesian reconstruction of *P*(*r*) directly from two-dimensional detector images *via* a Markov chain Monte Carlo method

Sudeshna Paul,[a] Alan M. Friedman,[b] Chris Bailey-Kellogg[c] and Bruce A. Craig[a]*

[a]Department of Statistics, Purdue University, 250 North University Street, West Lafayette, IN 47907, USA, [b]Department of Biological Sciences, Markey Center for Structural Biology, Bindley Biosciences Center and the Purdue Cancer Center, Purdue University, West Lafayette, IN 47907, USA, and [c]Department of Computer Science, Dartmouth College, Hanover, NH, USA. Correspondence e-mail: bacraig@purdue.edu

The interatomic distance distribution, $P(r)$, is a valuable tool for evaluating the structure of a molecule in solution and represents the maximum structural information that can be derived from solution scattering data without further assumptions. Most current instrumentation for scattering experiments (typically CCD detectors) generates a finely pixelated two-dimensional image. In continuation of the standard practice with earlier one-dimensional detectors, these images are typically reduced to a one-dimensional profile of scattering intensities, $I(q)$, by circular averaging of the two-dimensional image. Indirect Fourier transformation methods are then used to reconstruct $P(r)$ from $I(q)$. Substantial advantages in data analysis, however, could be achieved by directly estimating the $P(r)$ curve from the two-dimensional images. This article describes a Bayesian framework, using a Markov chain Monte Carlo method, for estimating the parameters of the indirect transform, and thus $P(r)$, directly from the two-dimensional images. Using simulated detector images, it is demonstrated that this method yields $P(r)$ curves nearly identical to the reference $P(r)$. Furthermore, an approach for evaluating spatially correlated errors (such as those that arise from a detector point spread function) is evaluated. Accounting for these errors further improves the precision of the $P(r)$ estimation. Experimental scattering data, where no ground truth reference $P(r)$ is available, are used to demonstrate that this method yields a scattering and detector model that more closely reflects the two-dimensional data, as judged by smaller residuals in cross-validation, than $P(r)$ obtained by indirect transformation of a one-dimensional profile. Finally, the method allows concurrent estimation of the beam center and $D_{\rm max}$, the longest interatomic distance in $P(r)$, as part of the Bayesian Markov chain Monte Carlo method, reducing experimental effort and providing a well defined protocol for these parameters while also allowing estimation of the covariance among all parameters. This method provides parameter estimates of greater precision from the experimental data. The observed improvement in precision for the traditionally problematic $D_{\rm max}$ is particularly noticeable.

1. Introduction

The interatomic distance distribution, $P(r)$, is a valuable tool for evaluating the structure of a molecule from small-angle solution scattering (SAS) data and is a common starting point for three-dimensional shape reconstructions. $P(r)$ is typically reconstructed from a one-dimensional profile of scattering intensities, $I(q)$ [$q = (4\pi/\lambda)\sin\vartheta$, where ϑ is half the scattering angle and λ is the wavelength of the incident radiation], on the basis of the Fourier relationship,

$$P(r) = (2/\pi) r \int\limits_0^\infty q I(q) \sin(qr)\,{\rm d}q. \qquad (1)$$

However, $I(q)$ is observed only in a finite interval [$q_{\rm min}$, $q_{\rm max}$]. Because of the missing information at $q < q_{\rm min}$ and $q > q_{\rm max}$ indirect Fourier transform (IFT) methods are employed to reconstruct $P(r)$. That is, a linear combination of basis functions is used to estimate $P(r)$,

$$P(r) = \sum_{i=1}^{N_{\rm max}} a_i \varphi_i(r), \qquad (2)$$

with the coefficients a_i determined by fitting the linear combination of Fourier-transformed functions, $\{\psi_i(q), i = 1, 2, \ldots, N_{\rm max}\}$, to the observed $I(q)$. Although there are several implementations of indirect transformation employing

doi:10.1107/S002188981300109X

different basis functions and fitting restraints (Glatter, 1977; Moore, 1980; Svergun *et al.*, 1988; Svergun & Koch, 2003), it is well known that this indirect approach is ill-conditioned because of the limited resolution of the scattering data (Svergun *et al.*, 1988). In fact, we have shown that a set of $P(r)$ curves better represents the measured data (Kavathekar *et al.*, 2010). However, as a practical matter, it is often advantageous to determine a single 'best' $P(r)$ curve for use in further analysis.

Current instrumentation for solution scattering employs a two-dimensional detector (typically CCD). These detectors give an orthogonal grid of finely spaced pixels, typically one to two thousand on an edge, that display a radially symmetric scattering pattern around a beam center. To remove the effect of scattering by the solution, a set of scattering images is obtained both for the molecule of interest dissolved in the solution and for the solution only. Typically 10–15 images of each type are generated at current synchrotron sources using short exposures (~1–5 s) to maximize total counts while limiting radiation damage.

To obtain the one-dimensional scattering profile, a beam center for all the data is first determined based on the scattering pattern of a standard sample with strong rings of scattering intensity (typically silver behenate or rat tail collagen). The center is generally determined from calibration images before data collection on the experimental samples and is possibly checked for consistency later. $I(q)$ profiles are then obtained from the experimental images by averaging the pixel intensities in concentric bands around the beam center, and a corresponding profile of standard errors is calculated. The $I(q)$ profiles from the 10–15 images of each type are averaged, and the solution average is subtracted from the molecule-in-solution average yielding a reduced molecule-only scattering profile. The corresponding standard errors are propagated accordingly.

Although well established and productive, these data reduction procedures result in the loss of some information, especially of the spatial relationship among the detector pixels. They are also based on strong *a priori* assumptions about the distribution of pixel intensities (*e.g.* proportionality between the mean and variance of a pixel intensity and also independence between neighboring pixels), which can result in incorrect standard errors and possibly a biased $P(r)$ estimate.

Considering these limitations, we propose determining $P(r)$ directly from the images rather than first reducing the two-dimensional image data to a one-dimensional $I(q)$ profile. Our approach can incorporate any indirect transform method, and the indirect transform coefficients are determined directly from the image pixel values without intermediate reduction. Bayesian inference, enabled by a Markov chain Monte Carlo (MCMC) method, is used to estimate the indirect transform coefficients. MCMC is a class of algorithms based on the construction of a Markov chain (where the future state depends only on the current state) that provide samples from a desired distribution (Brooks *et al.*, 2011). In this case, the MCMC method provides us with samples from the Bayesian posterior distribution of $P(r)$ coefficients. The posterior distribution can be used to summarize a family of $P(r)$ curves consistent with the data or to compute a single 'best' $P(r)$ curve. Using the image data directly allows us to account for spatial correlation among the pixel intensities (and potentially better model the pixel intensity and variance).

A Bayesian approach to the evaluation of solution scattering data has been applied to the estimation of $D_{\max}$, the magnitude of the longest interatomic vector in $P(r)$, and the smoothing parameter for the indirect transformation (Hansen, 2000; Vestergaard & Hansen, 2006). However this approach assumes a fixed form for the probability of these two parameters and this probability is only evaluated at the vector of IFT coefficients that maximizes the desired combination of smoothness and fit to the data. Since it starts with the $I(q)$ curve, this approach also does not consider all the information inherent in the two-dimensional detector images.

Our proposed Bayesian–MCMC approach extends the integration over the entire set of parameters (including $D_{\max}$ and beam center) without presupposing a particular form for their error structure. It allows us to directly estimate $D_{\max}$ and quantify its uncertainty by simply including it as an additional unknown parameter. Finally the location of the beam center can also be included as a model parameter, thereby reducing experimental effort and eliminating the possibility of a systematic difference between the data and calibration images.

Estimating $D_{\max}$, the beam center and the coefficients all at one time allows appropriate estimation of their uncertainty and their interactions (covariances), which can then be propagated into the $P(r)$ curve. Although we have simulated and tested this procedure on solution small-angle X-ray scattering (SAXS), it should apply equally well to solution small-angle neutron scattering, where the improved estimation of uncertainty may prove especially useful because of the lower flux and greater counting errors.

2. Methods

2.1. *P*(*r*) reconstruction from the two-dimensional images

In our Bayesian–MCMC approach (Fig. 1), we start with a molecule-in-solution and a solution-only detector image. These raw images are first normalized by the flux of the incident beam and masked appropriately to remove intensities from the four edges and the beam stop region of the image. Finally, corrections are applied to account for the flatness of the detector and varying distance from sample to pixel.

Initial alignment of the two images utilizes the radially symmetric scattering pattern inherent in each image. We first fit our Bayesian model, which assumes spatial independence between pixels (to be described later), on each image using a reduced data range ($q < 0.05$). At this stage, the integer pixel corresponding to the mean value of the MCMC sample for the beam center is used as that image's center pixel. (Note that, as an alternative, we have also searched for the pixel that minimized the variance within concentric rings around that center; this yielded similar results.) The solution-only image is then aligned with the molecule-in-solution image using their center pixels and subtracted to obtain the molecule-only two-dimensional image.

Using the molecule-only image, we then fit our spatially independent Bayesian model to the entire image to estimate the inherent spatial autocorrelation. After estimating the spatial autocorrelation (see §2.3), we fit a second Bayesian model, which accounts for spatial dependence between pixels, to obtain the final joint posterior distribution of parameters.

Our Bayesian approach uses an indirect transform model and accounts for spatial correlation among the pixels. Let $\mathbf{Y}_{\rm M}$ represent the vector of the n pixel intensities of the molecule-only image. We assume these intensities are multivariate normal with mean intensity vector $\boldsymbol{\mu}_{\rm M}$ and covariance matrix $\Sigma_{\rm M}$. Although any set of basis functions could be used to model the mean intensities, we here employ the basis functions of Moore (1980) for computational ease in this initial development of our approach. Thus for pixel i with associated scattering vector magnitude q_i,

$$\begin{aligned}\mu_{{\rm M},i} &= \sum_{j=1}^{N_{\max}} a_j \psi_j(q_i), \quad \text{where}\\ \psi_j(q) &= \pi j D_{\max}(-1)^{j+1}\sin(qD_{\max})\big[(\pi j)^2-(qD_{\max})^2\big]^{-1}.\end{aligned}\tag{3}$$

Here, $N_{\max}$ is determined by Shannon sampling (Shannon & Weaver, 1949).

To account for the possibility of spatial correlation among pixels, we consider a simultaneous autoregressive (SAR) model and define an $n \times n$ proximity matrix, W, whose entries, w_{ij}, identify neighboring pixels (Cressie, 1993). We assign w_{ij} = 1 if pixels i and j are considered neighbors and w_{ij} = 0 otherwise, and then standardize each row of W by dividing by the sum of the row, $\sum_j w_{ij}$. On the basis of observed experimental correlograms, we consider either 'queen' (lateral and diagonal) or 'rook' (lateral only) neighbors (Besag, 1974). Under the SAR model structure, the covariance matrix of $\mathbf{Y}_{\rm M}$ is

$$\Sigma_{\rm M} = (I-\alpha W)^{-1} V_{\rm M}\big[(I-\alpha W)^{-1}\big]^{\rm T},\tag{4}$$

where $V_{\rm M}$ represents the $n \times n$ diagonal matrix of pixel variances for the molecule-only difference image, and I is the $n \times n$ identity matrix. The parameter α denotes the first-order spatial autocorrelation parameter. The larger this value, the greater the correlation among the pixels.

Since the pixel variances in each image are assumed to be proportional to their expected intensity, the variance in the molecule-only image $V_{\rm M}$ equals $V_{\rm MS} + V_{\rm S}$, the sum of the variances in the observed molecule-in-solution and solution-only images, and is proportional to the sum of their expected intensities, $\hat{\mathbf{Y}}_{\rm MS} + \hat{\mathbf{Y}}_{\rm S}$. Since only the expected intensities in the molecule-only image, $\hat{\mathbf{Y}}_{\rm M}$, are modeled directly and not the separate $\hat{\mathbf{Y}}_{\rm MS}$ and $\hat{\mathbf{Y}}_{\rm S}$ intensities, we assume that $V_{\rm MS} + V_{\rm S}$ is approximately proportional to $2\mathbf{Y}_{\rm MS} - \hat{\mathbf{Y}}_{\rm M}$, where $\mathbf{Y}_{\rm MS}$ are the observed molecule-in-solution intensities. $V_{\rm M}$ thus becomes a function of the parameters used to estimate the mean intensity in the molecule-only image and the observed pixel intensities of the molecule-in-solution image.

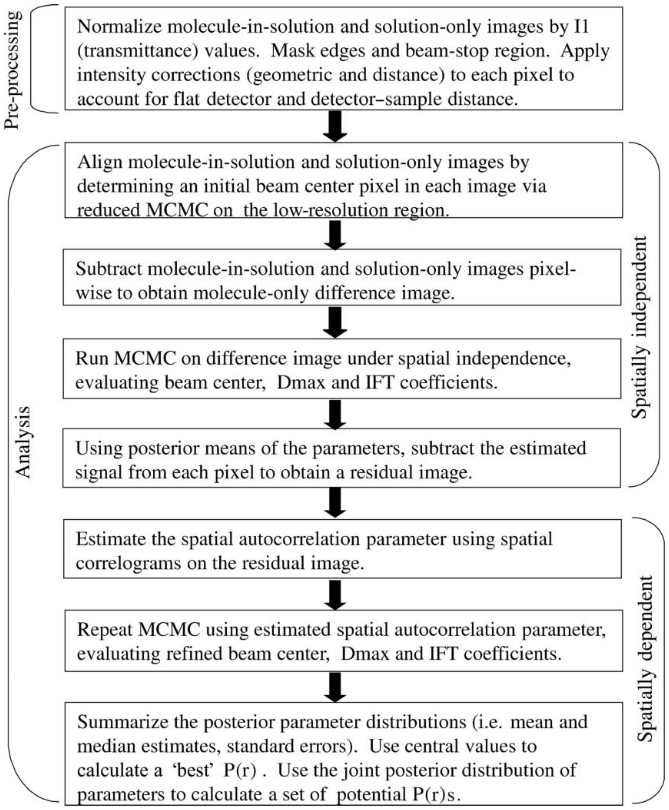
2.2. Estimation of model parameters

The parameters of our spatially dependent model, $\theta = \{a_n\colon n = 1, 2, \ldots, N_{\max}; D_{\max}; (c_x, c_y); \alpha\}$, include the indirect transform coefficients a_n, the maximum length of the molecule $D_{\max}$, the beam center (c_x, c_y) and the spatial autocorrelation parameter α. The exact log-likelihood of a set of parameter values can be expressed as

$$\begin{aligned}\ln\big[f(\mathbf{Y}_{\rm M} \mid \theta)\big] = {}& -\tfrac{1}{2}n\ln(2\pi) + \ln|I-\alpha W| - \tfrac{1}{2}\ln|V_{\rm M}|\\ & -\tfrac{1}{2}(\mathbf{Y}_{\rm M}-\boldsymbol{\mu}_{\rm M})^{\rm T}\Sigma_{\rm M}^{-1}(\mathbf{Y}_{\rm M}-\boldsymbol{\mu}_{\rm M}),\end{aligned}\tag{5}$$

where | | means determinant and $V_{\rm M}$, $\Sigma_{\rm M}^{-1}$ and $\mu_{\rm M}$ are functions of θ. We employ Bayesian inference, enabled by an MCMC method (Gilks *et al.*, 1995), to estimate the joint posterior distribution of all model parameters except the autocorrelation parameter α, which is estimated from the data using a two-step procedure (see §2.3). We adopt relatively non-informative lognormal priors for the indirect transform coefficients, a_n (needed to support positive coefficients for the Moore indirect transform), a more informative Gaussian prior (centered around the center of the physical beam stop) for the beam-center location (c_x, c_y) and a gamma prior for $D_{\max}$ (also positive). Analysis of the resulting MCMC samples showed these choices of priors had little influence on the final estimates; instead the data dominate (results not shown).

As a result of structural constraints in the basis functions, $D_{\max}$ and the first few coefficients (*e.g.* a_1, a_2, a_3) are highly correlated and their individual Metropolis–Hastings updates



Normalize molecule-in-solution and solution-only images by I1 (transmittance) values. Mask edges and beam-stop region. Apply intensity corrections (geometric and distance) to each pixel to account for flat detector and detector–sample distance.

Align molecule-in-solution and solution-only images by determining an initial beam center pixel in each image via reduced MCMC on the low-resolution region.

Subtract molecule-in-solution and solution-only images pixel-wise to obtain molecule-only difference image.

Run MCMC on difference image under spatial independence, evaluating beam center, Dmax and IFT coefficients.

Using posterior means of the parameters, subtract the estimated signal from each pixel to obtain a residual image.

Estimate the spatial autocorrelation parameter using spatial correlograms on the residual image.

Repeat MCMC using estimated spatial autocorrelation parameter, evaluating refined beam center, Dmax and IFT coefficients.

Summarize the posterior parameter distributions (i.e. mean and median estimates, standard errors). Use central values to calculate a 'best' P(r) . Use the joint posterior distribution of parameters to calculate a set of potential P(r)s.

Figure 1
Flowchart of our proposed data preprocessing and methodology.

suffered from poor mixing. To improve this, we adopted an adaptive joint update scheme by sampling from a multivariate normal proposal distribution (Gilks *et al.*, 1995; Hanson & Cunningham, 1998). We start with an MCMC algorithm using individual Metropolis–Hastings parameter updates and run the chain for a sufficient number of iterations ($\geq$25 000), discarding the first 5000 as burn-in. The elements of the covariance matrix of $D_{\max}$, a_1 and a_2 are estimated from the remaining post-burn-in samples and used in the multivariate normal proposal distribution. The chain then continues, using this joint proposal update for these three parameters and the remaining individual updates (further details of proposal distributions and acceptance ratios will be published elsewhere). Only the iterations from the joint proposal phase are used for further analysis. The joint proposal is employed in both the spatially independent and the spatially dependent Bayesian models.

2.3. Estimation of the autocorrelation parameter α

We estimate α and then hold it fixed during MCMC modeling for two reasons. First, simultaneous estimation of the spatial correlation parameter α within our MCMC algorithm is computationally expensive. High-resolution images of dimension 1024 × 1024 (or 2048 × 2048) result in a proximity matrix W of dimension over one (four) million by one (four) million with sparse nonzero off-diagonal terms. Computing $\ln|I - \alpha W|$ and $\Sigma_{\rm M}^{-1}$ in the log-likelihood is then extremely expensive computationally. However if α is fixed, $\ln|I - \alpha W|$ is simply a constant in the log-likelihood and can be ignored. We discuss $\Sigma_{\rm M}^{-1}$ under fixed α in the next section. Second, previous research has shown little change in the distribution of parameters for small differences in the correlation structure (Furrer *et al.*, 2006; Zhang & Du, 2008).

The estimate of α is obtained by first fitting our spatially independent Bayesian model [*i.e.* fix $\alpha = 0$ in equation (5) and estimate the remaining parameters using our MCMC approach] to the entire molecule-only image. The posterior means from this chain are used to calculate the standardized residual matrix. The spatial correlation among these standardized residuals is then quantified using the sp.correlogram function in the *spdep* package (Bivand, 2010) and the chosen neighborhood structure for W. Since second-order effects are weak in the data we have examined, the estimated correlation among neighboring pixels is used as our estimate of α.

2.4. Composite likelihood approximation

Another computationally intensive component in our log-likelihood is $\Sigma_{\rm M}^{-1}$. In order to improve the efficiency of our method, we employ a composite likelihood approximation to equation (5) that is frequently used in modeling high-dimensional geospatial data (Lindsay, 1988; Vecchia, 1988). With this method, the full likelihood is replaced by a product of pixel-specific likelihoods (or sum over pixel-specific log-likelihoods), each evaluated over a much smaller region of the image. Here, these likelihoods are based on the information for each pixel and its direct neighborhood, involving a total of k observed intensities. Thus,

$$\begin{gathered}\ln[f(\mathbf{Y}\mid\theta)] \simeq \sum_{i=1}^{\substack{\text{No. of}\\\text{pixels}}} \ln\big[f_i(\mathbf{Y}\mid\theta)\big], \quad \text{where}\\ f_i(\mathbf{Y}\mid\theta) \propto |\Sigma_{i(k)}|^{-1/2}\exp\big[-\tfrac{1}{2}(\mathbf{Y}_{i(k)} - \hat{\mathbf{Y}}_{i(k)})^{\rm T}\Sigma_{i(k)}^{-1} \qquad (6)\\ \times(\mathbf{Y}_{i(k)} - \hat{\mathbf{Y}}_{i(k)})\big] \quad \text{and}\\ \Sigma_{i(k)} = (I_{i(k)} - \alpha W_{i(k)})^{-1}V_{i(k)}[(I_{i(k)} - \alpha W_{i(k)})^{-1}]^{\rm T}.\end{gathered}$$

When using a 'rook' style neighborhood structure, $k = 5$; for 'queen' style, $k = 9$. We should note that a composite likelihood approach also allows for ready parallelization of this time-consuming step in the MCMC approach (see below).

2.5. Estimation of $D_{\max}$

Estimation of $D_{\max}$ from SAS data has been a challenging problem (Jacques & Trewhella, 2010). Previous methods for $P(r)$ reconstruction have shown that the indirect model parameter estimates and resulting $P(r)$ are sensitive to the choice of $D_{\max}$. Our Bayesian approach allows for easy inclusion of $D_{\max}$ as an additional parameter, thereby avoiding possible biases that can occur when fixing $D_{\max}$ before estimating the other parameters. Including $D_{\max}$ also allows incorporation of the additional uncertainty arising from this parameter into our results.

For the value of $N_{\max}$, the number of coefficients in our indirect transform model, we adopt a procedure similar to Moore (1980) and use Shannon sampling theory (Shannon & Weaver, 1949). Specifically, $N_{\max}$ is the largest integer less than $D_{\max}(q_{\max} - q_{\min})/\pi$. While this approach could potentially result in the number of parameters varying over our MCMC iterations, our experience has been that $D_{\max}$ varies little enough (provided we start with a reasonable initial $D_{\max}$) that $N_{\max}$ remains constant at the initial value.

2.6. Replicating traditional one-dimensional data reduction

For our experimental data analysis comparison, the one-dimensional scattering profile for an image is obtained after preprocessing (see §2.1) by calculating the unweighted average intensity among pixels within concentric rings around the specified beam center determined with a silver behenate standard. The widths of the rings were selected to match the q values employed in a traditional one-dimensional data reduction generated at BioCAT. Standard errors were calculated based on the assumption that the variance was proportional to the mean with a proportionality constant of 10 to roughly match the standard errors found in the reduced profiles generated at BioCAT. The solution-only profile was subtracted from the molecule-in-solution profile and the standard errors propagated to obtain the one-dimensional difference profile.

2.7. Simulating a molecule-only image

An initial $I(q)$ for the simulated data study was developed from the crystal structure of a monomer from orotidine monophosphate decarboxylase complex with XMP (PDB code 1lol; Wu & Pai, 2002). Using a $D_{\max}$ of 58.5 Å and a q range of 0.006–0.47 Å^{-1}, Shannon sampling suggests $N_{\max} = 8$ basis functions to appropriately model $P(r)$. These basis functions were fitted to the initial $I(q)$ curve to obtain the reference indirect transform coefficients, which were then used to generate a reference $P(r)$. These reference coefficient values were also used along with a spatial autocorrelation value of $\alpha = 0.5$, a queen-style neighborhood structure and an assumed beam center of (c_x, c_y) = (50, 50) to generate a 100 × 100 molecule-only simulated image $\mathbf{Y}_{\rm M}$, where

$$\begin{aligned} \mathbf{Y}_{\rm M} &= \boldsymbol{\mu}_{\rm M} + \Sigma_{\rm M}^{1/2}\mathbf{e}, \quad \text{with} \\ \boldsymbol{\mu}_{\rm M} &= \sum_{n=1}^{8} a_n \psi_n(q) \quad \text{and} \qquad (7) \\ \Sigma_{\rm M} &= (I - \alpha W)^{-1} V_{\rm M}[(I - \alpha W)^{-1}]^{\rm T}. \end{aligned}$$

The **e** vector is a set of independent random variables, each drawn from a standard normal distribution. To ensure an error structure similar to experimental data, $V_{\rm M}$ contains only diagonal elements calculated to match the relative errors found in experimental data from a 1.0 mg ml^{-1} sample of a 21 kDa protein collected at the BioCAT undulator beamline 18-ID at the Advanced Photon Source (APS) (Fischetti *et al.*, 2004) fitted with a high-sensitivity CCD detector (Phillips *et al.*, 2002). The diagonal elements of $V_{\rm M}$ are proportional to $\boldsymbol{\mu}_{\rm M}$, with the proportionality constant determined so that the relative standard errors in a ring were comparable to those observed in the real detector image. Computational limitations in simulating the spatial dependency make it difficult to simulate an image with more pixels. While the smaller size proved sufficient for developing and testing our methodology, calculating a one-dimensional reduction from the simulated image for comparison may lead to undersampling.

2.8. Implementation, availability and timing

The calculations were performed using software packages in R (http://www.r-project.org/) and *MATLAB* (The MathWorks Inc., Natick, MA, USA). Our R code is available as supplementary material.[1]

On a dual core 2.8 GHz Xeon processor with 12 Gbytes RAM, spatially independent MCMC simulation on a 1024 × 1024 detector image (25 000 iterations) took 1 h, while the spatially correlated MCMC (25 000 iterations) took 10 h. Since our code was not optimized for efficiency, several changes to the MCMC algorithm would probably improve it. For example, additional group updates of parameters may reduce this time. Also, the time could be shortened by parallelization, either by running several shorter independent Markov chains and combining results or by parallelization of a single chain, in particular the time-consuming inner loop of composite likelihood computations (Jacob *et al.*, 2011).

Table 1
Estimates of MCMC parameters for simulated data obtained under spatially independent and spatially dependent model assumptions.

		Spatially independent			Spatially dependent (α = 0.45)		
Parameter	True value	Mean†	Median†	SD†	Mean†	Median†	SD†
a_1	5.282	5.256	5.256	0.007	5.274	5.274	0.004
a_2	4.178	4.181	4.181	0.002	4.177	4.177	0.001
a_3	1.686	1.698	1.699	0.004	1.689	1.689	0.002
a_4	0.545	0.553	0.553	0.002	0.549	0.549	0.001
a_5	0.337	0.337	0.337	0.001	0.338	0.338	0.000
a_6	0.301	0.301	0.301	0.000	0.301	0.301	0.000
a_7	0.206	0.206	0.206	0.001	0.206	0.206	0.001
a_8	0.084	0.089	0.088	0.001	0.086	0.086	0.001
$D_{\max}$	58.5	58.712	58.720	0.056	58.566	58.567	0.032
c_x	50	50.000	50.000	0.003	50.001	50.001	0.002
c_y	50	49.999	49.999	0.003	50.000	50.000	0.002

† The mean, median and standard deviation of the posterior distribution of each parameter are based on 25 000 MCMC iterations with joint updating.

3. Results

We apply our method first to simulated and then to experimental data to test its performance relative to one-dimensional data reduction. Since one-dimensional data reduction followed by estimation of the basis function coefficients is a proven approach for estimating the IFT coefficients, we are particularly interested in our novel capabilities to estimate the beam center location (c_x, c_y), the maximum length of the molecule $D_{\max}$ and the detector spatial autocorrelation parameter α, in addition to changes in the standard errors of (c_x, c_y), $D_{\max}$ and the IFT coefficients. By calculating a set of potential $P(r)$ curves, we examine the impact of all the variables and their covariances on $P(r)$ reconstruction.

3.1. Simulated data study

Initially our model was fitted to the simulated image data under the assumption of spatial independence and the corresponding standardized residuals were obtained. As described in *Methods*, these residuals were used to determine the first-order autocorrelation parameter α, and then the model was run under spatial dependence to obtain the final parameters. Fig. 2 shows the standardized residuals (left) and the corresponding correlogram plot (right) for both the spatially independent model (top row) and the spatially dependent model (bottom row). Under the independent model, the neighboring residuals display the expected correlation, which quickly dies off with 'distance'. The first-order autocorrelation parameter α is estimated to be 0.45, close to the simulation value of $\alpha = 0.50$. This $\alpha = 0.45$ value was then used to fit our spatially dependent model.

Regardless of the choice of spatial model, all parameters were very close to their reference values (Table 1). The spatially dependent model estimates were more accurate for seven of the 11 parameters as judged by the mean and median of their MCMC distribution. Greater accuracy is seen especially in $D_{\max}$ and the first few indirect transform coefficients. The greater accuracy seen in this single simulation probably arises from the enhanced precision of the spatially dependent

[1] The code discussed in this paper is available from the IUCr electronic archives (Reference: HE5586). Services for accessing this material are described at the back of the journal.

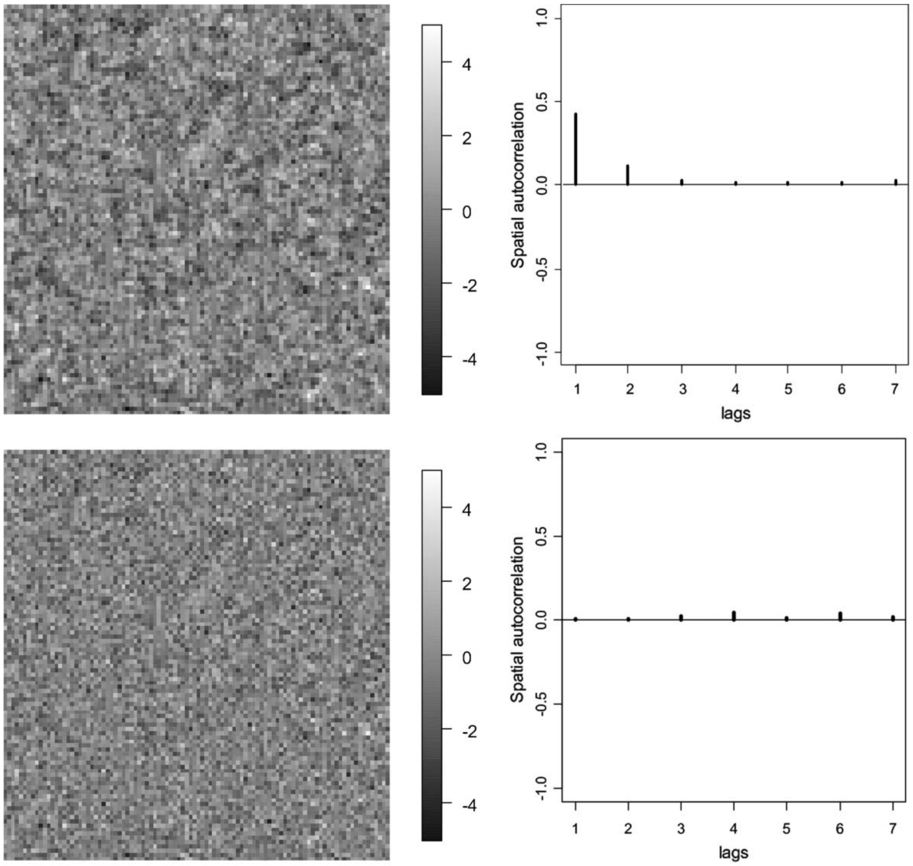


Figure 2
Standardized residuals of the simulated two-dimensional detector image (left) and the spatial correlogram of the standardized residuals (right), showing autocorrelation values for nearest neighbors (lag = 1), second-order neighbors (lag = 2) *etc.* Residuals after running MCMC simulations assuming spatial independence are shown on the top row and residuals after running MCMC simulations with spatial dependence and an autocorrelation parameter of α = 0.45 on the bottom row.

model as reflected in the reduction in uncertainty [Table 1, standard deviation (SD) of the posterior distribution] in the parameter estimates. Improved fit is also seen in the removal of spatial correlation from the standardized residuals (Fig. 2, compare top and bottom rows), although there is no improvement in the mean-square residual since both methods effectively model the underlying simulated signal. The $P(r)$ estimates based on samples of parameters from the spatially dependent MCMC model and the 99% credible interval of the $P(r)$ estimates show small variations around the reference (Fig. 3). Because of the coarse (100 × 100 pixel) sampling of the simulated data used in this developmental-stage simulation, comparison with one-dimensional reduction might be artificially unfavorable to the one-dimensional method. As a result, such comparisons are only made below using our experimental data, which involves images of 1024 × 1024 pixels.

For the spatially independent model, the MCMC chain was started at values far away from the true parameters, and this model demonstrates good convergence and stability (Fig. 4). These profiles also demonstrate strong correlations (both negative and positive) between $D_{\max}$ and the first three direct

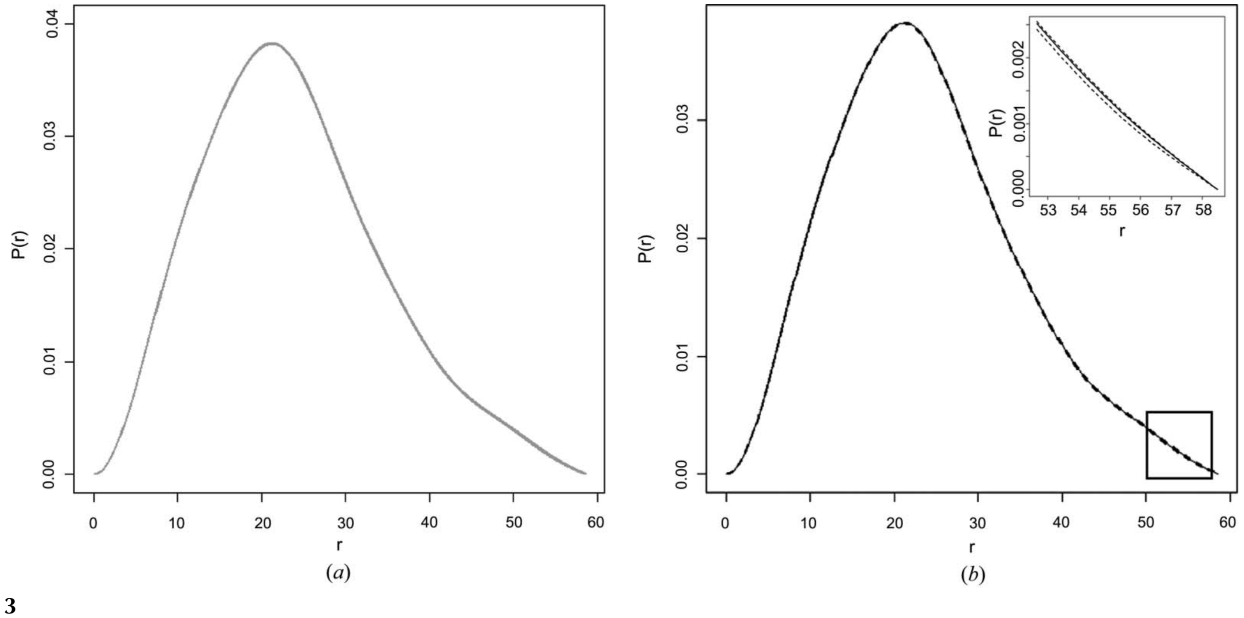


Figure 3
(*a*) A set of 10 000 $P(r)$ curves (gray) calculated from the parameters of the last 10 000 MCMC iterations for the simulated data using the spatially dependent model with α = 0.45. (*b*) The 99% credible interval based on the last 10 000 iterations (dashed lines) observed to encompass the reference $P(r)$ curve (solid gray).

transform coefficients ($r_{1D} = -0.992$, $r_{2D} = 0.890$ and $r_{3D} = 0.990$) but weak correlation among the other parameters. These strong correlations substantially reduce the efficiency of MCMC sampling. To improve mixing of the variables during the later stages of MCMC sampling, we jointly updated $D_{\max}$, a_1 and a_2 using a multivariate normal distribution with a covariance matrix based on the initial MCMC chain. The improvement in sample mixing is evident in the reduction of the correlation between $D_{\max}$ and the first three coefficients (Fig. 5) before and after the joint updating (left and right columns, respectively). Although the joint updating is shown here for the spatially independent model, it was applied to the spatially dependent model as well with an equally good effect.

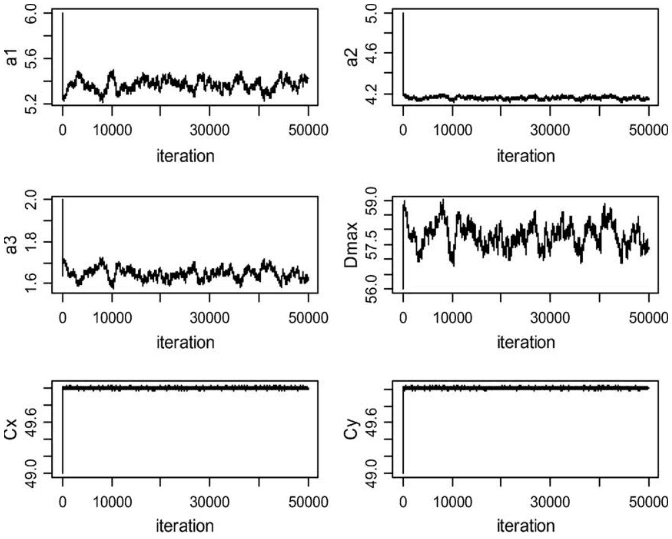


Figure 4
Trace plots of selected parameters of the spatially independent MCMC model using the simulated data. The chains started out at arbitrary initial values but quickly converged to the steady state.

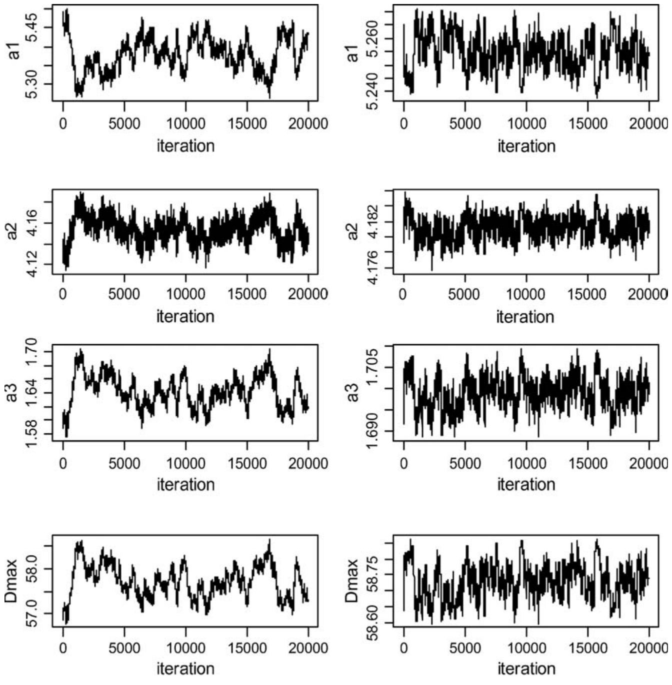


Figure 5
Trace plots of selected parameters of the spatially independent MCMC model using the simulated data after burn-in. The left panel shows the MCMC samples with individual parameter updates and the right panel shows the MCMC samples with a joint update of a_1, a_2 and $D_{\max}$.

3.2. Application to experimental data

As an example of high-quality experimental data, a set of 15 high-resolution (1024 × 1024 pixels) molecule-in-solution and solution-only SAXS images of horse heart myoglobin were collected using a mar165 detector with 2 × 2 binning (Fischetti *et al.*, 2004) installed at BioCAT at the Advanced Photon Source (http://www.bio.aps.anl.gov/facilities.html). A beam center (c_x, c_y) = (138, 362) for these images was obtained by the traditional method using a silver behenate standard. The images provided scattering intensities in a q range of 0.007–0.377 Å^{-1}.

In preparation for our Bayesian–MCMC analysis, the raw images were normalized by the flux of the transmitted beam (I1), the beam stop was masked, and corrections for distance and angle of incidence were applied to the images to correct for a flat detector geometry (Fig. 1). One image each was chosen from the 15 images of the molecule-in-solution and solution-only data sets. For initial alignment, the center pixels of the molecule-in-solution and solution-only images were determined by the MCMC method using only data with $q < 0.05$ Å^{-1}. In this case, the posterior means of the beam center in the two selected images were found to be almost identical. Since the center pixels matched, a difference image (Fig. 6) was then calculated by subtraction of corresponding pixels without adjustment or interpolation.

The one-dimensional scattering profile was calculated (see *Methods*) using the silver behenate center and fitted to Moore's basis functions for indirect transformation using

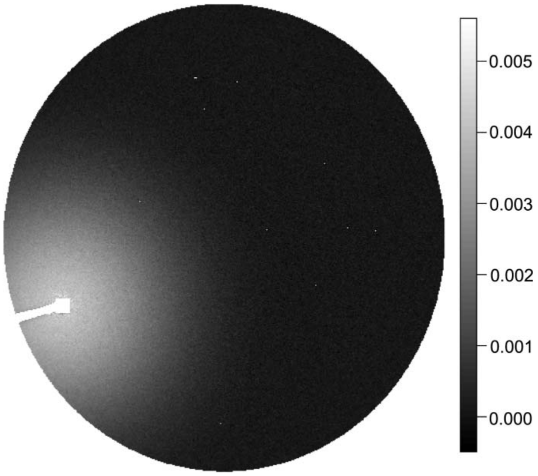


Figure 6
Molecule-only image of myoglobin of size 1024 × 1024 pixels used for MCMC evaluation. This image was obtained by taking 1024 × 1024 molecule-in-solution and solution-only images, applying corrections and normalizing, and then aligning these images to the nearest pixel and taking the difference.

Table 2
Comparison of Moore's indirect transform coefficient estimates for the experimental myoglobin data between our implementation of a weighted least-squares fit to concentric averages of the difference image (one-dimensional approach) and our two-dimensional spatially independent model.

	Concentric averaging		Spatially independent		
Parameters†	Estimate	SE	Mean	Median	SD
a_1	7.546	0.012	7.551	7.551	0.002
a_2	3.844	0.018	3.852	3.852	0.004
a_3	0.838	0.025	0.840	0.840	0.004
a_4	0.759	0.033	0.765	0.765	0.006
a_5	0.827	0.048	0.830	0.830	0.010

† Both approaches assume the same fixed center (c_x, c_y) = (138, 362) and $D_{\rm max}$ = 46.11 Å.

Table 3
Parameter estimates for the experimental myoglobin data fitted using our spatially independent and spatially dependent models (α = 0.5) with adjustable beam center (c_x, c_y) and $D_{\rm max}$.

	Spatially independent model			Spatially dependent model		
Parameters	Mean†	Median†	SD†	Mean†	Median†	SD†
a_1	7.533	7.532	0.013	7.506	7.506	0.008
a_2	3.853	3.854	0.010	3.870	3.871	0.006
a_3	0.841	0.842	0.005	0.846	0.846	0.003
a_4	0.764	0.764	0.007	0.759	0.760	0.004
a_5	0.831	0.831	0.009	0.830	0.831	0.005
C_x	138.654	138.650	0.080	138.568	138.563	0.054
C_y	361.070	361.072	0.070	361.045	361.043	0.038
$D_{\rm max}$	46.176	46.183	0.068	46.308	46.309	0.042

† The mean, median and standard deviation of the posterior distribution of each parameter are based on 25 000 MCMC iterations with joint updating.

various values of $D_{\rm max}$. The value $D_{\rm max}$ = 46.11 was selected on the basis of minimizing the χ^2 statistic for the fit.

To obtain parameter estimates by our MCMC method that would be comparable to those obtained using one-dimensional data reduction and indirect transformation, we first fixed the center at the silver behenate value and $D_{\rm max}$ = 46.11 (the values used for the one-dimensional procedure) and fitted our spatially independent model. This scenario is closest to approximating traditional one-dimensional data reduction, where the center and $D_{\rm max}$ are held constant and correlation among the pixels is ignored. Table 2 summarizes the fitted values and standard errors (SE) for the concentric averaging one-dimensional approach and the mean, median and standard deviation of the posterior distribution for the two-dimensional approach. The different estimates are close to each other, suggesting that under similar conditions the one- and two-dimensional methods will result in almost identical $P(r)$ reconstructions. Since we do not know the underlying true $P(r)$, we cannot say which is better.

The key observable difference between the one- and two-dimensional estimates lies in the uncertainties. The standard deviations of the IFT coefficients under the two-dimensional method are less than half the standard errors arising from the IFT fit under the one-dimensional model. Since an identical relationship between pixel intensity and variance is employed for both, the standard errors can still be compared fairly, even though the two procedures employ the data quite differently. This comparison thus reveals that information is retained and precision improved in the two-dimensional method.

Having completed our comparison to the one-dimensional method, we then fit our spatially independent model to obtain estimates of all parameters including $D_{\rm max}$ and beam center. Standardized residuals from the spatially independent model were used to estimate the spatial correlation. Because of the size of the image, 100 sub-images of size 64 × 64 covering the entire image were randomly selected to estimate the first-order autocorrelation parameter α. These estimates ranged between 0.45 and 0.56, with a mean of 0.50. The mean was used to fit our spatially dependent model.

Table 3 compares the parameter estimates under the spatially independent and spatially dependent models with both adjustable $D_{\rm max}$ and adjustable beam center. Similar to our earlier simulation study, the parameter estimates do not vary substantially between the two models. In that case, knowledge of the reference values showed improved accuracy under spatial dependence, suggesting that the spatially dependent procedure may be performing equally well here. The standard deviations of the parameters are also consistently lower when spatial dependence was incorporated into our model. Specifically, accounting for spatial correlation results in a further 1.5- to twofold reduction in the standard errors of the parameters (Table 3). It is also noteworthy that both analyses suggest a shift in the center of approximately one-half of a pixel in the **x** direction and one pixel in the **y** direction compared to the silver behenate value.

Since it is a persistently problematic parameter (Jacques & Trewhella, 2010), which our method evaluates directly, further examination of $D_{\rm max}$ is warranted. The value of $D_{\rm max}$ under our spatially independent model was estimated to be 46.18 Å with a 95% confidence interval of (46.05, 46.31) *versus* 46.31 Å (46.23, 46.37) under the spatially dependent model and *versus* 46.11 (confidence interval to be estimated below) previously obtained for the one-dimensional approach by minimizing the fit to Moore's basis functions. The value of $D_{\rm max}$ determined from the crystal structure of horse heart myoglobin (PDB code 1wla; Maurus *et al.*, 1997) was 46.49 Å based on all atoms, but this value provides only a crude comparison which may not reflect the state of the molecule and its associated solvent in solution.

While our earlier comparison between the one-dimensional and two-dimensional methods under a fixed beam center and $D_{\rm max}$ revealed significant reduction in the standard errors for the IFT parameters, a more sophisticated analysis is needed here to evaluate the uncertainty in $D_{\rm max}$ under the one-dimensional approach. We have adopted the profile likelihood confidence interval method (Venzon & Moolgavkar, 1988). This approach (Fig. 7) suggests a 95% confidence interval of (44.96, 47.22), which accords with the general intuition of experimenters in the uncertainty of this parameter. This interval (length 2.25 Å) is substantially wider than the credible region under the spatially dependent two-dimensional approach (length 0.14 Å).

The estimated $P(r)$ curves from our spatially dependent method as well as from the one-dimensional data reduction are shown in Fig. 8. The combination of different basis function coefficients and beam center and $D_{\rm max}$ parameters yield slight differences in the resulting $P(r)$ curves.

In the absence of known reference parameter values for comparison, we utilized a single-set cross-validation approach to compare the performance of our model *versus* the one-dimensional data reduction method. A random test sample (N = 17 363) of data points (detector pixels) was generated and parameters estimated (by both approaches) based on the remaining training data. The weighted squared error (WSE) for the predicted value of the pixels in the test data set was computed. The best linear unbiased predictors were used for both methods. This means that the neighboring pixel intensities and estimated correlation structure were utilized in the predictions from our two-dimensional approach, while this information is unavailable in the traditional one-dimensional approach. The WSE under the one-dimensional approach was 4.111×10^{-2}. This was reduced more than twofold using our two-dimensional approach to 1.326×10^{-2}.

It should be noted that the one-dimensional approach is using a $D_{\rm max}$ and beam center very close to those estimated from our two-dimensional approach. Since the estimation of $D_{\rm max}$ can be difficult (Jacques & Trewhella, 2010) and misestimation may artificially favor our method, we have explored the WSE of the test set as a function of $D_{\rm max}$ in the one-dimensional approach and found that the minimum WSE is indeed at the value used for the one-dimensional reduction.

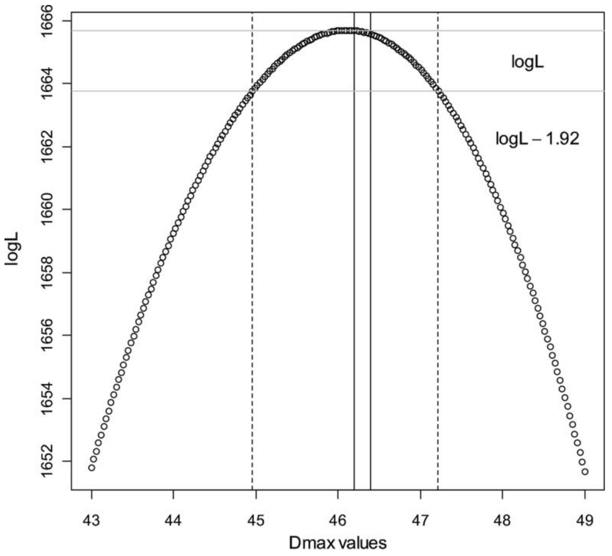


Figure 7
Plot of the profile log-likelihood *versus* $D_{\rm max}$ under the one-dimensional method (circles). For all considered values of $D_{\rm max}$, $N_{\rm max} = 5$. The maximum occurs at $D_{\rm max} = 46.11$. The 95% confidence interval for $D_{\rm max}$ (vertical dashed lines) is (44.96, 47.22). The 95% credible region based on the marginal posterior distribution of $D_{\rm max}$ from the two-dimensional method is (46.23, 46.37) (vertical solid lines).

3.3. Discussion

The Bayesian–MCMC approach that we describe here offers a flexible general framework for reconstructing $P(r)$ curves from SAS images. This approach is based on modeling the two-dimensional scattering images directly and obtaining beam characteristics and parameters for indirect transform basis functions directly from the detector pixels. Comparing the spatially dependent MCMC model with either the one-dimensional approach or the spatially independent MCMC model, we demonstrate here more accurate parameters for simulated data (Table 1), and more precise parameters for (smaller standard errors, Table 2 and Fig. 7) and a better fit (smaller WSE in cross-validation) to the experimental data.

Parameter estimation by our MCMC-based method is completely data dependent, with minimal prior information and restraints on the parameters. Simultaneous estimation of additional parameters including the beam center location (c_x, c_y) and the maximum length in the $P(r)$ distribution $(D_{\rm max})$ becomes possible in our method. As expected, introducing additional parameters increases the variability in the individual parameter estimates (compare Tables 2 and 3). In particular, by introducing $D_{\rm max}$ and the beam center, we see a fivefold increase in the standard error of a_1 and a threefold increase for a_2. However, the greater calculated uncertainty from the additional parameters better represents their true experimental uncertainty and avoids potentially biased estimation of the other parameters. Other beam and detector characteristics (*e.g.* beam divergence and detector tilt) could also be evaluated under this approach.

The adjustment for spatial correlation among neighboring pixels in our image model provides a way to account for the dependence among the pixel intensities. We can accurately recover a spatial correlation value built into simulated data

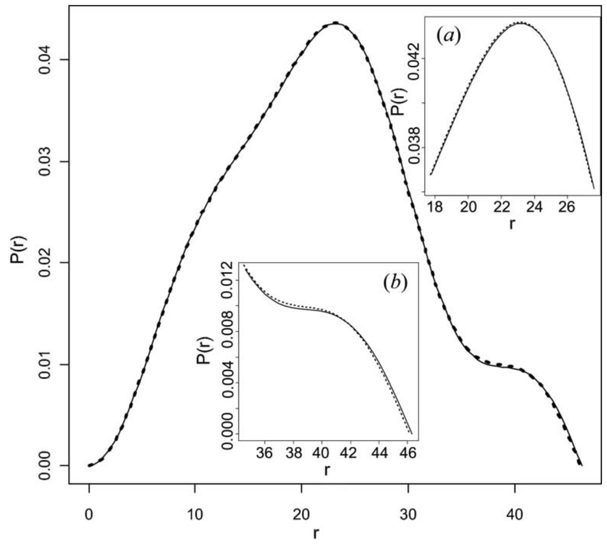


Figure 8
Estimated $P(r)$ using reduced data [(c_x, c_y) = (138, 362), $D_{\rm max}$ = 46.11 fixed] (dashed) and median $P(r)$ curve from the spatially dependent MCMC method (solid). In the insets, (*a*) the peak region ($17 < r < 28$) and (*b*) the (right) tail region ($34 < r < 47$) of the $P(r)$ curves are shown.

and use it to account for the effects on individual pixels. We have also demonstrated that a simple first-order spatially correlated error structure is sufficient to model pixel correlation in at least one real detector. Adjusting for spatially correlated errors (Table 3) that are present in the detector data reduces the standard errors of the individual parameters and thus reduces the width of the credible interval of (the potential uncertainty in) $P(r)$. Results from cross-validation further indicated that the spatially correlated model fitted the data well and resulted in smaller WSE compared to a standard weighted least-squares method based on reduced data.

Although this awaits explicit evaluation, we expect the present MCMC approach to be especially valuable in situations where the measurement error is greater (signal-to-noise ratio is low). This might include low concentrations of materials and weak sources, including neutrons. Furthermore, this approach may help better evaluate the small intensity differences resulting from small structural changes, such as those arising from ligand binding.

A challenge for any method of estimating $P(r)$ is the propagation of errors from the reciprocal space where the data are fitted to the real space where $P(r)$ is calculated. Using Bayesian estimation, we obtain a distribution of acceptable $P(r)$s from the MCMC iterations. This set of acceptable $P(r)$s helps quantify and visualize (Fig. 3) the joint effects of uncertainty in the model parameters on the $P(r)$ estimate. This set of $P(r)$s relates to our earlier work where a set of $P(r)$s is generated from the one-dimensional $I(q)$ curve by a linear programming-based method (Jacques & Trewhella, 2010). Unlike the earlier work, the present MCMC approach also permits calculation of a single 'best' $P(r)$ from central values of the posterior distribution. This 'best' $P(r)$ can be used for further analysis, *e.g.* three-dimensional reconstruction.

One limitation of using the two-dimensional images coupled to an MCMC-based parameter estimation procedure is that it is computationally more intensive in both time and memory. We develop here several approaches that reduce the computational complexity, including joint parameter updates using a multivariate normal proposal distribution, estimation of the autocorrelation parameter prior to MCMC sampling and the use of the composite likelihood. As a result, important additional parameters can be estimated and their uncertainty accounted for in the estimation of $P(r)$ at relatively low computing costs. An extensive MCMC evaluation of the real detector image can be done in less than a day on readily available software (even without parallelization).

We adopted Moore's basis functions to demonstrate our image-based spatial model as they are directly tied to $D_{\max}$ and are easily implemented. Our method is general, however, and in the future, we plan to utilize more sophisticated basis functions (Glatter, 1977; Svergun *et al.*, 1988) in our MCMC approach to improve $P(r)$ estimation directly from image data.

3.4. Conclusions

We have developed a methodology to directly model SAS image data to obtain the parameters needed for $P(r)$ estimation. Using simulated and experimental data, we show improvements in parameter accuracy and/or reproducibility. We also show the ability to evaluate spatial correlation among pixels and an improved fit to the pixel data. Finally, we demonstrate the ability to fit additional beam and detector parameters not previously evaluated directly from the data. While promising, the practical application of the method awaits further testing, especially in molecules of different shapes and sizes and under different error regimes.

We would like to thank Liang Guo, of Biological, Chemical and Physical Sciences at Argonne National Laboratory, for helpful discussions and supplying us with the myoglobin data images. We would also like to thank Bo Li of Purdue University for her advice on improving computational efficiency. We gratefully acknowledge support for this work from a grant from the National Science Foundation SEIII (IIS-0502801) to AMF, BAC and CBK. Use of the Advanced Photon Source, an Office of Science user facility operated for the US Department of Energy (DOE) Office of Science by Argonne National Laboratory, was supported by the US DOE under contract No. DE-AC02-06CH11357. Support for BioCAT was provided by grants from the National Center for Research Resources (2P41RR008630-17) and the National Institute of General Medical Sciences (9 P41 GM103622-17) from the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily reflect the official views of the National Center for Research Resources, National Institute of General Medical Sciences or the National Institutes of Health.

References

Besag, J. (1974). *J. R. Stat. Soc. Ser. B*, **36**, 192–236.

Bivand, R. (2010). *spdep: Spatial Dependence: Weighting Schemes, Statistics and Models.* Version 0.4−58. http://cran.r-project.org/web/packages/spdep.

Brooks, S., Gelman, A., Jones, G. & Meng, X.-L. (2011). Editors. *Handbook of Markov Chain Monte Carlo*. Boca Raton: CRC Press.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. New York: Wiley-Interscience.

Fischetti, R., Stepanov, S., Rosenbaum, G., Barrea, R., Black, E., Gore, D., Heurich, R., Kondrashkina, E., Kropf, A. J., Wang, S., Zhang, K., Irving, T. C. & Bunker, G. B. (2004). *J. Synchrotron Rad.* **11**, 399–405.

Furrer, R., Genton, M. G. & Nychka, D. (2006). *J. Comput. Graph. Stat.* **15**, 502–523.

Gilks, W. R., Richardson, S. & Spiegelhalter, D. (1995). *Markov Chain Monte Carlo in Practice: Interdisciplinary Statistics.* Boca Raton: Chapman and Hall, CRC Press.

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

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

Hanson, K. & Cunningham, G. S. (1998). *Proc. SPIE*, **3338**, 371–382.

Jacob, P., Robert, C. P. & Smith, M. H. (2011). *J. Comput. Graph. Stat.* **20**, 616–635.

Jacques, D. A. & Trewhella, J. (2010). *Protein Sci.* **19**, 642–657.

Kavathekar, P. A., Craig, B. A., Friedman, A. M., Bailey-Kellogg, C. & Balkcom, D. J. (2010). *J. Bioinf. Comput. Biol.* **8**, 315–335.

Lindsay, B. G. (1988). *Contemp. Math.* **80**, 221–239.

Maurus, R., Overall, C. M., Bogumil, R., Luo, Y., Mauk, A. G., Smith, M. & Brayer, G. D. (1997). *Biochim. Biophys. Acta*, **1341**, 1–13.

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

Phillips, W. C., Stewart, A., Stanton, M., Naday, I. & Ingersoll, C. (2002). *J. Synchrotron Rad.* **9**, 36–43.

Shannon, C. E. & Weaver, W. (1949). *The Mathematical Theory of Communication.* Urbana: University of Illinois Press.

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

Svergun, D. I., Semenyuk, A. V. & Feigin, L. A. (1988). *Acta Cryst.* A**44**, 244–250.

Vecchia, A. V. (1988). *J. R. Stat. Soc. Ser. B*, **50**, 297–312.

Venzon, D. J. & Moolgavkar, S. H. (1988). *Appl. Stat.* **37**, 87–94.

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

Wu, N. & Pai, E. F. (2002). *J. Biol. Chem.* **277**, 28080–28087.

Zhang, H. & Du, J. (2008). *Covariance Tapering in Spatial Statistics, Positive Definite Functions: From Schoenberg to Space–Time Challenges*, edited by J. Mateu & E. Porcu. Onda: Gráficas Castañ.