Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Introducing robustness to maximum-likelihood refinement of electron-microsopy data

Sjors H. W. Scheres* and José-María Carazo

Centro Nacional de Biotecnología – CSIC, Darwin 3, Cantoblanco, 28049 Madrid, Spain

Correspondence e-mail: scheres@cnb.csic.es

Received 23 July 2008
Accepted 31 March 2009

An expectation-maximization algorithm for maximum-likelihood refinement of electron-microscopy images is presented that is based on fitting mixtures of multivariate *t*-distributions. The novel algorithm has intrinsic characteristics for providing robustness against atypical observations in the data, which is illustrated using an experimental test set with artificially generated outliers. Tests on experimental data revealed only minor differences in two-dimensional classifications, while three-dimensional classification with the new algorithm gave stronger elongation factor G density in the corresponding class of a structurally heterogeneous ribosome data set than the conventional algorithm for Gaussian mixtures.

1. Introduction

Whereas maximum-likelihood approaches have become a gold standard in many areas of macromolecular X-ray crystallography (*e.g.* Bricogne, 1997; de La Fortelle & Bricogne, 1997; Read, 2001; Blanc *et al.*, 2004), in single-particle three-dimensional electron microscopy (3D-EM) such statistical approaches have only recently been started to be explored. An important characteristic of the maximum-likelihood approach is the natural way in which one may model the experimental noise in the data. Because the noise levels in 3D-EM data are typically extremely high, one would expect that 3D-EM refinement problems could greatly benefit from a proper error model. However, for many years image processing in 3D-EM has been addressed using methods that do not take the noisy character of the experimental data into account in a statistical way (Frank, 2006). Provencher and Vogel performed early work on a statistical model for the noise in 3D-EM data (Provencher & Vogel, 1988; Vogel & Provencher, 1988) and it was only in 1998 that Sigworth introduced a maximum-likelihood algorithm for the alignment of a set of two-dimensional images (Sigworth, 1998). Thereafter, Doerschuk and coworkers used the same principles for the three-dimensional reconstruction of icosahedral viruses (Doerschuk & Johnson, 2000; Yin *et al.*, 2003; Lee *et al.*, 2007), Pascual-Montano and coworkers introduced a maximum-likelihood algorithm for self-organizing maps (Pascual-Montano *et al.*, 2001) and Zeng and coworkers applied this approach to two-dimensional alignment of crystal images (Zeng *et al.*, 2007).

We were the first to address the problem of simultaneous two-dimensional image alignment and classification using maximum-likelihood principles (Scheres, Valle, Nuñez *et al.*, 2005) and we then extended this methodology to the general

case of three-dimensional reconstruction from structurally heterogeneous data (Scheres, Gao *et al.*, 2007). The latter is of special relevance for 3D-EM single-particle analysis, in which many samples constitute large and flexible macromolecular complexes. These complexes typically adopt multiple conformations that are often directly related to their function in living organisms. In principle, provided that one can sort the projections from distinct structures using a computer, multiple three-dimensional reconstructions of the particles in their distinct functional states may be obtained from a single 3D-EM experiment. However, this sorting is strongly intertwined with the orientational assignment of the projections and at present still represents one of the major challenges in single-particle image processing (Leschziner & Nogales, 2007).

We model structurally heterogeneous data as a finite mixture and treat the unavailable information about the orientation and the structural class of each experimental projection as missing data. We then tackle the mixture problem using expectation maximization, which can be shown to converge to the maximum-likelihood estimation of the mixture parameters under relatively mild conditions (Dempster *et al.*, 1977; McLachlan & Peel, 2000). The resulting algorithm is a multi-reference refinement procedure which is similar to conventional refinement approaches in the field (Radermacher, 1994; Penczek *et al.*, 1994). However, the most important difference of the maximum-likelihood approach is that the underlying statistical data model allows one to marginalize over the missing variables. That is, whereas conventional approaches assign a single orientation and class membership to each projection, the maximum-likelihood approach calculates probability-weighted assignments for all possibilities. This provides an intrinsic stabilization of the possibly unstable reconstruction problem. Together with the typical use of relatively small images (also to reduce computational costs) and an early stopping criterion in the underlying algrebraic reconstruction algorithm with smooth basis functions, or blobs (Marabini *et al.*, 1998; Scheres, Gao *et al.*, 2007), this yields a stable algorithm in practice which has been shown to be highly effective on multiple occasions (see, for example, Nickell *et al.*, 2007; Cuellar *et al.*, 2008; Julián *et al.*, 2008; Rehmann *et al.*, 2008).

Despite the importance of the underlying data model in statistical approaches, little work has been performed to explore alternative models for maximum-likelihood approaches in 3D-EM. All the approaches mentioned above share the assumption of additive white Gaussian noise in real space. A large part of the noise may result from shot noise owing to the small number of imaging electrons (10–20 per squared angstrom). The latter would require a multiplicative noise model with a Poisson distribution. However, in practice the additive Gaussian model is a good approximation when each pixel represents many squared angstroms (Sigworth, 2004). The pixel areas for the classifications described in this paper, for example, range from 12 to 30 squared angstroms. Moreover, several additional sources of noise exist such as structural noise arising from the surrounding ice and detector noise; the combination of these multiple independent sources of noise has been shown to follow a Gaussian distribution (Sorzano, de la Fraga *et al.*, 2004). The additive character of the Gaussian noise model results in a computationally attractive algorithm, but the assumption of whiteness is known to be a poor one for electron-microscopy projections. Therefore, we recently introduced an alternative data model in reciprocal space that allows the modelling of nonwhite, or coloured, Gaussian noise (Scheres, Nunez-Ramirez *et al.*, 2007). The Gaussian distribution still remains a common factor, while in other pattern-recognition fields a notable interest has developed in the use of alternative distributions. For many applied problems the tails of the Gaussian are shorter than required and mixtures of Gaussians may lack robustness in the presence of atypical observations. In particular, the use of multivariate *t*-distributions has repeatedly been proposed as a more robust alternative. The *t*-distribution has wider tails and its degree of freedom ν essentially plays the role of rejecting atypical observations. As ν tends to infinity, the *t*-distribution approaches the Gaussian, so that ν may be viewed as a robustness tuning parameter. Several contributions defining frameworks of expectation-maximization algorithms for mixtures of *t*-distributions have appeared and mixtures of *t*-distributions have been successfully applied to a range of different types of data (Lange *et al.*, 1989; McLachlan & Peel, 2000; Wang *et al.*, 2004).

In this contribution, we explore the suitability of modelling structurally heterogeneous 3D-EM data as a mixture of multivariate *t*-distributions. We derive the corresponding expectation-maximization algorithm in §2. In §3 we illustrate its intrinsic properties of providing robustness against outliers and compare the performance of the new algorithm with the conventional algorithm for Gaussian mixtures in two-dimensional and three-dimensional classification. We conclude this paper with a discussion on the potential usefulness of the proposed algorithm in §4.

2. Methods

2.1. The optimization problem

We model two-dimensional images $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$ as follows:

$$\mathbf{X}_i = R_{\Phi_i}\mathbf{V}_{\kappa_i} + \mathbf{G}_i, \qquad (1)$$

where

(i) $\mathbf{X}_i \in \mathbb{R}^J$ are the recorded data. Typical data sets comprise $N = 20\,000$ to $N = 200\,000$ images with $J = 50 \times 50$ up to $J = 100 \times 100$ pixels.

(ii) $\mathbf{G}_i \in \mathbb{R}^J$ is independent zero-mean (additive) noise.

(iii) κ_i is a random integer with possible values $1, 2, \ldots, K$. There are then K unknown three-dimensional structures, $\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_\kappa$ Note that K is fixed, *i.e.* user-determined.

(iv) $R_{\Phi_i}\mathbf{V}_{K_i} \in \mathbb{R}^J$ are the two-dimensional projection data (uncontaminated by noise) of the unknown object $\mathbf{V}_{\kappa_i}$ in an unknown random orientation in space and position in the plane. We parametrize the unknown orientation and position in the plane by a discretized distribution of $p = 1, \ldots, P$

projection directions (described by P combinations of two Euler angles) and $q = 1, \ldots, Q$ in-plane transformations consisting of $Q_{\rm rot}$ rotations and $Q_{\rm trans}$ translations and the corresponding discrete transformations are denoted as R_{pq}.

The reconstruction problem at hand is to estimate $\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_K$ from the observed data $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$. We view this estimation problem as a missing data problem, where the missing data associated with the observed data elements $\mathbf{X}_i$ are the position Φ_i and the random index κ_i. Thus, the complete data set is

$$(\mathbf{X}_i, \Phi_i, \kappa_i) \quad i = 1, 2, \ldots, N. \tag{2}$$

We solve the reconstruction problem by way of maximum-likelihood estimation, where we aim to find those parameters Θ^* that maximize the logarithm of the joint probability of observing the entire set of images $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$:

$$\Theta^* = \arg\max_{\Theta} \sum_{i=1}^{N} \log \sum_{k=1}^{K} \sum_{p=1}^{P} \sum_{q=1}^{Q} P(\mathbf{X}_i|k, p, q, \Theta) P(k, p, q|\Theta). \tag{3}$$

As described previously (see, for example, the Supplementary Note in Scheres, Gao *et al.*, 2007), we assume that particle picking has left a two-dimensional Gaussian distribution of residual translations q_x and q_y centred at the origin and with standard deviation ξ. Furthermore, we assume an even distribution of the $Q_{\rm rot}$ sampled in-plane rotations and a discretized distribution of estimated proportions α_{kp} of the data belonging to the pth projection of the kth underlying three-dimensional structure (with $\alpha_{kp} \geq 0$ and $\sum_{k=1}^{K} \sum_{p=1}^{P} \alpha_{kp} = 1$). Thereby, $P(k, p, q|\Theta)$ is calculated as follows:

$$P(k, p, q|\Theta) = \frac{\alpha_{kp}}{Q_{\rm rot} 2\pi\xi^2} \exp\left(\frac{q_x^2 + q_y^2}{-2\xi^2}\right). \tag{4}$$

In contrast to previous contributions where a Gaussian distribution was employed, we calculate $P(\mathbf{X}_i|k, p, q, \Theta)$ as a multivariate t-distribution, with a diagonal covariance matrix with all diagonal elements equal to σ^2:

$$\begin{aligned} P(\mathbf{X}_i|k, p, q, \Theta) &= t_J(\mathbf{X}_i; k, p, q, \Theta) \\ &= \frac{\Gamma\left(\frac{\nu+J}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\pi\nu\sigma^2)^{\frac{J}{2}}(1 + \delta_{ikpq}/\nu)^{\frac{(\nu+J)}{2}}}, \end{aligned} \tag{5}$$

with

$$\delta_{ikpq} = \frac{1}{\sigma^2} \|\mathbf{X}_i - R_{pq}\mathbf{V}_k\|^2 \tag{6}$$

and $\|\cdot\|$ denoting Euclidian distance.

Following McLachlan & Peel (2000), we notice that the multivariate t-distribution may be viewed as a weighted average Gaussian distribution with the weight given by the Gamma distribution:

$$t_J(\mathbf{X}_i; k, p, q, \Theta) = \int n_J(\mathbf{X}_i; k, p, q, u_i, \Theta) q(u_i)\, \mathrm{d}u_i. \tag{7}$$

Here, $q(u_i)$ is the p.d.f. of a Gamma distribution with equal scale and degrees of freedom, $G(\nu/2, \nu/2)$, and $n_J(\mathbf{X}_i; k, p, q, u_i, \Theta)$ is a multivariate Gaussian distribution centred at $R_{pq}\mathbf{V}_k$ and again with a diagonal covariance matrix, which has all diagonal elements equal to σ^2/u_i,

$$n_J(\mathbf{X}_i; k, p, q, u_i, \Theta) = \left(\frac{2\pi\sigma^2}{u_i}\right)^{-J/2} \exp\left(-\frac{u_i}{2}\delta_{ikpq}\right). \tag{8}$$

Therefore, it is convenient to introduce another set of 'missing' variables $u_1, \ldots, u_N$, which are defined such that

$$P(\mathbf{X}_i|k, p, q, u_i, \Theta) = n_J(\mathbf{X}_i; k, p, q, u_i, \Theta) \tag{9}$$

independently for $i = 1, \ldots, N$ and all u_i are independently distributed according to

$$P(u_i|k, \varphi, \Theta) = G\left(\frac{\nu}{2}, \frac{\nu}{2}\right). \tag{10}$$

Thus, the complete data set becomes

$$(\mathbf{X}_i, \Phi_i, \kappa_i, u_i), \quad i = 1, 2, \ldots, N \tag{11}$$

and the function to be optimized becomes

$$\begin{aligned} \sum_{i=1}^{N} \log \sum_{k=1}^{K} \sum_{p=1}^{P} \sum_{q=1}^{Q} P(\mathbf{X}_i|k, p, q, u_i, \Theta) \\ \times P(u_i|k, p, q, \Theta) P(k, p, q|\Theta). \end{aligned} \tag{12}$$

In analogy with (3) and with the previously introduced algorithm for Gaussian distributions (Scheres, Gao *et al.*, 2007), the reconstruction problem at hand is to find the parameter set Θ^* that maximizes (12). However, Θ now includes an additional parameter ν and the missing data vector has been augmented to not only include positions Φ_i and random indices κ_i but also variables u_i. In this way, atypical observations in the data (*i.e.* observations with relatively large residuals) may be accomodated by relatively wide Gaussian distributions (*i.e.* with small values of u_i) and the additional parameter ν is used to describe the assumed distribution of all u_i according to (10).

2.2. The optimization algorithm

This optimization problem may be solved by expectation maximization (Dempster *et al.*, 1977). This algorithm is used for finding maximum-likelihood estimates of parameters in probabilistic models that depend on unobserved or hidden variables. Expectation maximization is an iterative method that alternates between expectation (E) and maximization (M) steps. In the E-step one computes the expectation of the likelihood by including the hidden variables as if they were observed. In the M-step that follows, the maximum-likelihood estimate of the model parameters is computed by maximizing the expected likelihood found in the previous E-step. The parameters found in the M-step are then used to begin another E-step and the process is repeated. As stated above, the missing variables in this case are u_i, Φ_i and κ_i and the parameters to be estimated are contained in Θ.

In the E-step, again following McLachlan & Peel (2000), we calculate the expectation value of the log-likelihood function using the current estimates of the model parameters ($\Theta^{\rm old}$):

$$Q(\Theta;\Theta^{\rm old}) = \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{p=1}^{P}\sum_{q=1}^{Q}\tau_{ikpq}^{\rm old}\times[\log P(\mathbf{X}_i|k,p,q,u_i,\Theta) + \log P(u_i|k,p,q,\Theta)+\log P(k,p,q|\Theta)]. \quad (13)$$

Here, $\tau_{ikpq}^{\rm old}$ is the conditional probability distribution of k, p and q given $\mathbf{X}_i$,

$$\tau_{ikpq}^{\rm old} = \frac{P(k,p,q|\Theta^{\rm old})P(\mathbf{X}_i|k,p,q,\Theta^{\rm old})}{\sum_{k=1}^{K}\sum_{p=1}^{P}\sum_{q=1}^{Q}P(k,p,q|\Theta^{\rm old})P(\mathbf{X}_i|k,p,q,\Theta^{\rm old})}, \quad (14)$$

and for the conditional expectation of u_i, given $\mathbf{X}_i$, k, p and q, we obtain

$$u_{ikpq}^{\rm old} = \frac{\nu^{\rm old}+J}{\nu^{\rm old}+\delta_{ikpq}^{\rm old}}. \quad (15)$$

In the subsequent M-step of the algorithm, we maximize the lower bound (13) with respect to all model parameters in Θ. Since there exists no closed form for the update of $\nu^{\rm new}$, we will consider ν to be known (*i.e.* user-determined). The updates for ξ and the proportions $\alpha_{pk}^{\rm new}$ may be calculated independently from the updates of the other model parameters as follows:

$$\alpha_{pk}^{\rm new} = \frac{1}{N}\sum_{i=1}^{N}\sum_{q=1}^{Q}\tau_{ikpq}^{\rm old}, \quad (16)$$

$$\xi^{\rm new} = \left[\frac{1}{2N}\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{p=1}^{P}\sum_{q=1}^{Q}\tau_{ikpq}^{\rm old}(q_x^2+q_y^2)\right]^{1/2}. \quad (17)$$

For the updates of $\mathbf{V}$ and σ, we note that they are a weighted version of the corresponding updates in the case of Gaussian distributions, with standard deviations σ_i^2/u_i, ..., σ_N^2/u_N and with the weights being the additional missing variables u_1, ..., u_N. Therefore, as for the Gaussian case, updating $\mathbf{V}$ may be performed by separately solving K least-squares problems, for which we use a modified algebraic reconstruction algorithm (wlsART; see Scheres, Gao *et al.*, 2007). In this case, the least-squares problems are

$$\min\sum_{i=1}^{N}\sum_{p=1}^{P}\sum_{q=1}^{Q}\tau_{ikpq}^{\rm old}u_{ikpq}^{\rm old}\|\mathbf{X}_i - R_{pq}\mathbf{V}_k^{\rm new}\|^2 \quad (18)$$

and the updated σ is obtained as

$$\sigma^{\rm new} = \left(\frac{1}{NJ}\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{p=1}^{P}\sum_{q=1}^{Q}\tau_{ikpq}^{\rm old}u_{ikpq}^{\rm old}\|\mathbf{X}_i - R_{pq}\mathbf{V}_k^{\rm new}\|^2\right)^{1/2}. \quad (19)$$

2.3. Implementation

We implemented a total of four variants of the above-described algorithm in the open-source package *XMIPP* (Sorzano, Marabini *et al.*, 2004; Scheres *et al.*, 2008). The proposed algorithm for three-dimensional classification can be adapted with only minor changes to a two-dimensional classification algorithm. In this case, instead of optimizing (13) with respect to three-dimensional structures $\mathbf{V}_1$, ..., $\mathbf{V}_K$, one optimizes this function with respect to two-dimensional images $\mathbf{A}_1$, ..., $\mathbf{A}_K$. The algorithm remains basically the same, except for the fact that in this case R_{pq} represents an in-plane transformation (parametrized by a single rotation and two in-plane coordinates) and the least-squares problem in (18) is replaced by the following updated formula:

$$\mathbf{A}_k^{\rm new} = \frac{\sum_{i=1}^{N}\sum_{p=1}^{P}\sum_{q=1}^{Q}\tau_{ikpq}^{\rm old}u_{ikpq}^{\rm old}R_{pq}^{-1}\mathbf{X}_i}{\sum_{i=1}^{N}\sum_{p=1}^{P}\sum_{q=1}^{Q}\tau_{ikpq}^{\rm old}u_{ikpq}^{\rm old}}. \quad (20)$$

In addition, both the two-dimensional and the three-dimensional variants may also be expressed in reciprocal space. In this case, $\mathbf{X}_1$, ..., $\mathbf{X}_N$, $\mathbf{A}_1$, ..., $\mathbf{A}_K$ and $\mathbf{V}_1$, ..., $\mathbf{V}_K$ represent the Fourier transforms of the observed data and the two-dimensional or three-dimensional models, respectively, G_i is independent zero-mean additive noise in reciprocal space and R_{pq} represents the reciprocal-space equivalent of either a projection operation or an in-plane transformation in real space. In the former model one describes the noise by independent distributions on the real-space pixels, while in the latter the noise is modelled as being spatially stationary, which allows one to describe nonwhite or coloured noise. For a more extensive elaboration on these characteristics and their implementation, the reader is referred to Scheres, Nunez-Ramirez *et al.* (2007).

Finally, we mention that the summations over k, p and q are extremely computing-intensive operations. Therefore, we have implemented three deviations from the strict expectation-maximization algorithm that result in a considerable speed-up of the calculations without hampering the classification performance in practice. The first two deviations were also implemented as such in the algorithms using Gaussian distributions, whereas the third deviation is specific for the t-distribution case: (i) instead of integrating over the entire search space of k and q, we employ a reduced-space approach (Scheres, Valle & Carazo *et al.*, 2005), (ii) the update of σ is performed using $\mathbf{V}^{\rm old}$ instead of $\mathbf{V}^{\rm new}$ and (iii) following the proposal of McLachlan & Peel (2000), we replace the division by N in (19) by $\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{p=1}^{P}\sum_{q=1}^{Q}\tau_{ikpq}^{\rm old}u_{ikpq}^{\rm old}$.

3. Results

3.1. Robustness to outliers

We used a simplified two-dimensional test case to illustrate the potential of the t-distribution in providing robustness to outliers. The test data consisted of 1000 experimental cryo-EM projections of a 70S *Escherichia coli* ribosome particle in a single orientation. In 50 of the 1000 images, we positioned circles of constant density with radii varying uniformly between 10 and 15 pixels and with centres varying between −15 and 15 pixels from the image origin. The intensity of these circles was set to a constant value of 5, 10, 15 or 20 times the standard deviation of the original experimental images. We then performed two-dimensional real-space maximum-likelihood refinements with a single reference image for these data sets, comparing the performance of the Gaussian and

t-mixtures (Fig. 1). The resulting averages clearly showed the effect of the improved robustness to outliers provided by the *t*-mixture. For the data sets with the strongest outliers in particular, the averages obtained with the Gaussian mixture showed clear artefacts that were not visible in the averages obtained using the *t*-mixture model. Analysis of the converged estimates for the standard deviation in the noise indicated that the algorithm for the Gaussian case tries to accommodate the outliers by increasing the widths of the Gaussians. This is much less the case for the *t*-distribution case, where low values for $u_{ik\varphi}^{\rm old}$ downweight the contribution of the outliers in the calculation of the averages and the standard deviation of the noise. The stronger the outliers, the larger this downweighting effect and the larger the differences between the two algorithms.

3.2. Performance in two-dimensional classification

To explore the potential of the new algorithm in two-dimensional image classification, we performed maximum-likelihood multi-reference refinements on a cryo-EM data set of MCM top views (Gómez-Llorente *et al.*, 2005) and on a negative-stain data set of G40P top views (Nunez-Ramirez *et al.*, 2006). For each data set we performed four runs, using mixtures of Gaussians or of *t*-distributions with six degrees of freedom, and performing two-dimensional refinements in real or in reciprocal space. All four runs were started from identical seeds, which were obtained as average images over three random subsets of the data sets. Fig. 2 shows the resulting images of these runs, which show only minor differences between the two types of mixtures either in real or in reciprocal space. In all cases, the refined images look very similar to those obtained with a Gaussian mixture. Not only do the densities for the averaged particles in the centre of the images look very similar, the two mixture types even result in common characteristics in the noise background. The optimization path and the optimal orientation and classification parameters of the individual images upon convergence also showed only small differences (not shown).

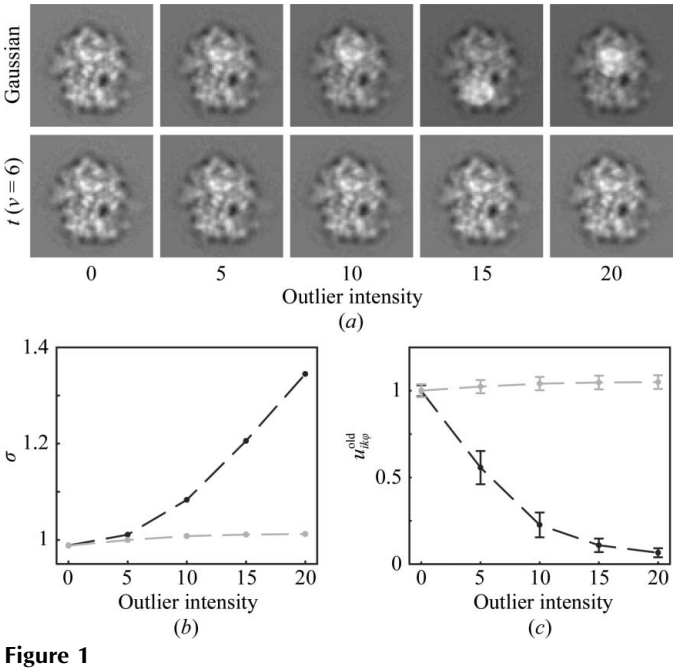

Figure 1
(*a*) Converged reference images for the runs with a mixture of Gaussians (top row) and *t*-distributions with six degrees of freedom (bottom row) for test sets containing outliers of increasing intensity. A value of zero for the outlier intensity is used to indicate the original test set without outliers. (*b*) Converged estimates for the standard deviation in the noise (σ) for the runs with a Gaussian (black) or a *t*-mixture (grey). (*b*) Average and standard-deviation values for the converged estimates for $u_{ik\varphi}^{\rm old}$ at maximal $\tau_{ik\varphi}^{\rm old}$ of the 50 outliers (black) and the remaining 950 images (grey) upon convergence for the runs with a *t*-mixture.

3.3. Performance in three-dimensional classification

For three-dimensional classification, we compared the performance of both types of mixtures using a data set of 20 000 ribosome particles. This data set was previously shown to be structurally heterogeneous as only part of the ribosomes are complexed with elongation factor G (EF-G; see Scheres, Nunez-Ramirez *et al.*, 2007). Refinements with four references typically converge to a single class corresponding to ribosomes in complex with EF-G and three classes of ribosomes without EF-G. We again performed four runs using real-space or reciprocal-space refinement and using a Gaussian or a *t*-distribution mixture with six degrees of freedom. The intensity of segmented EF-G density in the class corresponding to ribosomes in complex with EF-G may serve as an indicator of classification quality, since remaining heterogeneity will generally yield lower density levels for EF-G. Fig. 3 shows segmented EF-G densities for the four runs performed. Starting from identical seeds, the *t*-mixture model in real space gave somewhat higher EF-G densities than the Gaussian mixture and the corresponding classes overlapped by 87%. Refinement in reciprocal space yielded stronger EF-G densities than in real space for both types of mixtures. The differences between the Gaussian and the *t*-mixture were smaller in this case, as no obvious difference in the intensity of EF-G

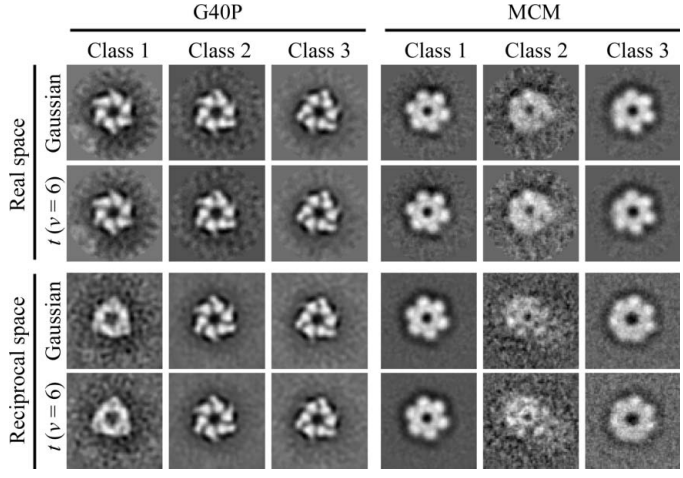

Figure 2
Class averages as obtained in two-dimensional classifications with three references for the G40P and MCM data sets using a Gaussian or *t*-mixture model in real or in reciprocal space.

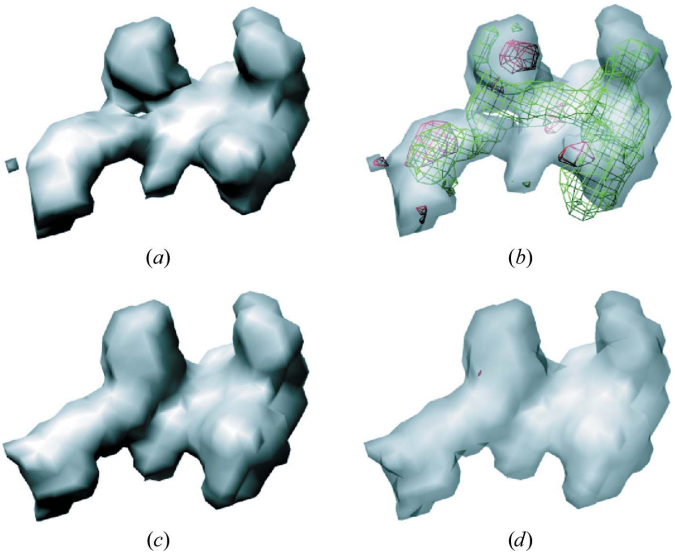

Figure 3
Segmented EF-G density from the maps obtained for the EF-G-containing class in three-dimensional maximum-likelihood refinements using a real-space (*a*, *b*) or reciprocal-space (*c*, *d*) target function and a Gaussian (*b*, *d*) or a *t*-mixture model with six degrees of freedom (*a*, *c*). Superimposed on the (transparent) densities obtained with the Gaussian mixture model are the positive (green) and negative (red) difference maps, *i.e.* the density obtained with the *t*-mixture minus the density obtained with the Gaussian mixture. All maps, including the difference maps, are rendered at the same isosurface value.

density was observed and the EF-G-containing classes overlapped by 94%.

4. Discussion

The selection of individual particles from electron micrographs, called particle picking, is typically a difficult task. For cryo-EM data on relatively small particles (200–500 kDa) in particular, automated procedures may have relatively high error rates and the collection of good data is often strongly dependent on the specialized skills of the electron microscopist (Zhu *et al.*, 2004). Therefore, it is relatively common for cryo-EM data sets to contain significant amounts of outliers. Atypical observations that were mistakenly assumed to be a particle of interest may deteriorate the quality of the three-dimensional reconstruction. In the best case scenario they only affect the resolution obtained. In the worst case scenario artefacts introduced by outliers may affect the interpretation of the structure itself. Conventionally, outliers have been dealt with by removing those particles with the lowest cross-correlation coefficients with the reference from the refinement process (Frank, 2006). Although effective in practice, such discrete decisions are hard to accommodate in the statistical framework of maximum-likelihood refinement.

The algorithm proposed in this contribution provides an alternative statistical solution to outlier removal. The problem of structurally heterogeneous projection data is modelled as a finite mixture of multivariate *t*-distributions with a given degree of freedom. In the resulting expectation-maximization algorithm, images with atypically large residuals contribute relatively little to the model estimates through lower values of $u_{ik\varphi}^{\rm old}$; see (15), (18) and (19). Note that the residuals used to calculate the weights $u_{ik\varphi}^{\rm old}$ are closely related to the cross-correlation coefficient, but instead of taking discrete decisions the statistical approach applies a continuous downweighting of outliers as their residuals increase. We illustrated this effect for a small experimental test set with artificially generated outliers. In the Gaussian model all particles contribute equally to the model estimates. Consequently, especially in the presence of strong outliers, the average images obtained showed outlier-related artefacts and the variance of the noise was overestimated. In contrast, the *t*-distribution model resulted in clean average images and reliable noise estimates through an effective downweighting of the outliers.

In practice, however, there are limits to the downweighting of outliers in the proposed algorithm. From (15) and the example in Fig. 1, we can see that even for a few degrees of freedom (*e.g.* six) significant downweighting is only achieved for images with squared residuals that exceed the standard deviation of the noise in the images several times. This may restrict the usefulness of the proposed algorithm in identifying aberrant particles. In practice, particles with such large residuals may be easily recognizable at earlier stages of image processing, whereas one would ideally want to downweight any particle that does not correspond to a projection of one of the K reference structures. Analyses of the two-dimensional classifications presented in §3.2 indeed did not reveal an obvious relation between low values of $u_{ik\varphi}^{\rm old}$ and what one would consider atypical images in terms of the underlying signal (results not shown).

The number of degrees of freedom is a free parameter of the proposed algorithm. Although in theory the optimal value of any free parameter should be tested, we performed all calculations presented in this work with a fixed value of six degrees of freedom. Because 3D-EM images typically contain many pixels, the right-hand side of both the numerator and the denominator in (15) will dominate the calculation of $u_{ik\varphi}^{\rm old}$ when using few degrees of freedom. Together with the observation made above that even for few degrees of freedom the effect of outlier downweighting may be relatively small, this suggests that in practice it may be sufficient to run this algorithm only with few degrees of freedom. This is confirmed by our calculations. When using three, nine or 30 degrees of freedom in the runs shown in Fig. 2, almost identical results were obtained (not shown) compared with using six degrees of freedom.

Improved image classification would be the ultimate aspiration of introducing a novel algorithm for maximum-likelihood refinement of 3D-EM data. Despite the fact that both the MCM and the G40P data set contained significant amounts of neighbouring particles and other artefacts that were not accounted for in the model, we did not observe any significant improvement in using the *t*-mixture model over the conventional Gaussian model in our two-dimensional classifications. One could attribute this to the observation that strong downweighting may only be achieved for outliers with very large residuals and that such strong artefacts were not present in these data. However, although the differences in the

$u_{ik\varphi}^{\rm old}$ weights may appear to be relatively small in practice, more subtle effects may still play an important role in the complicated convergence process. This may perhaps explain why three-dimensional classification of a structurally heterogeneous ribosome data set with a real-space *t*-mixture model may have given better classification results than the Gaussian mixture, as hinted at by a stronger signal for the complexed EF-G density.

We have presented too few tests to allow the drawing of general conclusions on the relative suitability of the *t*-mixture model and the conventional Gaussian model. A continuing application of the proposed algorithms on multiple test cases may provide further insights, but this falls beyond the scope of this contribution. Most probably, the optimal choice of algorithm will depend on the data set at hand. Therefore, we have made all algorithms described in this work accessible to the community by implementing them in our open-source package *XMIPP* (Sorzano, Marabini *et al.*, 2004; Scheres *et al.*, 2008). Apart from modifications to the maximum-likelihood classification approach as presented here, we also foresee the exploration of alternative algorithms, such as maximum *a posteriori* (MAP) estimation, which may offer signifant benefits in additional stabilization of the reconstruction problem through the incorporation of prior information.

We thank Dr Yacob Gómez-Llorente for providing the MCM data, Dr Rafael Núñez-Ramírez for providing the G40P data, Drs Haixiao Gao and Joachim Frank for providing the ribosome data and the latter for useful comments on an earlier version of this manuscript. We are grateful to the supercomputing centers of Barcelona (BSC-CNS) and Galica (CESGA) for providing computer resources. Funding was provided by the European Union (FP6-502828), the US National Institutes of Health (HL70472), the Spanish Ministry of Science (CSD2006-00023, BIO2007-67150-C03-1 and -3) and the Spanish Comunidad de Madrid (S-GEN-0166-2006).

References

Blanc, E., Roversi, P., Vonrhein, C., Flensburg, C., Lea, S. M. & Bricogne, G. (2004). *Acta Cryst.* D**60**, 2210–2221.
Bricogne, G. (1997). *Methods Enzymol.* **276**, 472–494.
Cuellar, J., Martin-Benito, J., Scheres, S. H. W., Sousa, R., Moro, F., López-Viñas, E., Gomez-Puertas, P., Muga, A., Carrascosa, J. & Valpuesta, J. (2008). *Nature Struct. Mol. Biol.* **15**, 858–864.
Dempster, A., Laird, N. & Rubin, D. (1977). *J. R. Stat. Soc. Ser. B*, **39**, 1–38.
Doerschuk, P. C. & Johnson, J. E. (2000). *IEEE Trans. Inf. Theory*, **46**, 1714–1729.
Frank, J. (2006). *Three-Dimensional Electron Microscopy of Macromolecular Assemblies.* Oxford University Press.
Gómez-Llorente, Y., Fletcher, R. J., Chen, X. S., Carazo, J. M. & San Martín, C. (2005). *J. Biol. Chem.* **280**, 40909–40915.
Julián, P., Konevega, A. L., Scheres, S. H. W., Lázaro, M., Gil, D., Wintermeyer, W., Rodnina, M. V. & Valle, M. (2008). *Proc. Natl Acad. Sci. USA*, **105**, 16924–16927.
La Fortelle, E. de & Bricogne, G. (1997). *Methods Enzymol.* **276**, 361–423.
Lange, K., Little, R. & Taylor, J. (1989). *JSTOR*, **84**, 881–896.
Lee, J., Doerschuk, P. C. & Johnson, J. E. (2007). *IEEE Trans. Image Process.* **16**, 2865–2878.
Leschziner, A. E. & Nogales, E. (2007). *Annu. Rev. Biophys. Biomol. Struct.* **36**, 43–62.
Marabini, R., Herman, G. T. & Carazo, J. M. (1998). *Ultramicroscopy*, **72**, 53–65.
McLachlan, G. & Peel, D. (2000). *Finite Mixture Models.* New York: John Wiley & Sons.
Nickell, S., Beck, F., Korinek, A., Mihalache, O., Baumeister, W. & Plitzko, J. M. (2007). *FEBS Lett.* **581**, 2751–2756.
Nunez-Ramirez, R., Robledo, Y., Mesa, P., Ayora, S., Alonso, J. C., Carazo, J. M. & Donate, L. E. (2006). *J. Mol. Biol.* **357**, 1063–1076.
Pascual-Montano, A., Donate, L. E., Valle, M., Bárcena, M., Pascual-Marqui, R. D. & Carazo, J. M. (2001). *J. Struct. Biol.* **133**, 233–245.
Penczek, P. A., Grassucci, R. A. & Frank, J. (1994). *Ultramicroscopy*, **53**, 251–270.
Provencher, S. W. & Vogel, R. H. (1988). *Ultramicroscopy*, **25**, 209–221.
Radermacher, M. (1994). *Ultramicroscopy*, **53**, 121–136.
Read, R. J. (2001). *Acta Cryst.* D**57**, 1373–1382.
Rehmann, H., Arias-Palomo, E., Hadders, M., Schwede, F., Llorca, O. & Bos, J. (2008). *Nature (London)*, **455**, 124–127.
Scheres, S. H. W., Gao, H., Valle, M., Herman, G. T., Eggermont, P. P. B., Frank, J. & Carazo, J. M. (2007). *Nature Methods*, **4**, 27–29.
Scheres, S. H. W., Nunez-Ramirez, R., Gomez-Llorente, Y., San Martin, C., Eggermont, P. P. B. & Carazo, J. M. (2007). *Structure*, **15**, 1167–1177.
Scheres, S. H. W., Nunez-Ramirez, R., Sorzano, C. O. S., Carazo, J. M. & Marabini, R. (2008). *Nature Protoc.* **3**, 977–990.
Scheres, S. H. W., Valle, M. & Carazo, J. M. (2005). *Bioinformatics*, **21**, *Suppl.* 2, ii243–ii244.
Scheres, S. H. W., Valle, M., Nuñez, R., Sorzano, C. O. S., Marabini, R., Herman, G. T. & Carazo, J. M. (2005). *J. Mol. Biol.* **348**, 139–149.
Sigworth, F. J. (1998). *J. Struct. Biol.* **122**, 328–339.
Sigworth, F. J. (2004). *J. Struct. Biol.* **145**, 111–122.
Sorzano, C. O. S., de la Fraga, L. G., Clackdoyle, R. & Carazo, J. M. (2004). *Ultramicroscopy*, **101**, 129–138.
Sorzano, C. O. S., Marabini, R., Velázquez-Muriel, J., Bilbao-Castro, J. R., Scheres, S. H. W., Carazo, J. M. & Pascual-Montano, A. (2004). *J. Struct. Biol.* **148**, 194–204.
Vogel, R. H. & Provencher, S. W. (1988). *Ultramicroscopy*, **25**, 223–239.
Wang, H., Zhang, Q. & Wei, S. (2004). *Pattern Recognit. Lett.* **25**, 701–710.
Yin, Z., Zheng, Y., Doerschuk, P. C., Natarajan, P. & Johnson, J. E. (2003). *J. Struct. Biol.* **144**, 24–50.
Zeng, X., Stahlberg, H. & Grigorieff, N. (2007). *J. Struct. Biol.* **160**, 362–374.
Zhu, Y. *et al.* (2004). *J. Struct. Biol.* **145**, 3–14.