[Short paper] EM-algorithm Enpowers Material Science: Application of Inverse Estimation for Small Angle Scattering

Akinori Asahara **Hidekazu Morita**
Hitachi Ltd.
Tokyo, 100-8280, Japan

Kanta Ono
High Energy Accelerator Research Organization
Tsukuba, 305-0801, Japan

Masao Yano **Tetsuya Shoji**
Toyota Motor Corporation
Toyota, 471-8572, Japan

Kotaro Saito
Paul Scherrer Institute
Villigen, 5232, Switzerland

Chiharu Mitsumata
National Institute for Materials Science
Tsukuba, 305-0047, Japan

Abstract

In this short paper, a machine-learning algorithm is applied to improve SAS (Small Angle Scattering) experimental analysis, which is commonly used in material science. In a SAS experiment, a particle beam incidenting to a material sample is scattered through the material sample. The distribution of the scattered beam indicates information about the grain-size distribution of the sample material; however, this distribution needs to be inversely estimated. Therefore, a stochastic model of the SAS experiment and EM (Expectation-Maximization)-algorithm to estimate the grain-size distribution in the material sample are proposed. While existing methods require much manual effort, the proposed EM-algorithm works automatically. Six simulation-generated datasets and two actual observed datasets were processed with the proposed method for examination. The result show that the proposed EM-based grain-size distribution estimation method is useful for automatically analyzing SAS data.

Introduction

Materials Informatics (MI) is an information technology intended for making material development faster that has been researched eagerly in recent years(National Institute of Standards and Technology 2019). MI will help material science researchers to discover new knowledge.

One such MI function is a data mining technique to find very small features of experimental data automatically. Traditionally, material science researchers carefully inspect experimental data to find small features because they might indicate new knowledge. The researchers however might take a long time to find such features or miss them. Therefore, automatic knowledge extraction from experimental data is attracting attention of the researchers.

This study focuses on small-angle scattering (SAS) experiments (Higgins and Benoît 1994) (Asahara et al. 2019), which are commonly conducted for observing microstructures of materials. There are various similar scattering experiments such as neutron-scattering, x-ray scattering, ion-beam scattering, etc. Their difference lies just in the particles to be scattered. The solution for the problem in SAS can be expected to be applied for these experiments also. Thus, the problem is crucial enough to need to be solved.

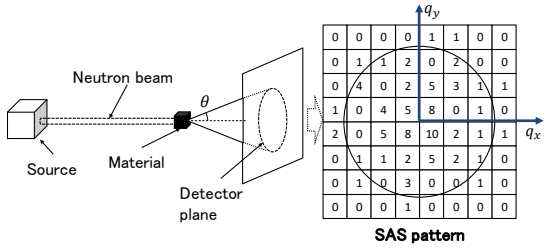

Figure 1: SAS Experiment

One of the SAS-experimental objectives is to estimate microscale-grain-size distributions in material samples. Neutrons detected on a plane during a SAS experiment make a pattern on the plane (called SAS pattern). Material science researchers with special knowledges observe SAS patterns carefully to find grain-size information about the microstructure of the sample material.

Accordingly, a method to automatically estimate grain-size distributions with SAS pattern data is presented in this paper. Several existing estimation methods are based on function optimization to fit the grain-size distribution to the SAS pattern, which requires much effort by maerial science researchers to adjust parameters. In contrast, our automatic estimation method is free from such effort because of probabilistic modeling of SAS experimental processes (that is, knowledges of the experimental settings). A maximum likelihood approach based on the stochastic modeling can be taken to estimate grain-size distribution without heuristic assumptions. In this paper, an expectation-maximization (EM) algorithm applicable to the estimation is shown and examined with simulation data and actual measurement data.

Problem settings

Small angle scattering

An experimental instrument setting of SAS is illustrated in Figure 1. In the experiment, a particle beam incident upon

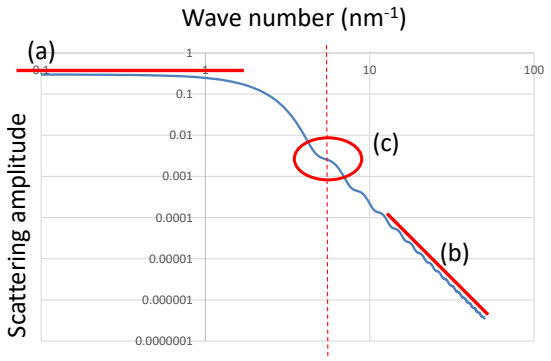

Figure 2: SAS pattern analysis with graphs

the sample interacts with the microstructures therein. The directions of the particles thus change due to the interactions. The angle θ between a straight beam and the changed direction of the scattered beam depends on the interaction. Finally detectors arranged on a plane detect the scattered beam. The counts of detection events form a pattern, called SAS pattern, on the plane. Thus, such microstructure causing the direction changes is called a "scattering body."

The particle behavior during the scattering experiment is modeled with a differential equation called the Schödinger equation. The solution of the Schödinger equation is a complex function called a wave function, of which the squared absolute value corresponds to the probability of detection. Because the distance L between the sample and the plane is large enough, the coordinate values on the plane $\mathbf{x} = (x, y)$ are approximately in proportion to $|\mathbf{x}| = L \sin\theta \simeq L\theta$. The probability density function (PDF) $P(\mathbf{x})$ of detection corresponds to the probability $P(\theta)$ that particle goes in the direction of θ, which is related to the microscopic structures called grains.

As the simplest setting, imagine a case in which the grains are balls. Intensity $I(r, q)$ of SAS pattern scattered by balls of radius r is in proportion to the following $\mathcal{I}(r, q)$

$$I(r,q) \propto \mathcal{I}(r,q) = \frac{1}{r^3}\left(\frac{\sin qr}{q^3} - \frac{r\cos qr}{q^2}\right)^2. \quad (1)$$

The q in the formula indicates a quantity called "wave number," which is the frequency of the wave function multiplied by 2π. The frequency of the wave function is three dimensional because it is derived with the Fourier transformation of the wave function in three dimensional space. The scattering angle θ depends on the frequency, so the size of $q = \mathbf{q}$ along the vertical vector to incident beam ("$\mathbf{q} = (q_x, q_y)$" in Fig 1) appears in the formula. Therefore a q indicates a location $\mathbf{x}$ on the detection plane, derived from distance between the incident beam center and the location. That is, we can obtain actual SAS intensity corresponding into $I(r, q)$ by converting $\mathbf{x}$ to q.

This formula is feasible in the case of a uniform grain size r. However actual grain sizes vary. The SAS pattern by multiple grain sizes is the weighted sum of $\mathcal{I}(r, q)$ over r and the weight is the grain-size distribution of the material, because the solutions of the Schödinger equation can be added together, accordingly scattering pattern $S(q)$ with a scattering body that is derived as

$$S(q) \propto \int f(r) I(r,q) \mathrm{d}r, \quad (2)$$

where the grain-size distribution is denoted as $f(r)$.

Expert-knowledge-based analysis

To estimate grain-size distribution, $S(q)$, which is the integration of $f(r)I(r, q)$, should be decomposed to the summation of $I(r, q)$; however this is difficult. Thus, material science researchers have tried to guess $f(r)$ with clues from small features latent in the plot of $\mathcal{I}(r, q)$ as shown in Fig. 2. The figure presents a log-log plot of a SAS pattern and it's domain is separated into three parts (a), (b) and (c). In (a), that is $q \to 0$, the power series of a trigonometric function with q

$$\mathcal{I}(r,q) \simeq \frac{1}{r^3}\left(\frac{qr}{q^3} - \frac{r}{q^2}(1 - \frac{1}{2}(qr)^2)\right)^2 = \frac{r}{4} \quad (3)$$

$S(q)$ is independent from q. Thus, it converges to a constant value. In (b), corresponding to $\mathcal{I}(r, q)$ under $q \to \infty$, is approximated as

$$\mathcal{I}(r,q) \simeq \frac{1}{r^3}\left(\frac{r\cos qr}{q^2}\right)^2. \quad (4)$$

Therefore, $S(q)$ is derived as

$$S(q) \simeq \frac{1}{q^4}\int r^2 f(r)\cos^2 qr \mathrm{d}r. \quad (5)$$

This behaves as the Fourier transform of $r^2 f(r)$ with decaying in the fourth power of q.

(c) is intermediate between (a) and (b). $\mathcal{I}(r, q)$ in the domain is the following.

$$\mathcal{I}(r,q) = \frac{1}{q^6}(\sin qr - qr\cos qr)^2. \quad (6)$$

$\mathcal{I}(r, q)$ is always non-negative and $\mathcal{I}(r, q) = 0$ when $\sin qr - qr\cos qr = 0$. Therefore $\mathcal{I}(r, q) = 0$ leads to $\sin qr/\cos qr = \tan qr = qr$. Figure 3 plots each side of this equation. The horizontal axis x of the graph indicates qr. The blue curve represents $y = \tan x$ and the orange line represents $y = x$. Their intersections, indicated by the circles in the figure, correspond to points satisfying $\tan qr = qr$, that is, $\mathcal{I}(r, q) = 0$. Therefore, the zero points appear periodically. Additionally local maximum points, which satisfy $\sin x = 0$, exist between the zero points. Thus $\mathcal{I}(r, q)$ oscillates and it's frequency depends on r. $S(q)$, which is the sum of the $\mathcal{I}(r, q)$, involves the oscillations of various phases, so the oscillations are gradually canceled by q becoming larger. Hence, only the oscillation at the small-q domain is readable.

The material science researchers accordingly look for the oscillation at the (c) domain because it gives implicit hints to understand $f(r)$. Therefore, $f(r)$ can be estimated only roughly. If $f(r)$ were estimated directly, the SAS experiment could give much more information of the sample. Consequently, a method to directly estimate $f(r)$ is highly needed. Thus, a machine-learning-based method is proposed in this paper.

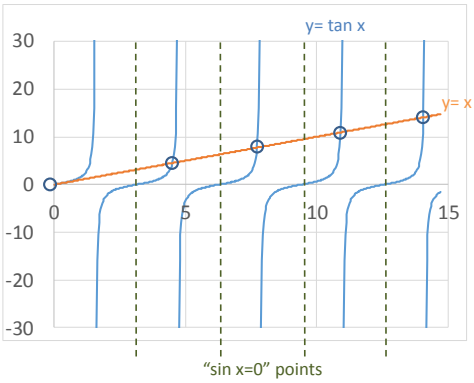

Figure 3: $\sin qr - qr\cos qr$ behavior

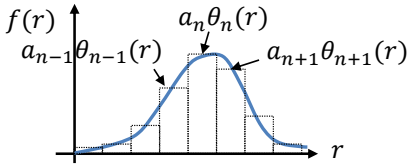

Figure 4: Indirect Fourier Transform (IFT)

Related works

One practicable method is parametric function fitting. Parameters of the function can be adjusted to fit to the obtained SAS pattern because their relationship is known (Joachim and Ingo 2018). However, for this approach, the form of $f(r)$ is required. The true $f(r)$ is generally unknown in actual situations. Material scientists therefore should assume many kinds of function forms to find the best estimation. Until the best estimation is achieved, many trials will be required, leading to a long calculation time.

To avoid such difficulty, a function having a more general formula should be used. One technique using such function is Indirect Fourier Transform (IFT) (Otto 1977). For IFT, summation of multiple stepwise functions $\theta_n(x)$ is used as the general function. The stepwise function $\theta_n(r)$ returns 1 when $r_n < r < r_{n+1}$, and 0 otherwise, where the domain of the function is separated into N small partitions $r_n < r < r_{n+1}$ $(1, \cdots n, \cdots N)$. Formula (2) is

$$S(q) \simeq \sum_n a_n \int \theta_n(r) I(r,q) \mathrm{d}r, \tag{7}$$

Under this assumption, the integral is decomposed into definite integrations in $r_n < r < r_{n+1}$. Because the definite integrals can be carried out analytically, $S(q)$ is described as a linear combination of a_n. After minimizing the difference between the linear combination of a_n and SAS pattern, the grain-size distribution $f(r)$ is obtained as the sum of $a_n\theta_n(r)$.

The resolution of the grain-size distribution is determined by θ_n in IFT as shown above. Therefore, the range of θ_n should be small to improve the resolution of grain size. Although many a_ns thus have to be determined for high resolution results, the SAS pattern must be highly accurate because

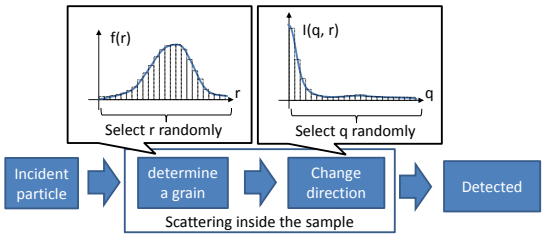

Figure 5: Probabilistic solution of scattering problems

the difference of a SAS pattern from $S(q)$ cannot be averaged enough in esthe timation of many a_ns. Accordingly the higher resolution setting makes estimation error larger. A technique to avoid this problem is to add regularization terms to suppress over fitting. However, the regularization terms is required to adjust manually. To automate regularization, complicated methods to determine the regularization terms have been proposed, but they are not common yet.

In this paper, an approach in which machine-learning algorithms are applied is taken against the problem. Specifically, the SAS-experimental process is modeled as a stochastic process with latent variables. After that, a likelihood function derived from the stochastic process is maximized to fit the SAS pattern. As the result, the grain-size distribution is obtained as the optimal model parameter of the stochastic process. No assumption is required for the method if a non-parametric model (that is, a very general stochastic model such as a Gaussian mixture) is applied for the SAS-experimental process. Generally an EM algorithm is applied to non-parametric models. Similarly a method using a non-parametric model and EM algorithm is proposed.

Such techniques are used in astrophysics (William 1972) (Leon 1974), bioinformatics (Lustig et al. 2008) (Lustig, Donoho, and Pauly 2007) and compressed sensing (Donoho 2006). However this kind of approach is not common in scattering experiments. Therefore, in this paper, algorithms suitable for SAS are proposed and examined using simulation and actual data.

Stochastic process of SAS

Approach

The process consists of dispersion and observation, which are modeled with two different probabilistic models. shown in Fig. 5.

At the first dispersion step, the incident beam interacts with grains. In Fig. 5, "determine a grain" represents the process. It can be interpreted as a stochastic process in which particles of the incident beam choose a scattering body in the sample material. The probability density function is consequently assumed in proportion to $f(r)$. That is, the dispersion step of N particles is modeled as a N-times iteration of random sampling from $f(r)$.

The second observation step, in which the incident beam changes its direction and arrives at a point on the detector plane, is also modeled as a random sampling process, shown as "change direction" in Fig. 5. The scattered parti-

cles choose a scattering angle randomly and are detected as a SAS pattern. This angle choice is stochastic due to the principle of quantum physics. Thus the probability distribution function is in proportion to $\mathcal{I}(r,q)$ defined in (3).

The entire process of SAS is modeled as the combination of these two stochastic processes. In the entire process, the size of the scattering body interacting with each particle is unobservable. When both latent variables and model parameters are unknown, that Bayes statistics works. The probability that q is chosen after determining r is described as a posterior $P(q|r)$ in Bayes statistics. Note that $P(q|r) \propto I(r,q)$ and the function to be estimated is $P(r|q)$ because only q is determined by the SAS pattern. These can be easily connected with Bayes theorem:

$$P(r|q) = \frac{P(q|r)P(r)}{P(q)}. \tag{8}$$

This formula includes two new parts ($P(r)$ and $P(q)$) though they do not cause problems. $P(r)$ is a prior about grain choosing. It can be set uniformly when no information about grain size is given. Moreover, $P(q)$ is a prior about the wavenumber. Being independent from grain-size, $P(q)$ will be canceled with a normalization constant of $P(r|q)$. Consequently, $P(r|q)$ equals $P(q|r)$, which is in proportion to $I(q,r)$, except the normalization constant.

This modeling is straightforward from a machine-learning-based viewpoint. However from the quantum-mechanics-based viewpoint, the incidenting particles are dealt with as a wave. Consequently, in the proposed approach, the model is simplified because of the aspect change from wave-like aspect to a particle-like one.

One-particle model

The formula about the scattering process of one particle should be precisely discussed as detailed above. The first process is to decide the grain causing scattering. The grain size r is continuous in domain $0 < r < R$. However, as mentioned above, it is separated into the L small partitions labeled by $0 \cdots L-1$. Assuming that the representative grain size in each partition is set as the center of the partition denoted as $r_0, \cdots r_L$ (that is $r_{n+1} = r_n + R/L$), we can write the grain size frequency as $f(r_0), \cdots f(r_{L-1})$. As the stochastic process, a particle randomly chooses a grain size for scattering with probability $P(r_i) \propto f(r_i)$. Accordingly,

$$P(r_l) = \frac{f(r_l)}{\sum_m f(r_m)}. \tag{9}$$

In the second process, the scattering angle is decided. Similarly the wavenumber domain $0 < q < Q$ is also separated into the K small partitions labeled by $0 \cdots K-1$ and the center of the partitions are denoted as q_k. The probability that the scattered particle is detected at the q_k detector is therefore described as $P(q_k|r_l)$, which is in proportion to $I(r_l, q_k)$. Although some particles will go outside of the detection plain, they are regarded as outside of the population distribution to be modeled. Consequently,

$$P(q_k|r_l) = \frac{I(r_l,q_k)}{\sum_m I(r_l,q_k)}. \tag{10}$$

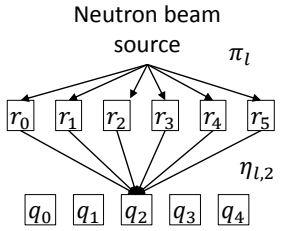

Figure 6: Marginalization of grain size

Algorithm 1: Estimation of grain size

Input: SAS pattern intensity $n_k \geq 0$, wavenumber $q_k \geq 0$ $(k = 0, 1, \cdots, K)$
resolution of grain size $r_l \geq 0$ where $(l = 0, 1, \cdots, L)$
Output: $\{\pi_l\}$
$N \Leftarrow \sum_k n_k$, $\{\eta_{l,k}\} \Leftarrow \{\frac{I(r_l,q_k)}{\sum_m I(r_l,q_k)}\}$, $\{\pi_l\} \Leftarrow 1/L$
repeat
$\{\pi_l\} \Leftarrow \sum_k \frac{n_k}{N} \frac{\pi_l \eta_{l,k}}{\sum_j \pi_j \eta_{j,k}}$
until convergence

For simplicity, $P(r_l) \equiv \pi_l$, $P(q_k|r_l) \equiv \eta_{l,k}$ hereafter. The probability that a particle is scattered at r_l and detected in the kth partition is derived as $\pi_i \eta_{l,k}$. To estimate the grain size distribution likelihood, we thus need $P(\{\pi_0, \cdots \pi_L\}|q_k)$.

The grain-size partition in which the particle is actually scattered is unobservable directly. Therefore, r_l should be marginalized as follows:

$$\begin{aligned} P(\pi_0, \cdots \pi_L|q_k) &= \frac{P(q_k|\pi_0, \cdots \pi_L)P(\pi_0, \cdots \pi_L)}{P(q_k)} \\ &\propto \sum_l P(q_k|r_l)P(r_l|\pi_0, \cdots \pi_L) \\ &= \sum_l \pi_i \eta_{l,k}, \end{aligned} \tag{11}$$

where priors $P(q_k)$ and $P(\pi_0, \cdots \pi_L)$ are regarded as constant parameters. Figure 6 illustrates this calculation. Even after a particle is detected at q_2, their possible paths are non-unique. Therefore the likelihood of the scatting process involves the sum of the all paths.

N-particles model

Although the likelihood of the 1-detection event is formulated as above, an actual SAS pattern includes many detection events. Because the SAS pattern is a set of counts of detection events, it is denoted as K integers: $\{n_0, \cdots n_K\}$. With the total number N of the events, $N = \sum_k n_k$. $\{\pi_0, \cdots, \pi_L\}$ maximizing the total likelihood $P(n_0, \cdots n_K|\pi_0, \cdots, \pi_L)$ is required, indicating the grain-size distribution.

For simplicity of the calculation, the following logarithmic likelihood is to be maximized by $\{\pi_k\}$.

$$\begin{aligned} &\ln P(\pi_0, \cdots, \pi_L|n_0, \cdots n_K) \\ &= \ln N! + \sum_k n_k \ln \sum_l \pi_l \eta_{l,k} - \sum_k \ln n_k! \end{aligned} \tag{12}$$

However, because the π_ks are probabilities of the random choice, they are restricted as $\sum \pi_k = 1$. Therefore, the maximization is carried out under the constraint with the Lagrange multiplier method.

$$\frac{\partial}{\partial \pi_l} \ln P(\pi_0, \cdots, \pi_L | n_0, \cdots n_K) = \frac{\partial}{\partial \pi_l} \sum_k n_k \ln \sum_l \pi_l \eta_{l,k} - \beta = 0, \quad (13)$$

where β is the Lagrange multiplier. This leads to the following L equations,

$$\frac{\partial}{\partial \pi_j} \sum_k n_k \ln \sum_l \pi_l \eta_{l,k} - \beta = \sum_k n_k \frac{\eta_{j,k}}{\sum_l \pi_l \eta_{l,k}} - \beta = 0. \quad (14)$$

After π is multiplied to both sides of the equations and the equations are summed,

$$\sum_k n_k \frac{\sum_j \pi_j \eta_{j,k}}{\sum_l \pi_l \eta_{l,k}} - \beta \sum_j \pi_j = 0$$
$$\beta = \sum_k n_k = N. \quad (15)$$

Therefore the equation

$$\sum_k n_k \frac{\eta_{j,k}}{\sum_l \pi_l \eta_{l,k}} = N \quad (16)$$

should be solved to obtain $\{\pi_l\}$.

To solve this problem, an iteration algorithm called an EM-algorithm (Bishop 2006) is generally applied (Zhang 1993)(Demoment 1989) (Nagata, Sugita, and Okada 2012). Because (10)(11) leads to

$$\frac{\pi_j \eta_{j,k}}{\sum_l \pi_l \eta_{l,k}} = \frac{P(q_k|r_l)P(r_l)}{P(q_k)} = P(r_l|q_k), \quad (17)$$

this part represents the probability that a particle detected at q_k is scattered at r_l. Therefore, the expectation value m_l of the number of such particles is $m_l = \sum_k n_k P(r_l|q_k)$ when n_k particles are detected at q_k. According to $P(r_l) = \pi_l$, additionally,

$$\sum_k n_k \frac{\eta_{j,k}}{\sum_l \pi_l \eta_{l,k}} = \frac{m_j}{\pi_j} = N. \quad (18)$$

The equation can be separated into the equation to lead π_ls and that to lead m_ls:

$$\pi_l = \frac{m_l}{N} \qquad m_l = \sum_k n_k \frac{\pi_l \eta_{l,k}}{\sum_j \pi_j \eta_{j,k}} \quad (19)$$

Consequently, E-step to obtain the expectation value m_l and M-step to obtain $\{\pi_l\}$ with the maximal likelihood are iteratively carried out to derive the solution of the equation (16). Algorithm 1 lists the procedures.

EXPERIMENTS

Experimental settings

Two different types of experiments were executed to evaluate whether the proposed algorithm automatically estimates grain-size distribution consistent with SAS pattern. In the first experiment (Experiment 1), simulation-generated data were processed because we can compare the results with ground truth. In the second experiment (Experiment 2), actual SAS pattern data with naive samples were processed to assess the actual feasibility of the proposed algorithm.

The two types of data were processed with the proposed algorithm, and IFT for comparison. For the proposed algorithm, 10,000 iterations of the EM algorithm were carried out instead of checking convergence. That is because the processing time is limited in an experiment but is unlimited until convergence. The processing time is expected to be limited when the iterations are limited.

The IFT executed in the experiments involves the L1 and L2 regularization. The weight parameters of the regularization terms are tuned for IFT to return reasonable estimation result. This tuning is carried out twice, that is, for Experiments 1 and 2, because the best setting depends on the total event number of the SAS pattern.

Experiment 1: simulation data

In Experiment 1, six types of grain-size distributions were defined. Each pattern is one Gamma distribution or the sum of two Gamma distributions having the most frequent point around 10nm. The grain-size distribution is discretized by 0.2 nm, and its domain is set from 0 to 20 nm (i.e., 100 values), corresponding to $f(r)$ in (2). The $S(q)$ was calculated by evaluating integration of (2). Because $S(q)$ indicates the probability of the detection, by multiplying the detection event number to $S(q)$, the most probable SAS patterns can be generated. The q of SAS pattern is also discrete and its domain is from $0.1\mathrm{nm}^{-1}$ to $5\mathrm{nm}^{-1}$. For the experiment, the detection event number was set as 10,000, and the SAS patterns of the grain-size distributions were generated and named Patterns 1-6.

Figures 7 and 8 show the results. In both figures, (a) plots the SAS pattern by log-log plot, (b) plots the grain-size distribution estimated by the proposed method, and (c) plots the grain-size distribution estimated by IFT for comparison. The blue lines in (b) and (c) plot the truth, i.e. the original grain-size distribution. In (b), all estimation results are highly similar to ground truth. In contrast, in (c), estimation results are generally inaccurate.

The grain-size distribution of Pattern 1 has a small peak at the foot of a large peak. The two peaks should be separately estimated. The ML results are so accurate that the small peak appears clearly, whereas the small peak in the IFT results is difficult to recognize.

The grain-size distribution of Pattern 2 also has a small peak, but it is located on the opposite side to that in Pattern 1. The IFT results do not accurately estimate the small peak, whereas the ML results do.

The grain-size distribution of Pattern 3 has only one peak. The IFT results of this pattern are similar to those of Pat-

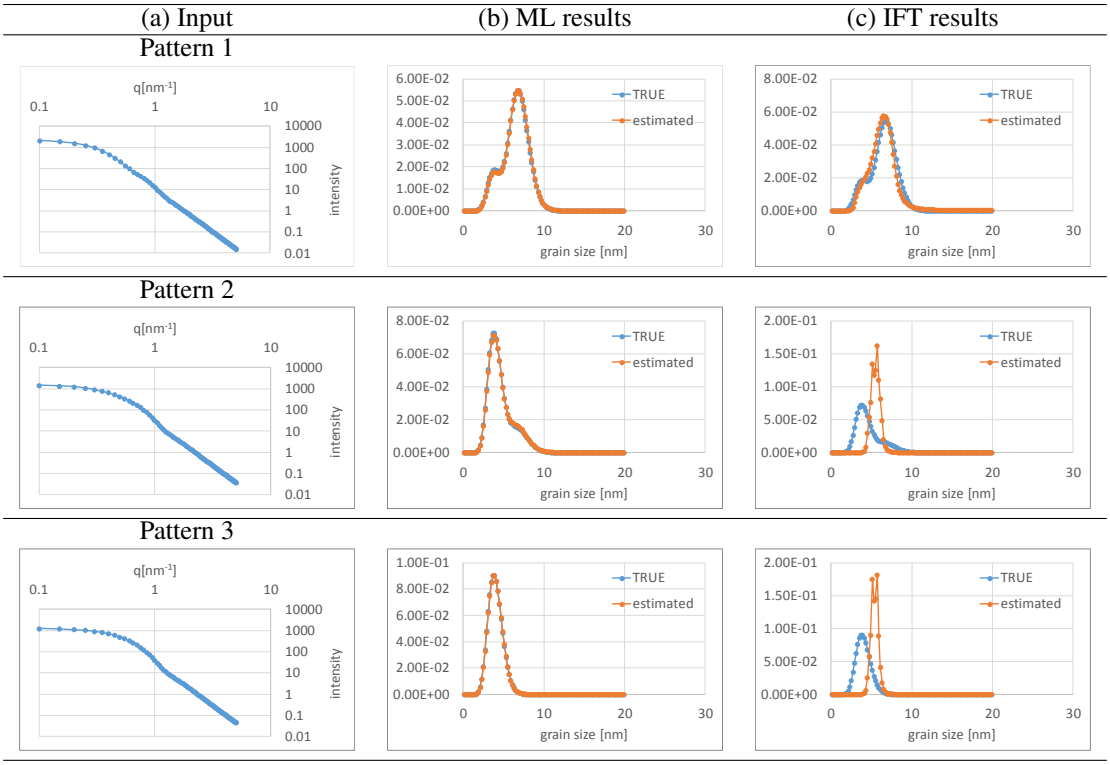

Figure 7: Results of Exp. 1 Pattern 1, 2, 3

tern 2. Both Patterns 2 and 3 have a large peak at a small grain size. The small grain size corresponds to a large wave number due to $I(q, r)$. The features are very small as shown in Fig. 7 (a) because the $S(q)$ in the high-q area decays q^{-4}. Function-fitting-based techniques such as IFT cannot handle such small components, whereas stochastic techniques such as the proposed method take into account small probabilities.

One large peak in the intermediate grain-size is shown in Pattern 4. Pattern 4 is so simple that the estimation is easy. Indeed, both the ML and IFT results are very accurate. However, the ML results are more accurate the IFT results.

Two comparable peaks appear closely in Pattern 5. Because the IFT results do not detect these two peaks, one peak instead appears between them. In contrast, ML results detect both peaks accurately.

Three peaks are shown in Pattern 6. Similar to Pattern 5, the IFT results did not extract the three peaks, whereas the ML results did.

The SAS patterns of (a) input are quite similar for humans. Therefore, material scientists have to make an effort to obtain their difference, which reflects radical changes in the grain-size distribution. According to the results, the proposed method is helpful and reliable. This shows that the SAS experiment can become more useful for observing microstructures of materials.

Figure 9 plots processing time of the pattern estimation. For this experiment, a computer loading Intel(R) Core(TM) i3-4150 CPU 3.50GHz and 11 GB RAM and Cent OS. The implementation is based on Python 3.6.5 and numpy library (Oliphant 2006) is used to improve efficiency of the process.

The proposed method takes around 1.2 seconds, which is much shorter than the experimental time of SAS (for neutron scattering, around 20 minutes). In comparison, IFT takes around 6.0 seconds, 5 times as long as the proposed method. IFT is not much slower; however, this difference can became important if material science researchers have to conduct many iterations during trial-and-error experiments. This shows the proposed method is quite useful for SAS data analysis.

According to the results, the proposed method enables the grain-size distribution to be estimated accurately. IFT makes large errors when the grain size is small, whereas the proposed method works well for such cases. In actual situations, we cannot know whether the grain size of a sample is low (i.e., IFT applicable) or not. Therefore, IFT requires much effort by material scientist but the proposed method does not. This shows that the proposed method is suitable for automatically processing SAS patterns.

Experiment 2: actual measurements

In Experiment 2, SAS patterns of neutrons with a polystyrene ball (radius 18 nm) sample and a silica ball (radius 25 nm) sample were examined. Figure 10 shows the results ((a), (b) and (c) are the same as in Experiment 1). The SAS pattern are more noisy than those of Experiment 1.

The most frequent radius of (b) and (c) is around the sample true radius. This shows that both the proposed method

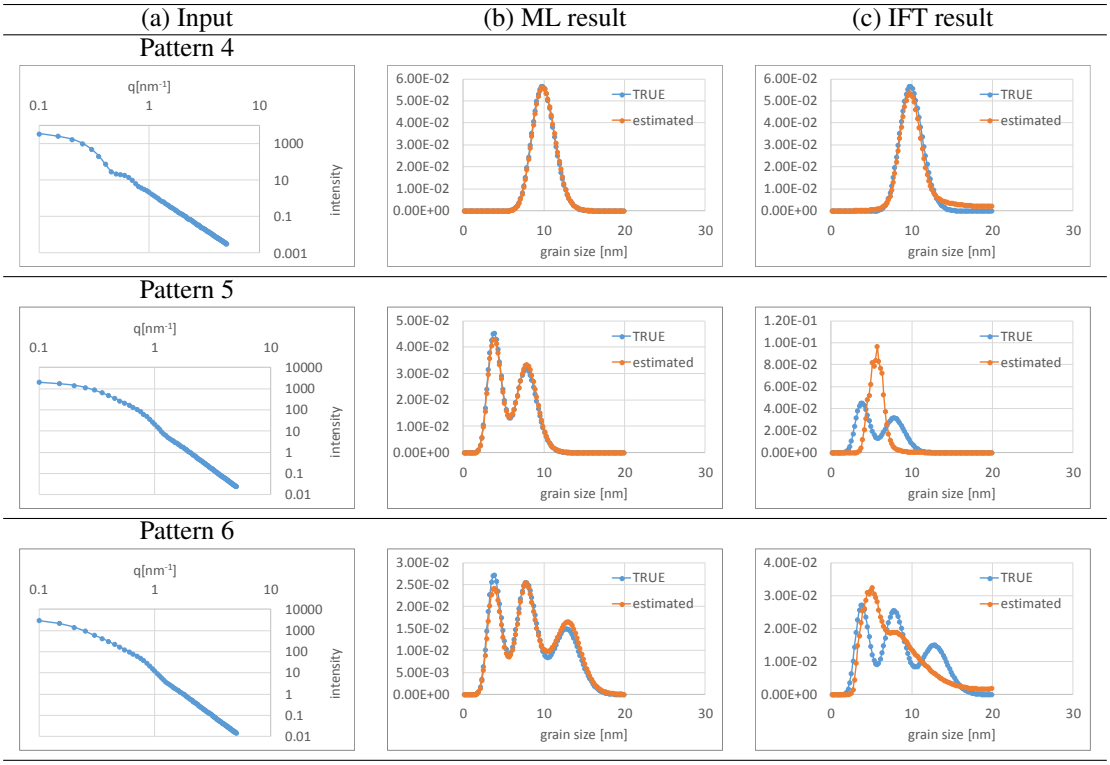

Figure 8: Results of Exp. 1 Patterns 4, 5, 6

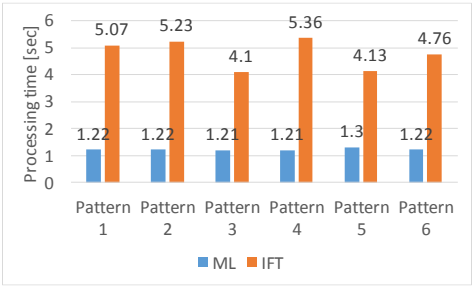

Figure 9: Comparison of processing time

and IFT can be used. The difference between the ML results and IFT results is that small peaks appear at the integer-multiplied true radius. This is considered to be because clusters of the multiple balls are detected.

The results show the proposed method is feasible for actual SAS pattern analysis. Moreover small material-inside behaviors might be observable. Thus this implies that the proposed method will extract information leading to new knowledge.

Conclusion and Future Works

An expectation-maximization (EM)-based grain-size distribution estimation method was proposed for the automatically analyzing small angle scattering (SAS) patterns. Experimental results showed that the proposed method can accurately estimate the original grain-size distribution from SAS patterns. Moreover, the proposed method does not require parameter tuning to obtain good results, whereas the existing method (Indirect Fourier Transform) does.

The stochastic model that is the base of the proposed method does not assume priors. However, with priors, the estimation might be made more accurate and detection events required to estimate the grain-size might be made fewer. In addition, non-ball scattering bodies should be taken into account. Such extensions are possible future works.

References

Asahara, A.; Morita, H.; Mitsumata, C.; Ono, K.; Yano, M.; and Shoji, T. 2019. Early-stopping of scattering pattern observation with bayesian modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 9410–9415.

Bishop, C. M. 2006. *Pattern Recognition and Machine Learning*. New York: Springer.

Demoment, G. 1989. Image reconstruction and restoration: overview of common estimation structures and problems. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 37(12):2024–2036.

Donoho, D. L. 2006. Compressed sensing. *IEEE Transactions on information theory* 52(4):1289–1306.

Higgins, J. S., and Benoît, H. 1994. *Polymers and neutron scattering*. Clarendon press Oxford.

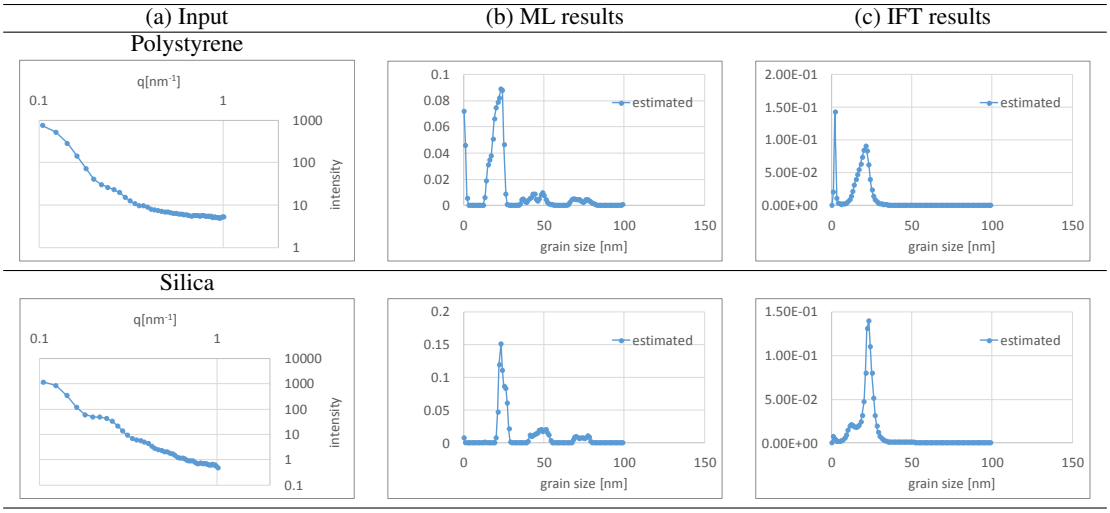

Figure 10: Results of Exp. 2

Joachim, K., and Ingo, B. 2018. SASFit. https://www.psi.ch/en/sinq/sansi/sasfit.

Leon, B. L. 1974. An iterative technique for the rectification of observed distributions. *The astronomical journal* 79(6):745–754.

Lustig, M.; Donoho, D. L.; Santos, J. M.; and Pauly, J. M. 2008. Compressed sensing mri. *IEEE Signal Processing Magazine* 25(2):72–82.

Lustig, M.; Donoho, D.; and Pauly, J. M. 2007. Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine* 58(6):1182–1195.

Nagata, K.; Sugita, S.; and Okada, M. 2012. Bayesian spectral deconvolution with the exchange monte carlo method. *Neural Networks* 28:82 – 89.

National Institute of Standards and Technology. 2019. mgi. https://www.nist.gov/mgi(viewed at Oct. 2019).

Oliphant, T. E. 2006. *A guide to NumPy*, volume 1. Trelgol Publishing USA.

Otto, G. 1977. A new method for the evaluation of small-angle scattering data. *Journal of Applied Crystallography* (10):415–421.

William, Hadley, R. 1972. Bayesian-based iterative method of image restoration. *Journal of the Optical Society of America* 62(1):55–59.

Zhang, J. 1993. The mean field theory in em procedures for blind markov random field image restoration. *IEEE Transactions on Image Processing* 2(1):27–40.