Journal of Innovative Optical Health Sciences
Vol. 2, No. 3 (2009) 303–313

World Scientific
www.worldscientific.com

EARLY CATARACT DETECTION BY DYNAMIC LIGHT SCATTERING WITH SPARSE BAYESIAN LEARNING

SU-LONG NYEO*,‡ and RAFAT R. ANSARI†,§
*Department of Physics, National Cheng Kung University
Tainan, Taiwan 701, ROC
†Bioscience and Technology Branch
NASA Glenn Research Center at Lewis Field
21000 Brookpark Road, Cleveland, OH 44135, USA
‡t14269@mail.ncku.edu.tw
§Rafat.R.Ansari@nasa.gov

Dynamic light scattering (DLS) is a promising technique for early cataract detection and for studying cataractogenesis. A novel probabilistic analysis tool, the sparse Bayesian learning (SBL) algorithm, is described for reconstructing the most-probable size distribution of α-crystallin and their aggregates in an ocular lens from the DLS data. The performance of the algorithm is evaluated by analyzing simulated correlation data from known distributions and DLS data from the ocular lenses of a fetal calf, a Rhesus monkey, and a man, so as to establish the required efficiency of the SBL algorithm for clinical studies.

Keywords: Cataract; dynamic light scattering; diagnostic algorithm; sparse Bayesian learning (SBL).

1. Introduction

Cataract is the leading cause of impaired vision worldwide, especially in developing countries. Cataracts usually develop slowly but eventually interfere with vision. Therefore, early diagnosis of cataracts and study of cataractogenesis can provide the appropriate preventive measures with much attention on the development of drugs and investigation of their effects on the medical treatment of cataract.

With the invention of a compact fiber optic dynamic light scattering (DLS) probe,[1,2] early cataract detection is possible. DLS has been proven to be an efficient non-invasive technique for measuring the sizes of colloidal particles and found applications in the study of dynamical properties of biological materials *in vivo*.[3]

DLS has been shown to be well suited for the diagnosis of early cataracts,[1,2,4] which are caused by aggregation of crystallin proteins in the lens that strongly scatters light. Lens opacity increases with the amount of large protein aggregates and eventually affects vision. In clinical studies, DLS has been proven to be far more superior (several orders of magnitude) in sensitivity than the gold standard imaging techniques used currently in the evaluation of cataractogenesis.

In a DLS experiment, light scattered off the diffusing colloidal particles is detected and processed at discrete delay times τ_i, $i = 1, 2, \ldots, M$ to yield an intensity time-autocorrelation function (TCF), $\langle I(t)I(t+\tau_i)\rangle \equiv G^{(2)}(\tau_i)$, a second-order TCF for the electric field, where $I(t) = E^*(t)E(t)$ and $E(t)$ denote, respectively, the intensity and electric fields of the scattered light at time t, and the braces indicate averaging over t. The TCF provides the data for an inverse problem for inferring the translational diffusion coefficient or size distribution of the particle system. For Gaussian signals, the first-order TCF for the electric field, $\langle E(t)E^*(t+\tau_i)\rangle \equiv g^{(1)}(\tau_i)$

is related to the intensity TCF by Siegert's relation:

$$G^{(2)}(\tau_i) = A(1 + \beta|g^{(1)}(\tau_i)|^2), \quad (1)$$

where A is the measured baseline and β is an instrument coherence factor.

The electric field TCF is related to the distribution $P(\Gamma)$ of the particle system by

$$|g^{(1)}(\tau_i)| = \int_0^\infty P(\Gamma)e^{-\Gamma\tau_i}d\Gamma, \quad (2)$$

where Γ is the decay constant, which is related to the diameter d of the particles by

$$\Gamma = \frac{k_B T q^2}{3\pi\eta d}. \quad (3)$$

Here,

$$q = \frac{4\pi n_0}{\lambda}\sin\frac{\theta}{2} \quad (4)$$

is the momentum transfer, k_B is Boltzmann's constant, T is the absolute temperature, η and n_0 are, respectively, the viscosity coefficient and refractive index of the fluid, λ is the wavelength of the laser light and θ the scattering angle at which the signal is collected.

Equation (2) is very ill-posed, since it is extremely sensitive to experimental uncertainties, such as noise in the data.[5,6] To retrieve the information about a particle system, various analysis tools have been developed. Notably, the cumulant method,[7] exponential sampling method,[8,9] CONTIN[10] and maximum entropy method (MEM)[11–15] are some common analysis tools that enjoy reasonable reliability and efficiency in many applications.

In this paper, we describe a fully probabilistic analysis tool, the SBL algorithm,[16] with easy implementation for determining particle size distribution (PSD) from DLS data and, in particular, for the diagnosis of early cataract in humans. In view of the fact that experimental data resulted from a measurement process which is necessarily statistical in nature, the use of an analysis technique based on Bayesian statistical learning theory is suitable. In passing, we note that Bryan's maximum entropy algorithm[14,15] is based on Bayesian inference and is also probabilistic in nature. It uses the maximum-entropy function as a regularization term, whereas the SBL algorithm uses a quadratic function and leads to a set of simple update equations for easy implementation.

The SBL algorithm[16] is a flexible version of the relevance vector machine (RVM), which is known to enjoy more advantages than the state-of-the-art technique of the support vector machine (SVM).[17] Thus, it is useful to understand how the SBL algorithm can be applied to analyze DLS data to extract reliable results, specifically, for the application of DLS data in clinical studies, where Bayesian interpretation of the results with strong statistical evidence is required. Moreover, reliable and robust diagnostic algorithms are essential, since the algorithm is less sensitive to initial input parameters, such as the domain of the distribution, which is related to resolution as dictated by the data quality. For a detailed discussion on the experimental considerations for obtaining quality data and also on various data analysis methods, please see Refs. 18 and 19.

In Sec. 2, we briefly describe the SBL algorithm and apply it to DLS. In Sec. 3, simulated data generated from unimodal and bimodal distributions will be analyzed using the SBL algorithm, and DLS data from the ocular lenses of a fetal calf, a monkey, and a man will be analyzed to demonstrate the efficiency of the SBL algorithm for early cataract detection and monitoring.

2. Application of the SBL Algorithm to DLS

The SBL algorithm was introduced by Tipping[16] for solving regression and classification problems and has found many promising applications.[20–23]

To use the SBL algorithm for DLS, Eq. (2) is discretized as

$$y_i = \sum_{j=1}^{N}\Phi_{ij}\omega_j + \epsilon_i,$$

$$\Phi_{ij} \equiv e^{-\Gamma_j\tau_i}, \quad i = 1, \ldots, M, \quad (5)$$

where ϵ_i are the errors in the data y_i. Solving Eq. (5) for ω_j amounts to defining an "inverse" for the kernel Φ_{ij} in some representation. Here, we are dealing with an inverse problem for an oversampled DLS dataset where the amount of independent information is much less than the number of measured data points ($M \gg N$).

SBL is a probabilistic approach to finding a sparse solution to the supervised learning problem [Eq. (5)]. Here, $\boldsymbol{\epsilon}$ is the noise from a zero-mean Gaussian process. For many practical problems, the challenge is to determine the most sparse representation of the reconstruction coefficients $\boldsymbol{\omega} = [\omega_1, \omega_2, \ldots, \omega_N]^T$ with the fewest nonvanishing ω_j,

i.e., only few "relevance vectors" or basis functions are significant.

For our DLS data analysis, we consider using the SBL algorithm to find smooth PSD. Essentially, this requires drawing an inference. At the heart of statistical inference lies Bayes' theorem, which relates the posterior and prior probabilities of two events. Thus, solution to Eq. (2) is more appropriately determined with a probability-based algorithm, and SBL provides a fully probabilistic algorithm.

In the context of Bayesian statistics, we calculate the posterior or conditional probability $p(\boldsymbol{\omega}|\mathbf{y})$ of $\boldsymbol{\omega}$ given the DLS data $\mathbf{y}$. The SBL approach assumes that the noise in the data $\mathbf{y}$ is normally distributed with zero mean and variance σ^2, i.e.,

$$p(\epsilon_i|\sigma^2) \sim \mathcal{N}(0, \sigma^2), \quad i = 1, 2, \ldots, M. \tag{6}$$

Then the likelihood of the data reads

$$p(\mathbf{y}|\boldsymbol{\omega}, \sigma^2) = (2\pi)^{-M/2}|\mathbf{B}|^{1/2} \times \exp\left(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\omega})^T\mathbf{B}(\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\omega})\right), \tag{7}$$

where $\mathbf{B} = \sigma^{-2}\mathbf{I}$ is a global parameter in SBL and its adjustments for cases where errors depend on the data values may be considered. Now, maximizing the likelihood (7) amounts to minimizing its exponent or chi-square. By requiring the weights to be positive, we are led to solving the inverse problem conveniently with the nonnegative least squares (NNLS) algorithm.[24] In general, maximum-likelihood estimation from Eq. (7) will lead to severe overfitting, and so for smooth functions, an automatic relevance determination (ARD) Gaussian prior[25] for the weights is introduced:

$$p(\boldsymbol{\omega}|\boldsymbol{\alpha}) = (2\pi)^{-N/2}|\boldsymbol{\alpha}|^{1/2}\exp\left(-\frac{1}{2}\boldsymbol{\omega}^T\boldsymbol{\alpha}\boldsymbol{\omega}\right), \tag{8}$$

where $\boldsymbol{\alpha} = \mathrm{diag}(\alpha_1, \alpha_2, \ldots, \alpha_N)$ is a regularization matrix with diagonal elements and plays essentially the same role as Marquardt's parameter in other analysis methods, such as MEM and CONTIN, where the Marquardt's parameter is a scalar and the data have different errors. The introduction of a hyperparameter α_j for each weight ω_j leads to a sparse model. However, it is to be noted that since the degree of smoothing is to a large extent controlled by the choice of kernel, a possibility of severe overfitting or oversmoothing can still result.

We note that $\boldsymbol{\alpha}$ and σ^2 are nuisance parameters, having Gamma distributions as suitable priors[16]:

$$p(\boldsymbol{\alpha}) = \prod_{i=0}^{N}\mathrm{Gamma}(\alpha_i|a, b), \tag{9}$$

$$p(\sigma^2) = \mathrm{Gamma}(\sigma^2|c, d), \tag{10}$$

with $\mathrm{Gamma}(\alpha|a, b) = \Gamma(a)^{-1}b^a\alpha^{a-1}e^{-b\alpha}$, and $\Gamma(a)$ being Euler's Gamma function. For non-informative and uniform (over a logarithmic scale) priors, we might fix $a = b = c = d = 0$. For given data $\mathbf{y}$ and the hyperparameters, the posterior distribution of the weights can be evaluated analytically in view of the fact that the likelihood of the data and the priors for the weights are of the Gaussian form. Indeed, by using the probability distributions, $p(\mathbf{y}|\boldsymbol{\omega}, \sigma^2)$ and $p(\boldsymbol{\omega}|\boldsymbol{\alpha})$, and Bayes' theorem

$$p(\boldsymbol{\omega}|\mathbf{y}, \boldsymbol{\alpha}, \sigma^2) = \frac{p(\mathbf{y}|\boldsymbol{\omega}, \sigma^2)p(\boldsymbol{\omega}|\boldsymbol{\alpha})}{p(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2)}, \tag{11}$$

we obtain the marginal likelihood, or evidence, for the hyperparameters:

$$p(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2) = \frac{1}{(2\pi)^{M/2}}\frac{1}{|\sigma^2\mathbf{I} + \boldsymbol{\Phi}\boldsymbol{\alpha}^{-1}\boldsymbol{\Phi}^T|^{1/2}} \times \exp\left\{-\frac{1}{2}\mathbf{y}^T(\sigma^2\mathbf{I} + \boldsymbol{\Phi}\boldsymbol{\alpha}^{-1}\boldsymbol{\Phi}^T)^{-1}\mathbf{y}\right\}, \tag{12}$$

and the weight posterior:

$$p(\boldsymbol{\omega}|\mathbf{y}, \boldsymbol{\alpha}, \sigma^2) = \frac{1}{(2\pi)^{N/2}|\boldsymbol{\Sigma}|^{1/2}} \times \exp\left\{-\frac{1}{2}(\boldsymbol{\omega} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\omega} - \boldsymbol{\mu})\right\}, \tag{13}$$

where

$$\boldsymbol{\Sigma} = \left(\boldsymbol{\Phi}^T\mathbf{B}\boldsymbol{\Phi} + \boldsymbol{\alpha}\right)^{-1} \tag{14}$$

denotes the variance matrix and

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{B}\mathbf{y}, \tag{15}$$

the estimate of $\boldsymbol{\omega}$. From Eqs. (5) and (15), we have the estimate of data as follows:

$$\hat{\mathbf{y}} = \left[\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{B}\right]\mathbf{y} \equiv \mathbf{S}\mathbf{y}, \tag{16}$$

where $\mathbf{S}$ is known as the smoothing or hat matrix.

Returning to the marginal likelihood (12) and the weight posterior (13), we note that they depend on the hyperparameters. In the context of Bayesian inference, we define the hyperpriors for $\boldsymbol{\alpha}$ and σ^2

and integrate out these hyperparameters. To do so, we maximize Eq. (12) or equivalently $p(\boldsymbol{\alpha}, \sigma^2|\mathbf{y})$ with respect to $\boldsymbol{\alpha}$ and σ^2 with their assumed non-informative priors.[16] Thus, to find the values of $\boldsymbol{\alpha}$ and σ^2 that maximize Eq. (12), we follow an iterative re-estimation approach to finding the hyperparameters that maximize the hyperparameter posterior $p(\boldsymbol{\alpha}, \sigma|\mathbf{y})$, such that appropriate update equations can be derived. Given Bayes' rule

$$p(\boldsymbol{\alpha}, \sigma^2|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2)p(\boldsymbol{\alpha}, \sigma^2)}{p(\mathbf{y})}$$

$$\propto p(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2)p(\boldsymbol{\alpha})p(\sigma^2) \qquad (17)$$

with priors given by (9) and (10), the maximization of (17) is obtained by differentiating $\log(p(\boldsymbol{\alpha}, \sigma^2|\mathbf{y}))$ with respect to $\log(\alpha_i)$ and $\log(\sigma^2)$. Leaving out the details of the derivations, which can be found in Ref. 22, we list the update equations for the hyperparameters:

$$(\alpha_i)^{\text{new}} = \frac{\gamma_i}{\mu_i^2}, \quad i = 1, 2, \ldots, N, \qquad (18)$$

$$(\sigma^2)^{\text{new}} = \frac{\|\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\mu}\|^2}{M - \sum_{i=1}^{N} \gamma_i}, \qquad (19)$$

where

$$\gamma_i = 1 - \alpha_i \Sigma_{ii}, \quad i = 1, 2, \ldots, N. \qquad (20)$$

We now summarize the iterative steps for obtaining the most-probable values $\boldsymbol{\alpha}_{\text{MP}}$ and σ^2_{MP} and hence the estimates of the weights and data:

(1) Assign initial values to $\boldsymbol{\alpha}$ and σ^2. e.g., $\boldsymbol{\alpha} = 10^{10}$ and $\sigma^2 = 10^{-6}$.
(2) Compute Σ in Eq. (14).
(3) Compute $\boldsymbol{\mu}$ in Eq. (15).
(4) Compute γ_i in Eq. (20).
(5) Compute the hyperparameters in Eqs. (18) and (19).
(6) Reset μ_i to a positive value at every iteration if μ_i becomes negative by multiplying it by a small factor, e.g., -0.01, so as to ensure the nonnegativity of the solution $\{\mu_i\}$.
(7) Go to steps 2–6 until the hyperparameters $\boldsymbol{\alpha}$ and σ^2 converge.

Due to its iterative property, the SBL algorithm may be applied to analyze DLS data of ocular lenses in two ways. First, we can study the solutions generated by the algorithm at various iterations. The solutions at the initial iterations are smooth distributions, while those at larger iterations are sparser. By using the SBL algorithm, we first consider simulated data generated from smooth distributions and reconstruct the most-probable or optimal solutions. We define the most-probable or optimal, smooth solution at some iteration by some goodness-of-fit criterion.

To study DLS data of an ocular lens, we may consider using the SBL algorithm to reconstruct sparse solution, which gives values only at the "relevant" particle sizes, and the distribution provides a categorized list of the α-crystallin of few particle sizes. However, the PSD is not smooth and does not properly reflect the true distribution.

To analyze DLS data by the SBL algorithm, we have to assign the boundaries and the grid points of the PSD, which depends on the particle system and the data quality. In most analysis methods, one uses N logarithmically-spaced grid points for PSD. This is particularly efficient to describe highly polydisperse systems. We define the first and last values of the range of the decay constants (or particle sizes), respectively, by $\Gamma_1 = \Gamma_{\min}$ and $\Gamma_N = \Gamma_{\max}$. The number of points N can be chosen to be large, but larger N requires more computation time to invert the $N \times N$ regularization matrix for the kernel Φ_{ij} at each iteration.

For a fixed resolution, an increase in the ratio Γ_N/Γ_1 should be accompanied by an increase in the number of grid points N. Accordingly, to obtain a distribution, some operational steps are as follows:

(1) Determine the Γ values for the distribution. To obtain the reconstructions from simulated data, we shall use the first value $\Gamma_1 = 1\,\text{s}^{-1}$ and the last value $\Gamma_N = 10^4\,\text{s}^{-1}$. We may initially use a broader range and appropriately adjust it after some computational trials.
(2) Choose a reasonable value for the number of Γ values. We shall use $N = 100$ in this paper.
(3) Check the goodness-of-fit to the data, since we are essentially performing a regression analysis on the data. We evaluate the residuals

$$r_i \equiv (y_i - \overline{y}_i), \quad \overline{y}_i = \sum_{j=1}^{N} \Phi_{ij}\mu_j, \qquad (21)$$

where μ_j denotes the estimate of ω_j. It is useful to calculate the residual norm ($Rnorm$)

$$\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\mu}\|_2 = \sqrt{\sum_{i=1}^{M} r_i^2}, \qquad (22)$$

which is a measure of the goodness-of-fit. When each data point has its error σ_i, we should use the residuals

$$r_i \equiv \left(\frac{y_i - \overline{y}_i}{\sigma_i}\right), \quad i = 1, 2, \ldots, M. \tag{23}$$

In the SBL algorithm, most of the parameters α_i will be pushed to infinity with their corresponding ω_i being zero, resulting in only few nonvanishing ω_i's. Normally, 15 or fewer iterations are needed to get a stable sparse distribution.

To obtain smooth PSD, we shall evaluate the reconstructions at various iterations. One of the reconstructions with reasonable goodness-of-fit will be the optimal solution. However, how to obtain the optimal distribution requires a criterion for justifying our choice. The common quantities that change with iterations are the residual norm and the solution norm. The behaviors of these quantities with iterations are dictated by the data quality.

In regularization methods, the L-curve[26–28] provides a visual tool to understand the trade-off between the magnitude of the regularized solution (solution norm) and the quality of the fit to the data (residual norm). The optimal regularization parameter lies on the corner (vertex), the point of maximum curvature, of an L-curve, which is defined by connecting consecutive points in the sequence

$$(\log \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\mu}_k\|_2, \log \|\boldsymbol{\mu}_k\|_2), \quad k = 1, 2, \ldots \tag{24}$$

by straight lines, where k denotes the sequence of regularization parameters. By using the L-curve method, we take k to be the iteration number and determine the required iteration for an optimal, smooth solution from the L-curve [Eq. (24)]. This method can be very useful when the errors in the data are not known. However, in many situations, computational round-off errors and data quality can cause L-curves to differ from an "L" shape.[29]

Alternatively, if an estimate of the variance of the noise vector is known, Morozov's[30,31] criterion can be considered and is related to the chi-square statistic

$$\chi^2 = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{y_i - \overline{y}_i}{\sigma_i}\right)^2, \tag{25}$$

which measures the misfit between actual noisy data y_i and the data $\overline{y}_i$ that would be observed in the absence of noise. Typically, we choose a criterion such that $\chi^2 \approx 1$. At values of $\chi^2 > 1$, the calculated and the observed data are not in agreement, while at values of $\chi^2 < 1$ we fit the noise in the observed data y_i. However, when the errors in the observed data, such as experimental DLS data, are not exactly known, the use of Morozov's or χ^2 criterion fails to give a reliable solution. In this case, the optimal regularization parameter (the iteration number) may then be determined by considering the L-curve.

We also note that in the SBL algorithm, the statistical errors are taken to be the Gaussian type: zero mean and constant standard deviation, while the actual errors in the DLS data are different. For experimental DLS data, we consider the squared deviations of the data to have a Poisson profile, then the deviations of the data y_i are given by[10–12]

$$\begin{aligned}\sigma_i^2 &\approx \frac{1 + y_i^2}{4Ay_i^2} \\ &\equiv \tilde{\sigma}^2\hat{\sigma}_i^2,\end{aligned} \tag{26}$$

where the measured baseline $A = 1/\tilde{\sigma}^2$ is taken as an adjustable parameter in the L-curve, and $\hat{\sigma}_i = \sqrt{(1 + y_i^2)/4y_i^2}$ are data dependent and will be used to rescale our data. The expression (26) shows a reasonable feature of a correlation function that data at longer delay times have larger errors.

By taking the errors to have the form (26), we perform a rescaling procedure on the error term in Eq. (5) and the rescaled (weighted) residual norm is given by

$$\left\|\frac{\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\mu}}{\hat{\sigma}}\right\|_2 = \sqrt{\sum_{i=1}^{M}\left(\frac{y_i - \overline{y}_i}{\hat{\sigma}_i}\right)^2}. \tag{27}$$

3. Data Analysis and Results

Since SBL is a novel method for reconstructing solutions to inverse problems, it is necessary to examine its performance in the analysis of DLS data. We evaluate its efficiency in reconstructing PSD for polydisperse systems from simulated data and also from experimental DLS data.

3.1. *Simulated data*

We simulate several sets of data from log-normal distributions with known mean values and standard deviations. These data provide references to setting the criterion for selecting the optimal PSD. Gaussian noise of $RN/\sqrt{10^7}$ is added to the first-order TCF, where RN denotes Gaussian random noise of zero mean and unit standard deviation. Unimodal and bimodal distributions are considered for our

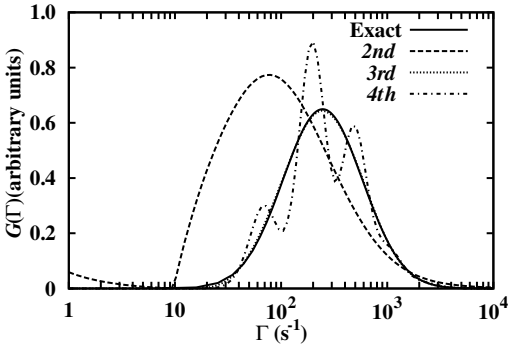

Fig. 1. The solid curve denotes the exact distribution of variance 0.973 used in the simulation. The optimal distribution is given at the 3rd iteration. At the 4th iteration and beyond, the distributions exhibit oscillatory behaviors or multimodal structures.

simulations. For unimodal distributions, polydispersities or normalized variances of $\mu_2/\overline{\Gamma}^2 = 0.973$ and 0.05 are used. For the bimodal distribution, we consider peak ratio of $\overline{\Gamma}_2/\overline{\Gamma}_1 = 2$, where $\overline{\Gamma}_1$ and $\overline{\Gamma}_2$ are the mean values of the first and second peaks, respectively. Both peaks have small variances: $\mu_2^{(1)}/\overline{\Gamma}_1^2 = \mu_2^{(2)}/\overline{\Gamma}_2^2 = 0.01$. The first-order correlation data are generated at 136 delay times, so that the χ^2 statistic [Eq. (25)] defines a residual norm of $\|\boldsymbol{\Phi\mu} - \boldsymbol{y}\|_2 = \sqrt{136/10^7} \approx 3.688 \times 10^{-3}$.

From the three sets of simulated data, several reconstructions at specified iterations from the SBL algorithm are plotted in Figs. 1–3. The L-curves from the reconstructions of the iterations $k = 1, 2, \ldots, 100$ are given in Figs. 4–6. By comparing their χ^2 values at the first few iterations, we would choose the reconstruction at the 3rd iteration in Fig. 1, which has a relatively small *Rnorm* of 3.926×10^{-3}. This value becomes smaller with

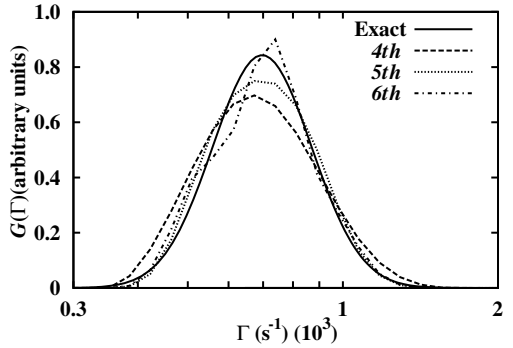

Fig. 2. The solid curve denotes the exact unimodal distribution of variance 0.05. The reconstruction at the 5th iteration provides the optimal distribution. The distribution at the 6th iteration exhibits an irregular feature.

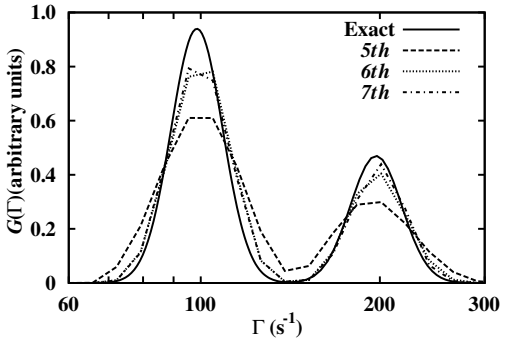

Fig. 3. The solid curve denotes the exact bimodal distribution used in the simulation. The 6th iteration gives the optimal distribution.

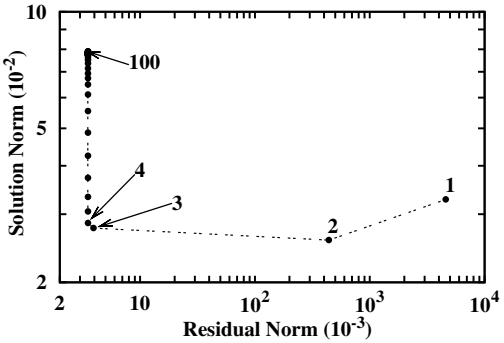

Fig. 4. The L-curve from the reconstructions of iterations $k = 1, 2, \ldots, 100$ for the broad unimodal distribution of variance 0.973. At the 3rd iteration, $Rnorm = 3.926 \times 10^{-3}$, while at the 4th iteration, $Rnorm = 3.518 \times 10^{-3}$.

increasing $N > 100$ and forms the vertex of the L-curve in Fig. 4. The optimal distribution fits nicely with the exact distribution. From the reconstructions in Figs. 2 and 3, their respective L-curves in Figs. 5 and 6 indicate that the choices of optimal distributions are specified by the 5th iteration in

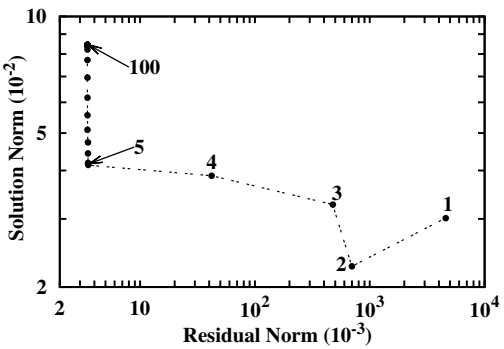

Fig. 5. The L-curve from the reconstructions of iterations $k = 1, 2, \ldots, 100$ for the narrow distribution of variance 0.05. At the 4th iteration, $Rnorm = 4.205 \times 10^{-2}$, while at the 5th iteration, $Rnorm = 3.539 \times 10^{-3}$.

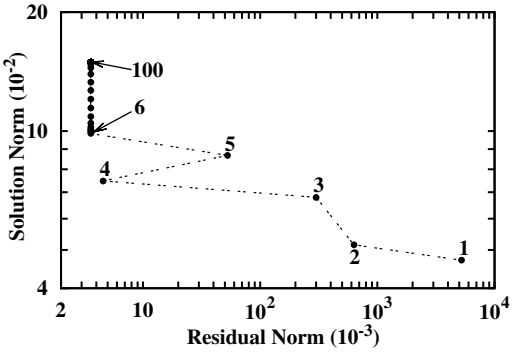

Fig. 6. The L-curve from the reconstructions of iterations $k = 1, 2, \ldots, 100$ for the bimodal distribution of peak ratio $\overline{\Gamma}_2^2/\overline{\Gamma}_1^2 = 2$. At the 4th, 6th and 7th iterations, $Rnorm = 4.554 \times 10^{-3}, 3.597 \times 10^{-3}$ and 3.584×10^{-3}, respectively.

Fig. 2 and 6th iteration in Fig. 3. We note that the fluctuations of the L-curves at small iterations are due to computational round-off errors and data quality. The choices of the optimal distributions agree very well with the exact distributions. We also note that, in the SBL algorithm, an L-curve defines a trajectory that approaches a fixed point of definite solution norm with iterations.

3.2. *Experimental data*

In this section, we apply the SBL algorithm to analyze the DLS data from the ocular lenses of a fetal calf, a Rhesus monkey (two years old), and a man (about 40 years old). We use the diameter range with $d_{\min} = d_1 = 1\,\text{nm}$ and $d_{\max} = d_N = 10^5\,\text{nm}$ for reconstructing the PSD. We consider the statistical errors in the DLS data to have a Poisson profile and use Eq. (27) to define the L-curve. The measurements of the DLS data were made in the nuclear region of the lenses at room temperature using the compact fiber optic DLS probe.[1,2] The TCF data from the lenses of the fetal calf, monkey, and man were measured at 100 delay times from $20\,\mu$s to $8 \times 10^3\,\mu$s.

Figures 7–9 show the reconstructions of the size distributions of α-crystallin and aggregate from the DLS data from the lenses of the fetal calf, monkey, and man, and Figs. 10–12 depict the L-curves with iteration numbers. By following the analysis in the previous section, we choose the iteration numbers for the optimal reconstructions with relatively small *Rnorm*. From Figs. 10–12, we see that for acceptable residual norms, the numbers of iterations needed for the optimal reconstructions are: eight for the calf, six for the monkey, and six for

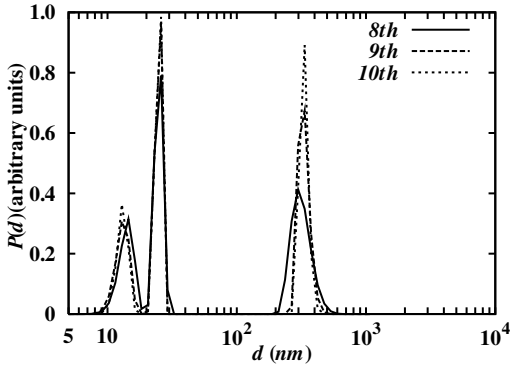

Fig. 7. The solid curve denotes the reconstruction of the optimal size distribution of α-crystallin and aggregate in the lens of the fetal calf. The distributions for $d > 150\,\text{nm}$ have been rescaled by a factor of 30.

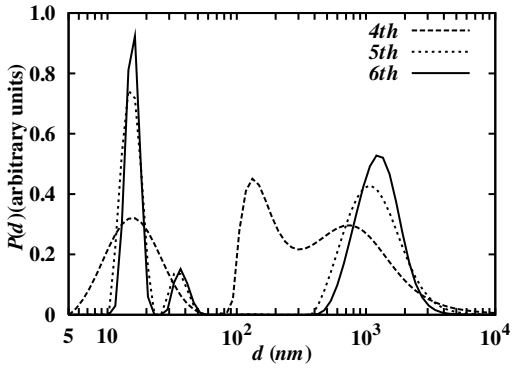

Fig. 8. The solid curve denotes the reconstruction of the optimal size distribution of α-crystallin and aggregate in the lens of the monkey. The distributions for $d > 80\,\text{nm}$ have been rescaled by a factor of 200.

the man. These optimal distributions exhibit multimodal structures. By comparing these distributions, we see that there are peaks of α-crystallin

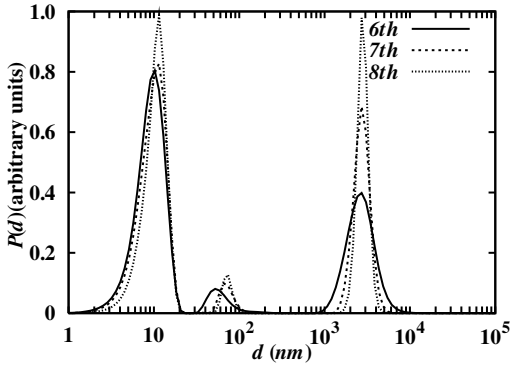

Fig. 9. The solid curve denotes the reconstruction of the optimal size distribution of α-crystallin and aggregate in the lens of the man. The distributions for $d > 500\,\text{nm}$ have been rescaled by a factor of 50.

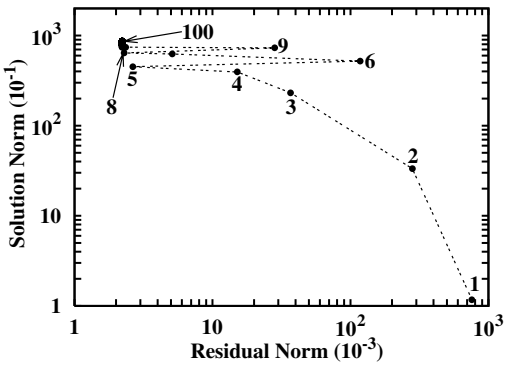

Fig. 10. The L-curve for the DLS data from the lens of the fetal calf. The residual norms at the 5th, 8th, and 10th iterations are respectively: $Rnorm = 2.640 \times 10^{-3}, 2.229 \times 10^{-3}$, and 2.349×10^{-3}.

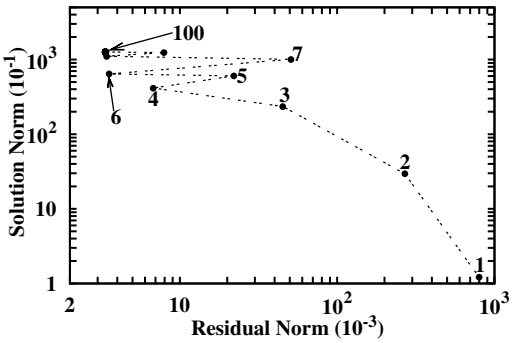

Fig. 11. The L-curve for the DLS data from the lens of the monkey. The residual norms at the 4th, 6th, and 8th iterations are respectively: $Rnorm = 6.776 \times 10^{-3}, 3.545 \times 10^{-3}$, and 3.416×10^{-3}.

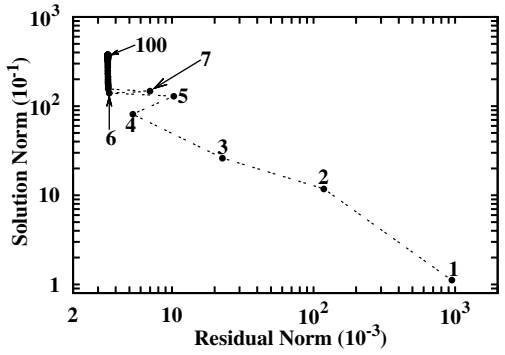

Fig. 12. The L-curve for the DLS data from the lens of the man. The residual norms at the 4th, 6th, and 8th iterations are respectively: $Rnorm = 5.264 \times 10^{-3}, 3.608 \times 10^{-3}$, and 3.545×10^{-3}.

aggregates of large diameters. A significant peak appears in the diameter range $200\,\mathrm{nm} \lesssim d \lesssim 600\,\mathrm{nm}$ for the fetal calf, $400\,\mathrm{nm} \lesssim d \lesssim 4 \times 10^3\,\mathrm{nm}$ for the monkey, and $800\,\mathrm{nm} \lesssim d \lesssim 10^4\,\mathrm{nm}$ for the man. The peak contribution is the smallest for the fetal calf and largest for the man, indicating that aggregation of α-crystallin increases with age. The characteristics of the peaks in the distributions in Figs. 7–9 are listed in Table 1.

To understand quantitatively the implication of protein aggregation in an ocular lens in early cataractogenesis, we refer to a clinical study[4] where

Table 1. The results of the PSD from the SBL analysis are listed. For the iterations corresponding to the reconstructions shown in Figs. 7–9, the first, second, third, and fourth rows list, respectively, the areas (unnormalized), the average diameters (in nm), the standard deviations, and the variances for the peaks in the reconstructions. The α-crystallin indices (ACI) are also given.

Lens	Iteration	Peak			ACI (%)
Fetal calf	8th	0.0933	0.2725	0.1241	74.67
		14.18	25.03	329.7	
		1.81	1.89	61.63	
		0.016	0.006	0.035	
	9th	0.0835	0.2830	0.1320	73.52
		13.24	24.99	334.93	
		1.60	1.39	35.34	
		0.015	0.003	0.011	
	10th	0.0777	0.2858	0.1273	74.06
		13.13	25.02	333.45	
		1.30	1.35	26.61	
		0.01	0.003	0.006	
Monkey	4th	0.3961		0.2152	64.80
		21.05		5865	
		9.10		14000	
		0.187		5.66	
	5th	0.3052	0.1213	0.2040	67.64
		15.75	42.70	1726	
		2.41	25.24	1149	
		0.023	0.35	0.44	
	6th	0.2827	0.1149	0.2031	64.05
		15.96	40.11	1539	
		1.86	13.94	650	
		0.014	0.12	0.18	
Man	6th	0.0813	0.0430	0.2460	33.57
		10.60	76.47	3253	
		3.29	40.61	1399	
		0.096	0.28	0.19	
	7th	0.0839	0.0402	0.2481	33.34
		11.03	75.10	2917	
		3.25	15.23	627	
		0.087	0.041	0.051	
	8th	0.0825	0.0384	0.2467	32.89
		11.35	74.40	2880	
		2.92	10.28	450	
		0.066	0.019	0.024	

an index called the α-crystallin index [ACI (%)] was introduced to provide a measure of unbound α-crystallin in a lens. The α-crystallin, which acts as a molecular chaperone protein, has been shown to bind other damaged lens proteins to prevent their aggregation. In their study, Datiles *et al.* have shown that DLS is able to clinically detect loss of low-molecular-weight α-crystallin proteins even in clinically clear lenses. ACI measurements may be useful in identifying patients at high risk for developing cataract. In particular, from the lenses of humans aged 7–86 years, they showed that ACI was significantly lower in lenses with higher levels of nuclear opacity and with age.

From our reconstructions of the size distributions of α-crystallin and their aggregates in the lenses of the fetal calf, monkey, and man, we calculate the ACI of these lenses and list them in Table 1. These ACI values of (clear) lenses show signs of protein aggregation. In particular, the human lens has greater protein aggregation than either the fetal calf or the monkey lens. Also, referring to the histogram plot for the different age groups,[4] we see that an ACI of 33% for the human lens is much higher than the average ACI value for the 36–45 age group.

As an alternative, simple approach to evaluate the efficiency of the SBL algorithm for early cataract detection, we can obtain the sparse distributions from the DLS data. A sparse distribution is the solution given by the fixed point of an L-curve. Figure 13 shows the sparse distributions of α-crystallin and their aggregates for the calf, monkey, and man. The significant contributions in percentage are given at the dominant particle diameters. The sparse solutions agree strongly with the ACI values in Table 1.

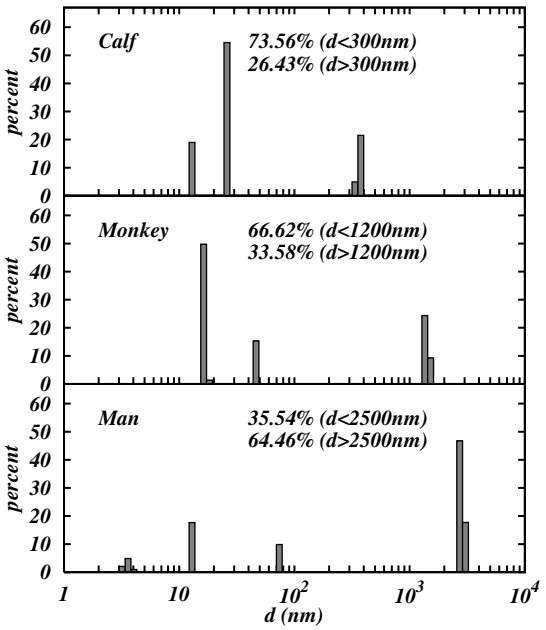

Fig. 13. The percentage of α-crystallin and aggregate in the ocular lenses of the fetal calf, monkey and man.

4. Discussion and Conclusion

In this paper, we have described the SBL algorithm, which is a fully probabilistic analysis tool and can be easily implemented, for reconstructing solutions to inverse problem in DLS. We have shown how to apply the algorithm to analyze DLS data to obtain efficiently reliable estimates of PSD for particle systems.

To evaluate the efficiency of the SBL algorithm, we have considered the simulated data from log-normal distributions and experimental data from the ocular lenses of a fetal calf, a Rhesus monkey, and a man. By using the iterative nature of the SBL algorithm, we have developed an approach to reconstruct optimal solution based on the L-curve method. For the simulated data at a given noise level, the algorithm has been shown to give optimal distributions that agree well with the exact distributions. The algorithm was then applied to reconstruct smooth optimal PSD of α-crystallin and their aggregates in the lenses of the fetal calf, monkey and man. From their optimal size distributions, we see that the size and amount of α-crystallin aggregate increase with age. Similar results can also be obtained (Fig. 13) by noting that the SBL algorithm gives sparse solution to an inverse problem.

Finally, we should note that the use of the SBL algorithm can be computationally demanding. This is particularly the case when we are interested in very smooth, optimal distributions, which are determined iteratively with a large number of grid points N that each iteration requires inversion of an $N \times N$ matrix. However, for the purpose of determining the α-crystallin index of a lens, it suffices to use $N = 30$ to obtain the sparse distribution. In view of its efficiency, the SBL algorithm should receive more attention as an alternative to other analysis tools for the analysis of DLS data in biological studies and, especially, in clinical studies of the lens, where results of high reliability are essential.

Acknowledgments

Part of this work was carried out at the School of Health Information Sciences of the University of Texas Health Science Center at Houston while Dr Nyeo was a visiting Professor and Dr Ansari was a full Professor at UTHSC. The authors are very grateful to Dr J. S. Zigler Jr. for providing the lens samples. Dr Nyeo would like to acknowledge the financial support from the National Science Council of the Republic of China under the Contract No. NSC-97-2112-M-006-006.

References

1. R. R. Ansari, M. B. III. Datiles, "Use of dynamic light scattering and Scheimpflug imaging for the early detection of cataracts," *Diabetes Technology & Therapeutics* **1**, 159–168 (1999).
2. R. R. Ansari, "Ocular static and dynamic light scattering: A noninvasive diagnostic tool for eye research and clinical practice," *J. Biomed. Opt.* **9**, 22–37 (2004).
3. M. Van Laethem, B. Babusiaux, A. Neetens, J. Clauwaert, "Photon correlation spectroscopy of light scattered by eye lenses in *in vivo* conditions," *Biophysical J.* **59**, 433–444 (1991).
4. M. B. III. Datiles, R. R. Ansari, K. I. Suh, S. Vitale, G. F. Reed, J. S. Jr. Zigler, F. L. III. Ferris, "Clinical detection of precataractous lens protein changes using dynamic light scattering," *Arch. Ophthalmol.* **126**, 1687–1693 (2008).
5. D. A. Ross, N. Dimas, "Particle sizing by dynamic light scattering: Noise and distortion in correlation data," *Part. & Part. Syst. Charact.* **10**, 62–69 (1993).
6. H. Ruf, E. Grell, E. H. K. Stelzer, "Size distribution of submicron particles by dynamic light scattering measurements: Analyses considering normalization errors," *Eur. Biophys. J.* **21**, 21–28 (1992).
7. D. E. Koppel, "Analysis of macromolecular polydispersity in intensity correlation spectroscopy: The method of cumulants," *J. Chem. Phys.* **57**, 4814–4820 (1972).
8. J. G. McWhirter, E. R. Pike, "On the numerical inversion of the Laplace transform and similar Fredholm integral equations of the first kind," *J. Phys. A* **11**, 1729–1745 (1978).
9. N. Ostrowsky, D. Sornette, P. Parker, E. R. Pike, "Exponential sampling method for light scattering polydispersity analysis," *J. Mod. Optics* **28**, 1059–1070 (1981).
10. S. W. Provencher, "CONTIN: A general purpose constrained regularization program for inverting noisy linear algebraic and integral equations," *Comput. Phys. Commun.* **27**, 229–242 (1982).
11. A. K. Livesey, P. Licinio, M. Delaye, "Maximum entropy analysis of quasielastic light scattering from colloidal dispersions," *J. Chem. Phys.* **84**, 5102–5107 (1986).
12. S.-L. Nyeo, B. Chu, "Maximum-entropy analysis of photon correlation spectroscopy data," *Macromolecules* **22**, 3998–4009 (1989).
13. C.-D. Sun, Z. Wang, G. Wu, B. Chu, "An improvement on maximum entropy analysis of photon correlation spectroscopy data," *Macromolecules* **25**, 1114–1120 (1992).
14. R. K. Bryan, "Maximum entropy analysis of oversampled data problems," *Eur. Biophys. J.* **18**, 165–174 (1990).
15. J. Langowski, R. Bryan, "Maximum entropy analysis of photon correlation spectroscopy data using a Bayesian estimate for the regularization parameter," *Macromolecules* **24**, 6346–6348 (1991).
16. M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.* **1**, 211–244 (2001).
17. V. N. Vapnik, *Statistical Learning Theory*, Wiley, New York (1998).
18. B. Chu, *Laser Light Scattering: Basic Principles and Practice*, 2nd Edition, Academic Press Inc., Boston (1991).
19. R. Xu, *Particle Characterization: Light Scattering Methods*, Springer-Verlag, Netherland (2000).
20. D. P. Wipf, B. D. Rao, "Sparse Bayesian learning for basis selection," *IEEE Trans. Signal Process.* **52**, 2153–2164 (2004).
21. S. K. Majumder, N. Ghosh, P. K. Gupta, "Relevance vector machine for optical diagnosis of cancer," *Lasers Surg. Med.* **36**, 323–333 (2005).
22. J. T. Flake, *Application of the Relevance Vector Machine to Canal Flow Prediction in the Sevier River Basin*, MSc. Thesis, Utah State University, Logan, Utah (2007).
23. J.-Y. Peng, J. A. D. Aston, R. N. Gunn, C.-Y. Liou, J. Ashburner, "Dynamic positron emission tomography data-driven analysis using sparse Bayesian learning," *IEEE Trans. Med. Imag.* **27**, 1356–1369 (2008).
24. C. L. Lawson, R. J. Hanson, *Solving Least Squares Problems*, Prentice-Hall Inc., Englewood Cliffs, (1974); also available as *Classics in Applied Mathematics*, Vol. 15, SIAM, Philadelphia, PA (1995); and reprinted with a detailed "new developments" appendix in 1996 by SIAM.
25. D. MacKay, "The evidence framework applied to classification networks," *Neural Computation*, **4**, 720–736 (1992).
26. P. J. McCarthy, "Direct analytic model of the L-curve for Tikhonov regularization parameter selection," *Inverse Problems* **19**, 643–663 (2003).

27. P. C. Hansen, "Regularization tools version 4.0 for Matlab 7.3," *Numerical Algorithms* **46**, 189–194 (2007).
28. L. Reichel, H. Sadok, "A new L-curve for ill-posed problems," *J. Comput. Appl. Math.* **219**, 493–508 (2008).
29. G. Rodriguez, D. Theis, "An algorithm for estimating the optimal regularization parameter by the L-curve," *Rendiconti di Matematica, Serie VII* **25**, 69–84 (2005).
30. V. A. Morozov, "On the solution of functional equations by the method of regularization," *Soviet Math. Doklady* **7**, 414–417 (1966).
31. O. Scherzer, "The use of Morozov's discrepancy principle for Tikhonov regularization for solving nonlinear ill-posed problems," *Computing* **51**, 45–60 (1993).