Department of Physics & Astronomy
Ithaca College

pydls

A Python Suite for the Bayesian Inferrence of Particle Size Distributions in Dynamic Light Scattering Experiments

Thy Doan Mai Le
Advisor: Jerome Fung
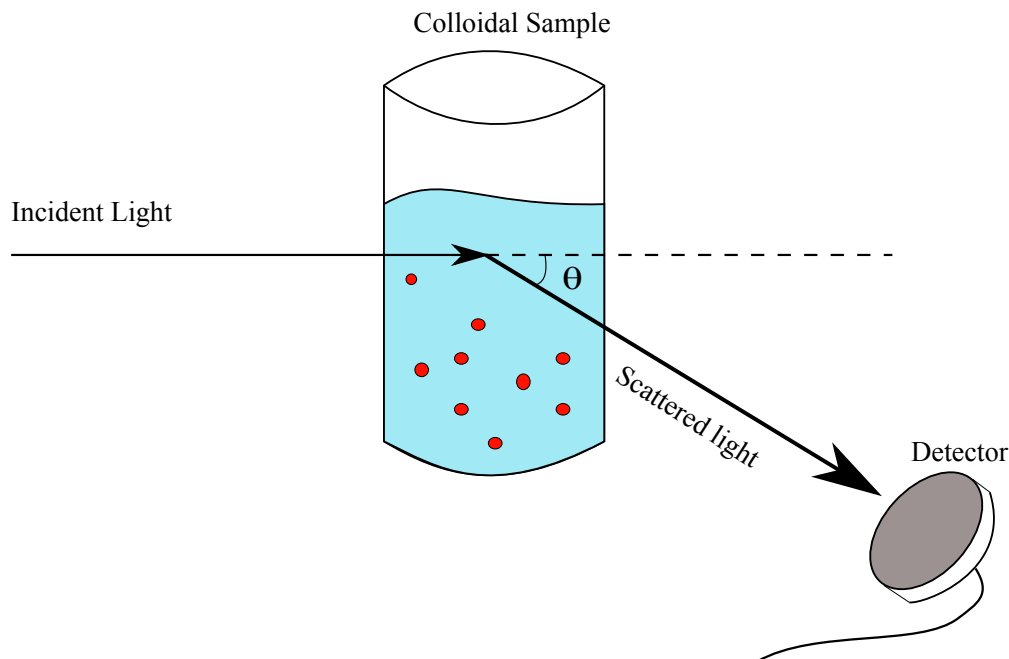
Date: August 28, 2019

Contents

Figure 1: When incident light encounters the colloidal sample, it is scattered by each individual particulate molecule within the sample and upon exiting the sample, all individually scattered light rays come together to give an interference pattern that is then received by the detector. Because the individual particles within the sample all experience random Brownian motion, the interference pattern is predicted to have a fluctuating intensity as the molecules move into and out of the incident light beam.

1   Dynamic Light Scattering

Dynamic ligh scattering (DLS) is a technique used to determine the size distribution of a sample of particles suspended in an optically transparent medium, also known as a colloidal sample. When incident light is directed towards a colloidal sample, each particle within the sample scatters the incident light independently and this same process repeats for all of the particles within the sample. A sample schematic of a DLS experiment is found in Fig. 1. Because each molecule experiences random Brownian motion while suspended in the sample, the intensity of the scattered light fluctuates as these molecules move about. A similar phenomenon occurs with starligh in the night sky. When ligh from distant stars reaches the Earth's atmosphere, the particles in the atmostphere scatter this light. These particles are not stationary, however, but move about constantly in a randomized fashion. As a result, the scattered starlight is sometimes brighter, sometimes dimmer and our eyes register this fluctuation as "twinkling" stars.

For colloidal samples, the scattered light also "twinkles" and the "twinkling" pattern is actually distinct depending on the size, shape and other physical properties of the particles in the sample. Larger particles experience more drag during Brownian motion due to their larger surface area. Due to the larger drag, their Brownian motion occurs in a more gradual fashion over time. Because these larger particles are moving much more slowly, they produce scattered light that "twinkles", or fluctuates in intensity, very gradually over time. Smaller particles, because they experience a much smaller drag due to their reduced size, move about very quickly, changing their position very
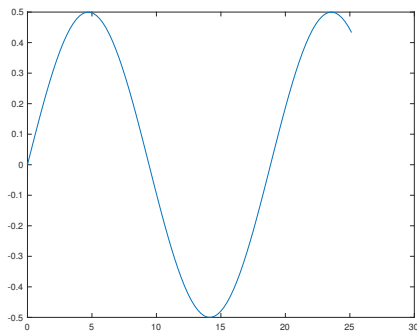
Figure 2: Large particles experience more drag and change their positions much more slowly than smaller particles. As a result, the scattered light's intensity changes very slowly with time. If the particle was moving in a sinusoidal motion, then the scattered light's intensity would have a similar, slowly varying sinusoidal pattern.
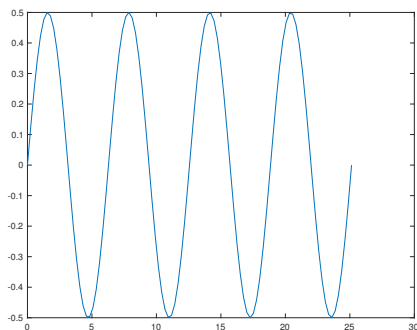

Figure 3: Smaller particles experience less drag and thus change their positions much faster than larger particles. Because of that, the light that they scatter has a very rapidly changing intensity. If the small particle was moving in a sinusoidal motion with a very high frequency (which is typical of a fast moving particle), the scattered light intensity will have a sinusoidal pattern with a high frequency as well.

rapidly in time. As a result, the scattered light from smaller particles fluctuates in intensity very quickly.

In real-world situations, particles in Browninan motion do not have perfectly sinusoidal and periodic trajectories that scatter light so that the resultant intensity is perfectly periodic. Typically, scattered light from colloidal samples often have an aperiodic pattern with a lot of noise added in. In order to extract any useful information from the fluctuating scattered light, the scattered light's intensity that is detected by the detector in Fig. 1 is used to produce an autocorrelation function. A correlation function demonstrates the degree of similarity between two consecutive data points in a data set. In the case of DLS experiments, the correlation function is expected to be an exponential decrease because over time, fluctuations in the detected signal cause the similarity between consecutive data points to dramatically fall. As a result, the correlation function takes a nosedive and resembles an exponential drop. In order to extract useful information about the colloidal sample, mathematical models are fit against the correlation function. In this project,

the mathematical model proposed by Buoalem, Jabloun, Ravier, Naiim and Jalocha is used in an implementation of Monte Carlo Markov chain to infer the particle size distribution of a colloidal sample, given its autocorrelation function from a DLS experiment at one single angle.

2 Theoretical Background

2.1 Prior Distribution

According to Buoalem, Jabloun, Ravier, Naiim and Jalocha, the prior distribution of this non-parametrized method is as follows:

$$p(\mathbf{f}) \propto exp\left(\mathbf{f}^T\mathbf{L_2}^T\mathbf{L_2}\mathbf{f}\right), \text{if } \mathbf{g} \geq 0 \tag{1}$$

2.2 Likelihood Distribution

The likelihood distribution is modeled according to the $g^{(2)}$ function as

$$p(\tilde{\mathbf{g}}^{(2)}|\mathbf{f}, \sigma_r^2) = \frac{1}{(2\pi)^{\frac{M_r}{2}}\sigma^{M_r}} exp\left(-\frac{\chi_r(\mathbf{f})}{2\sigma_r^2}\right) \tag{2}$$

if $\mathbf{f} \geq 0$.

The χ-squared $\chi_r(\mathbf{f})$ is defined as

$$\chi_r(\mathbf{f}) = \sum_{j=1}^{M_r} \left(\tilde{g}_{\theta_r}^{(2)}(\tau_j) - g_{\theta_r}^{(2)}(\tau_j)\right)^2 \tag{3}$$

2.3 Posterior Distribution

With the prior and likelihood distributions described in Eq. 1 and Eq. 2, the posterior distribution for the particle size distribution with the noise variance given the measured autocorrelation data $(g^{(2)})$ is:

$$p(\mathbf{f}, \sigma_1^2, \ldots, \sigma_r^2|\tilde{\mathbf{g}}^{(2)}) \propto exp\left(-\mathbf{f}^T\mathbf{L_2}^T\mathbf{L_2}\right) \prod_{r=1}^{R}\left[\frac{1}{\sigma_r^{M_r+2}} exp\left(-\frac{\chi_r(\mathbf{f})}{2\sigma_r^2}\right)\right] \tag{4}$$

Using the classical Jeffrey prior density for the noise variance in $\tilde{\mathbf{g}}^{(2)}$,

$$p(\sigma_r^2) \propto \frac{1}{\sigma_r^2} \tag{5}$$

the posterior distribution can be integrated over the noise variance with the following integral:

$$\int_0^\infty \frac{1}{\sigma_r^{M_r+2}} exp\left(-\frac{\chi_r(\mathbf{f})}{2\sigma_r^2}\right) d\sigma_r \tag{6}$$

In order to integrate this integral, let $\sigma_r = \frac{1}{t}$ and thus $d\sigma_r = -\frac{1}{t^2}dt$. From this, we can rewrite expressions from the integrand in Eq. 6 as follows.

$$\frac{1}{\sigma_R^{M_r+2}} = t^{M_r+2}$$
$$\frac{\chi_r(\mathbf{f})}{2\sigma_r^2} = \frac{\chi_r(\mathbf{f})}{2}t^2 \tag{7}$$

Thus the integral from Eq. 6 becomes:

$$\int_\infty^0 t^{M_r+2}\left(-\frac{1}{t^2}\right) exp\left(-\frac{\chi_r(\mathbf{f})}{2}t^2\right) dt$$
$$= \int_0^\infty t^{M_r} exp\left(-\frac{\chi_r(\mathbf{f})}{2}t^2\right) dt \tag{8}$$

The limits switched from $(0, \infty)$ in Eq. 6 to $(\infty, 0)$ in Eq. 8 because of the substitution $\sigma_r = \frac{1}{t}$. The reciprocal relationship of the substitution switched the limits of the integration. In order to make the integral easier to integrate, the minus sign on the second line of Eq. 8 is exchanged for the re-switching of the integration limits from $(\infty, 0)$ back to $(0, \infty)$.

In order to actually integrate Eq. 8, a second substitution has to be made. Let $\tau = t\sqrt{\chi_r}$ and $d\tau = \sqrt{\tau}dt$. From this, we can rewrite the expressions in the integrand of Eq. 8 as follows.

$$\chi_r(\mathbf{f})t^2 = \tau^2$$
$$t^{M_r} = \left(\frac{\tau}{\sqrt{\chi_r}}\right)_r^M$$
$$= \tau^{M_r}\chi_r^{-\frac{M_r}{2}} \tag{9}$$

From the above substitutions, the integral becomes:

$$\int_0^\infty \chi_r^{-\frac{1}{2}}\tau^{M_r}\chi_r^{-\frac{M_r}{2}} exp\left(-\frac{\tau^2}{2}\right) d\tau$$
$$= \int_0^\infty \chi_r^{-\frac{1}{2}(M_r+1)}\tau^{M_r} exp\left(-\frac{\tau^2}{2}\right) d\tau \tag{10}$$

Since the χ_r term does not have τ dependence, we can pull that out of the integrand, giving us the following integral at the end:

$$\chi_r^{-\frac{1}{2}(M_r+1)} \int_0^\infty \tau^{M_r} exp\left(-\frac{\tau^2}{2}\right) d\tau \tag{11}$$

Recall that the χ_r expression was previously

$$\chi_r(\mathbf{f}) = \sum_{j=1}^{M_r} \left(\tilde{g}_{\theta_r}^{(2)}(\tau_j) - g_{\theta_r}^{(2)}(\tau_j)\right)^2 \tag{12}$$

where $\tilde{g}^{(2)}$ is modeled to be:

$$\tilde{g}^{(2)}_{\theta_r}(\tau_j) = 1 + \beta_r \left(k_{\theta_r} \sum_{i=1}^{N} f(D_i) C_{I,\theta_r}(D_i) exp \left(-\frac{\Gamma_{0,\theta_r}}{D}\tau_j \right) \Delta D_i \right)^2$$

$$\text{where } \Gamma_{0,\theta_r} = \frac{16\pi n^2 \sin^2(\theta/2) k_B T}{2\lambda_0^2 \eta} \tag{13}$$

with

$$
\begin{aligned}
&\text{n : refractive index} \\
&\theta : \text{angle at which scattered light was measured} \\
&k_B : \text{Boltzmann constant} \\
&T : \text{temperature in Kelvin} \\
&\lambda_0 : \text{incident wavelength value in vacuum} \\
&\eta : \text{viscosity} \\
&D_i : \text{diameter of particle size} \\
&\beta_r : \text{instrumental factor} \\
&k_{\theta_r} : \text{normalization constant that can be found via k} = \frac{1}{\int_0^\infty f(D)C_{I,\theta}(D)dD} \\
&\tau_j : \text{delay times} \\
&f(D) : \text{the particle size distribution function}
\end{aligned}
\tag{14}
$$

The posterior distribution is now defined to be:

$$p(\mathbf{f}|\tilde{\mathbf{g}}^{(2)}) \propto exp\left(-\mathbf{f}^T \mathbf{L_2}^T \mathbf{L_2} \mathbf{f}\right) \prod_{r=1}^{R} [\chi_r(\mathbf{f})]^{-\frac{M_r}{2}} \tag{15}$$

Since $\chi_r(\mathbf{f}) = \sum \left(\tilde{g}^{(2)}_{\theta_r}(\tau_j) - g^{(2)}_{\theta_r}(\tau_j) \right)^2$, the posterior distribution is essentially:

$$p(\mathbf{f}|\tilde{\mathbf{g}}^{(2)}) \propto exp\left(-\mathbf{f}^T \mathbf{L_2}^T \mathbf{L_2} \mathbf{f}\right) \prod_{r=1}^{R} \left(\sum (\tilde{g}^{(2)}_{\theta_r}(\tau_j) - g^{(2)}_{\theta_r}(\tau_j))^2 \right)^{-\frac{M_r}{2}} \tag{16}$$

When the log of the posterior is taken, the above equation becomes:

$$\ln(p) = -\mathbf{f}^T \mathbf{L_2}^T \mathbf{L_2} \mathbf{f} + \left(-\frac{M_r}{2} \right) \ln \left(\sum \left(\tilde{g}^{(2)}_{\theta_r}(\tau_j) - g^{(2)}_{\theta_r}(\tau_j) \right)^2 \right) \tag{17}$$

3 Development

In order to test the functionality of pydls, the package was used to infer simulated monodisperse and bidisperse Gaussian particle size distributions. Each simulated particle size distribution (PSD)
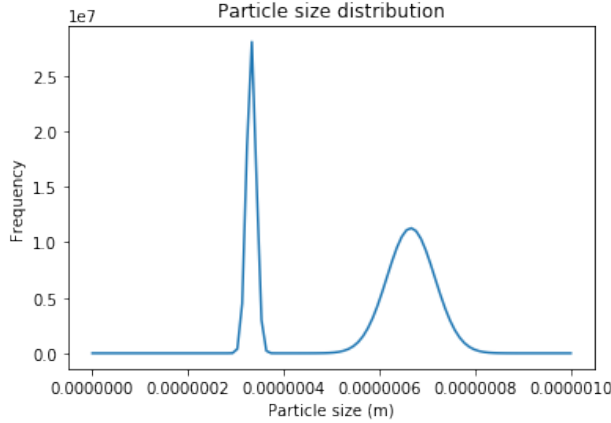
Figure 4: The distribution of particle size (not normalized here) has a peak at 3.33×10^{-7} meter and another peak at 6.66×10^{-7} meter

is first used to generate a g^2 function as described in Eq. 13, which is then is fitted using the exponential decay model to estimate the approximate mean and formulate a Gaussian distribution around this mean, as if we had no idea what the original PSD looked like. The g^2 data is now fed into the infer function in the pydls package with the approximated PSD as starting point for the MCMC walkers. The infer function will then give results from all of its walkers and utility functions within pydls can be used to extract the final particle size distribution, which will then be compared to the original, simulated PSD.

3.1   Bidisperse Distribution

The following self-generated particle size distribution with two peaks of different full-width-half-max (FWHM) values was used to test the validity of the $g^{(2)}$ function on a bimodal distribution. The reason for this test is to verify that the implementation written is able to work with a bimodal size distribution, since one of the advantages of the Direct Bayesian Method (DBM) is that it is able to work with almost any particle size distribution.

From this distribution, the $g^{(2)}$ autocorrelation function of intensity v. time was found to be:

In order to verify the validity of this autocorrelation function, I used SciPy's optimization for curve fits to extrapolate, from this autocorrelation function, the single exponential decay and the method of cummulants models. The goal of this method of testing is to verify that both of the single exponential decay and cummulants models fail upon the emergence of a bimodal particle size distribution. In order to detect this anticipated failure, the residuals between the $g^{(2)}$ model and each of the tested models (single exponential/cummulants) are plotted. Failure of each of the tested models is present when overshooting and/or undershooting are present in the residuals plot.

When tested against the single exponential decay model, the residuals plot is depicted in Fig. 6.

When tested against the cummulants model, the residuals plot is as depicted in Fig. 7.

In conclusion, the residuals plots have shown that both the single exponential and the cummulants model were unable to fully capture the information displayed in the $g^{(2)}$ autocorrelation
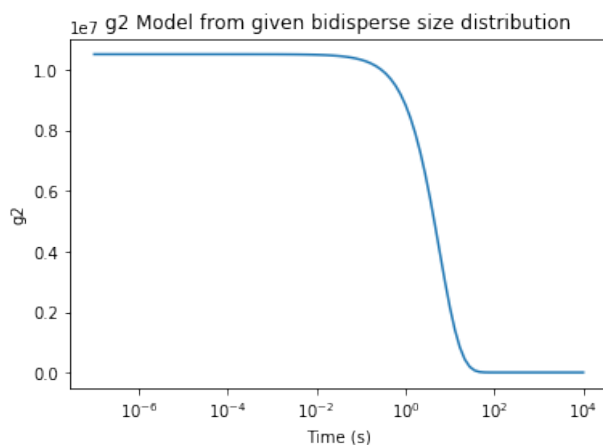
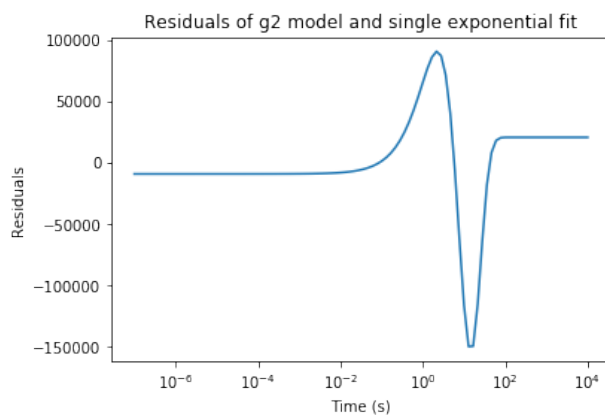Figure 5: The $g^{(2)}$ autocorrelation function extrapolated from the implementation.

Figure 6: The undershooting and overshooting are both apparent around the region of exponential decay, implying that the exponential fit is unable to capture certain aspects of the autocorrelation. This is congruent with the given data. Since single exponentials are only able to capture information of systems with 1 single particle size, a bimodal distribution of particle size should render the single exponential model inept.
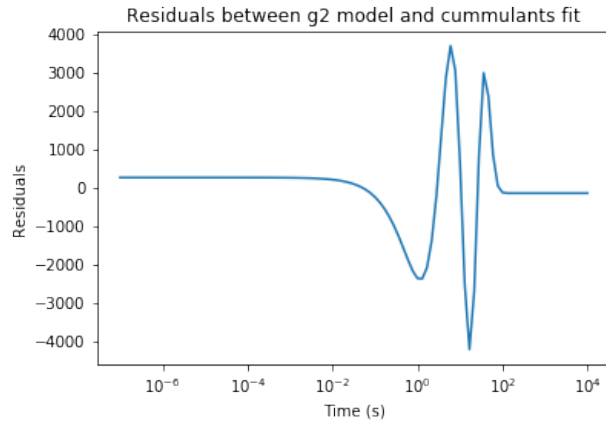
Figure 7: The residuals between the autocorrelation and the cummulants fit model showed dramatic breakdown around the exponential decay region, signifying that the cummulants fit was not able to truly capture the $g^{(2)}$ information.
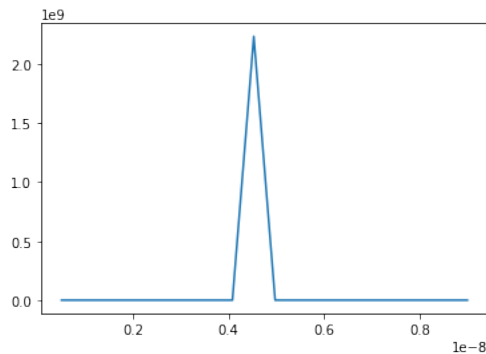

Figure 8: The simulated particle size distribution with a mean of 4.488×10^{-9} and a standard deviation of 6.67×10^{-11}.

and therefore should not be used in analyzing dynamic light scattering information of samples with know multimodal particle size distributions.

3.2 Single Gaussian distribution

Write about tests of the g(2) model using a single Gaussian distribution of particle sizes. In this test, the expected results are that the method of cumulants and the single exponential model both fail to completely capture the physics, shown by a residual plot that exhibits overshooting and undershooting of results.

3.3 Bayes Take 20 - July 8, 2019

After 19 previous rounds of debugging, the package was finally stable and ready for testing using a variety of particle size distributions: narrow, wide, monodisperse or bidisperse. In Bayes Take 20, performed on July 08, 2019, the simulated particle size distribution has a mean at 4.488×10^{-9}
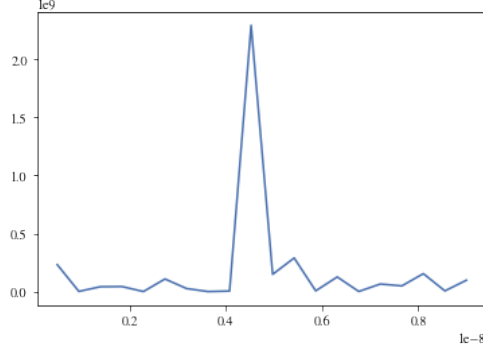
Figure 9: The inferred particle size distribution using pydls on the g2 data in Fig. 10 and the environmental factors above.
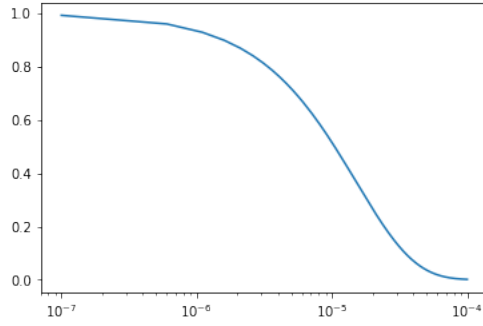

Figure 10: The g2 function using the stimulated particle distribution and the environmental factors as described above.

and a standard deviation of 6.67×10^{-11}. Throughout all following tests, environmental conditions are kept to be:

- m = 20

- c = 1

- $\eta = 1 \times 10^{-3}$

- $\theta = \pi/2$

- n = 1.33

- $k_B = 1.38 \times 10^{-23}$

- T = 298.15K

- $\lambda_0 = 638 \times 10^{-9}$

- $\beta = 1$

Using the g2 data as well as the simulated PSD, the inference was able to successfully yield a PSD, as depicted in Fig. 9, that resembled the original simulated PSD, albeit with some sharp, small magnitude peaks in locations that previously had zero probability in the original PSD.

3.4 Further Testing

Further testing can be found at `https://github.com/thydmle/pydls/tree/master/Test`, starting with Bayes Take 21 and finishing at Bayes Take 33.

4 Conclusion

Comprehensive testing showed that pydls is successfully able to handle DLS autocorrelation functions from single angle data with very high accuracy. Further development should target the expansion of this package into a code base that is able to handle multi-angle DLS data with accurate (instead of an approximation) Mie scattering coefficient calculations.