Automation of Guinier Analysis in SAXS

2016-07-11

Nobutaka SHIMIZU and Masatsuyo TAKAHASHI

Photon Factory

Institute of Materials Structure Science

High Energy Accelerator Research Organization, KEK

Contents

0 Introduction

a It is not a simple task to programmatically determine a radius of gyration and other parameters in Guinier analysis from experiment data in SAXS.

b This document describes how we tried to solve the problem in a sufficiently simple way and implement it into a maintainable set of programs.

c By “simple”, we mean here mainly that it is composed of a simple combination of well-established methods with fewer customizations.

1 Understanding of the Task

a Guinier Plot, as shown in , is one of the most fundamental steps in the analysis of the SAXS experiment data.

b I(0) and Rg Estimation

c Quality Assessment

2 Goal of the Program

a The goal of the program is set to establish a good balance of the sub-goals suggested by the following key words.

- ✧ accuracy of the estimation of radius of gyrations, etc.
- ✧ usable quality indicator on the input data
- ✧ stability and robustness
- ✧ feasible calculation speed

3 Solution Scheme

3.1 From Guinier-Porod to Guinier

a Our solution scheme mainly consists of two parts. The first part is the preliminary fitting to the Guinier-Porod model, followed by Guinier analysis based on the fitting results.

b From the nature of two-region Guinier-Porod model, the boundary of the two basic regions can be calculated simply from the fitting parameters, namely R_g (radius of gyration) and d (Porod exponent). See appendices A.5 for details.

c If a set of data fits to the model, then we can use the calculated boundary as the approximate end point of the Guinier interval.

d Otherwise, the set of data may be of too low quality with large dispersion, or it may be suspected to have some kind of anomalies such as aggregation or inter-particle interference.

e In short, we can get useful information from Guinier-Porod model fitting both in success or in failure.[1]

f Since our goal here is the automation of Guinier analysis, we will finally stick to the pure Guinier approximation, which is implemented as a linear regression of a certain selected interval.

3.2 Why this scheme is simple and stable

a If you have determined a Guinier region interval, the remaining work is, basically, a linear regression and routine calculations.

b So, the difficulty of the problem lies rather in the determination of an appropriate interval.

c Determining an optimal interval from the shape information of the curve is a complex process consisting of several self-made steps such as smoothing the measurement data curve, computing its derivatives and curvatures, etc.

d So, it is error-prone even when the condition of the data is relatively good and almost impossible when it is bad.

e On the other hand, the above mentioned scheme begins with a well-established single step — the least squares fitting to a relatively simple data model — and is die-hard in the approximate determination of the interval. In addition, it rarely results in completely wrong estimations. (See A.5 f on page 20 for the risk of wrong estimations)

f So, it is easier to handle uniformly from bad to good, corresponding to the quality of the data.

g The uniformity or continuity of the process to data quality is essential for the stable calculation of the quality index, especially for judging bad when it is bad.[2]

---

[1] Even when the fit for the whole interval fails, it can be fitted for a sufficiently narrower interval and we can get information from the fit as well.

[2] R_g estimation can get really bad when the data is really bad. Nevertheless, it is better than nothing to give some information as long as it is given correctly marked bad with the quality indicator index. What must be avoided are the cases where bad results are given with relatively high quality scores.

4 Formulation of the Optimizations

a In this chapter, we will describe the implemented optimizations conceptually, leaving the precise description or implementation details to the later chapters.

4.1 Guinier Porod Fitting

a Although the Guinier Law is observed in logarithmic intensity versus scattering variable squared (q^2) scaling, we let the program perform the Guinier-Porod model fitting in the simple plain scaling as follows without logarithm and q squaring.

$$\text{minimize: } S(G, R_g, d) = \sum_{i=1}^{N} \frac{\left(y_i - F_{G,R_g,d}(x_i)\right)^2}{{e_i}^2},$$

where N : number of obsevations,
x_i : ith value of scattering variable (q),
y_i : average intensity at x_i, (4-1)
e_i : standard error of y_i,
$F_{G,R_g,d}$: value of the model with G, R_g and d,
G : scale factor of the model,
R_g: radius of gyration,
d : Porod exponent

b Note that, in this formulation, the least squares optimization evaluates the deviations in the lower q ranges quite stronger because the intensity is exponentially higher in the region.

c So, this plain scaling, rather than Guinier Plot scaling, is considered more suitable for this program's purpose in the following sense.

d Because the ultimate goal of the program is Guinier analysis, it is not appropriate to be overly affected by the shape information from the Porod region.

e In other words, we want it to be fitted roughly as well for those data that are considered to fit badly to the Guinier-Porod model. Data from samples in the form of surface fractals are said to be such cases. See Appendixes A.5 and C for details.

f The range (or interval) in the q-axis for the fitting is initially determined after removing the head anomalies. See Section 6.1 for details.

g When the fit is too bad, the program retries fitting after narrowing the interval in several ways. See Section 6.2 for details.

4.2 Guinier Interval Determination

a After Guinier-Porod fitting we have an approximate knowledge of the boundary of the two basic regions.

b Since the automatically calculated boundary Q_1 tends to exceed the Guinier interval, an approximate value of our wanted upper boundary is basically determined by the $qR_g < 1.3$ constraint.

c I.e., we can get an approximate value $\hat{Q}$ of the upper boundary from the following formula.

$$\hat{Q} = \min\left(\frac{1.3}{R_g}, Q_1\right), \quad \text{where } Q_1 = \frac{1}{R_g}\left(\frac{3d}{2}\right)^{1/2}. \tag{4-2}$$

d And the value tends to be determined depending only on R_g, since $q = 1.3/R_g$ is often smaller than Q_1. See Appendixes A.5 and C for details.

e The lower boundary is determined approximately by the following optimization consisting of three factors.

$$\text{minimize: } D(q_1, q_2) = w_1\left(F_{consistency}\right)^2 + w_2(F_{size})^2 + w_3(F_{start})^2,$$

where

$$F_{consistency} = \left(R_g(q_1, q_2) - \widehat{R_g}\right)^2, \quad F_{size} = \frac{1}{(q_2 - q_1)^2}, \quad F_{start} = {q_1}^2 \tag{4-3}$$

q_1 : the lower boundary of Guinier interval,

q_2 : the upper boundary not exceeding Q_1 determined by formula (4-2),

$R_g(q_1, q_2)$: R_g determined approximately on the interval $[q_1, q_2]$,

$\widehat{R_g}$: Gunier-Porod fitted R_g,

w_i : weights for evaluation.

f We said "approximately" above because the real implementation is more complex to cope with the dispersion of the experiment data. See 6.5 for $F_{consistency}$ details.

g The real factors and weights in implementation need to be adjusted considering various situations although not specified here.

4.3 Rg Estimation

a After the determination of the Gunier interval, R_g calculation is a relatively simple task of weighted linear regression formulated as below.

$$\text{minimize: } E(A,B) = \sum_{i=1}^{N} \frac{\left(Y_i - L_{A,B}(X_i)\right)^2}{{e_i}^2},$$

where

N : number of obsevations in the interval, (4-4)

$X_i = {x_i}^2$: ith value of scattering variable (q) squared,

$Y_i = \log(y_i)$: natural logarithmic average intensity at x_i,

e_i : standard error of y_i,

$L_{A,B}(X_i) = A + BX_i$.

b And the corresponding I_0 and R_g are calculated respectively from the intercept A and the slope B as follows.

$$I_0 = exp(A). \quad (4\text{-}5)$$

$$R_g = (-3B)^{1/2}. \quad (4\text{-}6)$$

c Since the $qR_g < 1.3$ constraint had been only approximately satisfied in the previous stage and the linear regression does not consider the constraint, the final R_g must be slightly modified by re-calculating after adjusting the interval so as to satisfy the constraint. See 6.6 for details.

5 Quality Index Design

5.1 Quality Factors and their Proportions

a In order to achieve the usability mentioned at the early part of the document, we designed the quality index as the sum of those quality factors whose scoring proportions are listed in Tab. 5-1.

Tab. 5-1 Quality Factors and their Proportions

No	Factor Name	Proportion	Description
1	intensity score	0.1	Higher if the GP-fitted I_0 is higher.
2	positive score	0.1	Higher if the data contains less negative values.
3	QP-fitting cover ratio	0.1	Higher q-range that fits well to GP model is wider.
4	QP-fitting score	0.0	Used internally, but weighted zero for itself.
5	R_g stdev score	0.2	Higher if the R_g stdev is smaller.
6	fit consistency	0.4	Higher if the estimated R_g is closer to the GP-fitted R_g.
7	qR_g range score	0.1	Higher if the minimum qR_g is smaller.

b The listing order of factors above is intended to represent the order of logical dependence among them, although the actual computation is slightly different.

c The value of each factor is defined to fall in the interval between zero and one, letting it to mean that the greater the value the better the quality.

d Explanation for each factor will be given in the following sections.

5.2 Intensity Score

a Until the intensity of the x-ray beam attains a certain level, the contrast is not enough to distinguish the sample and the buffer, and tends to result in a very low quality of data.

b In this sense, this score is the most fundamental.

c We chose to evaluate this score from the I_0 obtained in the preliminary GP-fitting.

d It would have been more desirable if we could calculate it from the more primitive value such as the intensity itself. However, that was found to be inappropriate because it is observed that there are some cases where weaker intensities produce better contrasts than stronger intensities.

e The intensity score is calculated linearly, flattening the values outside of the evaluation range, according to Tab. 5-1 below.

f The boundary values at both ends were chosen so that the ignorance of differences in less or grater values than those is insignificant in the primitive comparison of the qualities.

Tab. 5-2 Definition of Intensity Score

Range of I_0	Intensity Score
$I_0 < 0.0005$	0
$0.0005 \leq I_0 \leq 0.002$	$(I_0 - 0.0005)/(0.002 - 0.00005)$
$I_0 > 0.002$	1

5.3 Positive Score

a Negative intensity values result from subtraction in low contrast, and they are inconvenient to logarithmic treatment employed in the analysis.

b Therefore, they are counted to get their counterpart proportion to the whole data before removing to make the rest available in the later stages.

c The positive score is defined from the proportion of positive intensities as follows.

$$positive\ score = 2(\max(0.5, positive\ ratio) - 0.5),$$

where

$$positive\ ratio = \frac{number\ of\ positive\ intensities}{number\ of\ measured\ intensities}.$$

5.4 GP-fitting Cover Ratio

a For brevity, we assume here that method to distinguish whether the fit is acceptable or not is somehow known. See 6.2 for details

b If the whole data does not fit well to the Guinier-Porod model, the program tries to find a narrower interval within which the data fits well to the model.

c In this way, there always exists an interval where the fit is acceptable.

d Therefore, the GP-fitting cover ratio is defined as follows.

$$GP\text{-}fitting\ cover\ ratio = \frac{number\ of\ positive\ intensities\ in\ the\ well\text{-}fitted\ interval}{number\ of\ positive\ intensities}.$$

5.5 GP-fitting Score

a This score represents how well the set of data fits the model.

b It is defined as follows based on the AIC — Akaike Information Criterion — of the fit result.

$$GP\text{-}fitting\ score = \frac{aic\ score}{10 - 8},$$

$$aic\ score = \min\left(10, \max\left(8, \ln(-aic)\right)\right),$$

$$\text{where } \ln(-aic)\ is\ interpreted\ as\ zero\ when\ aic > 0$$

c The range of the $aic\ score$ between 8 and 10 was chosen from the observation of actual data so that the score adequately represent the quality of the model fitting.

5.6 Rg Stdev Score

a This score represents the reliability of the estimated R_g in the Guinier interval.

b It is defined as follows based on the stdev ratio of the estimated R_g in the interval.

$$R_g\ stdev\ score = \exp\left(-50 \times \frac{R_g\ stdev}{R_g}\right)$$

c The constant 50 was chosen so that the value of the score varies between zero and one roughly reflecting the human evaluation of the actual data.

5.7 Fit Consistency

a This score represents the degree of consistency between the values of R_g's from the Guinier-Porod fitting and the final evaluation in the Guinier interval.

b It comes after the previous two factors because you cannot tell the consistency correctly unless both of the R_g's are reliable.

c In other words, there is a risk that the score may become accidentally high when both of them get close together in a coincidence, and we have to avoid such cases.

d The score is defined as a product of the three factors as follows.

$$fit\ consistency = raw\ fit\ consistency \times GP\text{-}fitting\ score \times R_g\ stdev\ score,$$

where

$$raw\ fit\ consistency = 0.4 \times base\ score + 0.6 \times bonus\ score$$

$$\text{base score} = diff\ score$$

$$bonus\ score = \max(0, diff\ score - 0.9\ )$$

$$diff\ score = \max\left(0, 1 - \frac{\text{abs}\left(estemated\ R_g - GP\text{-}fitted\ R_g\right)}{GP\text{-}fitted\ R_g}\right)$$

e Note that the "$raw\ fit\ consistency$" is defined so as to amplify the differences in cases when the "$diff\ score$" is higher than 0.9 (, i.e., the difference ratio is less than 0.1).

5.8 qRg Range Score

a The maximum value of qR_g in the Guinier region is usually limited by the $qR_g < 1.3$ restriction.

b On the other hand, the minimum value better represent the quality to show that the interval is chosen widely enough to get a reliable estimation of R_g.

c The score is defined as follows from the minimum value of qR_g.

$$qR_g\ range\ score = primitive\ range\ score \times intensity\ score,$$

where *primitive range score* is defined to linearly interpolate the following table.

Tab. 5-3 Definition Table of the qRg primitive Range Score

Minimum value of qR_g	qR_g primitive range score
～0.3	1
0.4	0.8
0.5	0.6
0.6	0.4
0.7	0.2
0.8～	0

d Multiplication by the *intensity score* in the definition is added to avoid over evaluation for low quality data with insufficient intensities.

e As for the reliability of the R_g estimation, the score gives a different point of view other than the $R_g\ stdev\ score$.

6 Outline of the Program

a Removing starting-q-range anomalies

b Guinier-Porod model fitting

c Determination of a sufficiently wider interval

d Smoothing the curve in the approximate interval

e Determination of an optimal Guinier interval

f Estimation of the Rg

g Estimation of the errors

h Calculation of the quality index

6.1 Removing starting-q-range Anomalies

6.2 Guinier-Porod Model Fitting

6.3 Determination of a sufficiently wider Interval

6.4 Smoothing the Curve in the approximate Interval

6.5 Determination of an optimal Guinier Interval

a So far, we have two smooth curves, i.e., one is a Guinier-Porod fitted curve on the whole interval and the other is a spline curve obtained in the approximate partial interval. (See Fig. 6-1 for example curves.)

b The former curve is linear when seen in Guinier Analysis axes, i.e., in q^2 for x-axis and $Ln(I)$ for y-axis, so that it can be represented by its slope[3], which is transformed into the Rg.

c We use the latter curve to make the objective function smooth and faster to get an optimal interval for Guinier analysis.

d In a simplified formation for illustration, the objective function and its optimization problem is defined as follows, using R_g's from two divided sub-intervals (See Fig. 6-1).

$$\text{minimize: } F(q_0, q_2) = \left(R_g(q_0, q_1) - \widehat{R_g}\right)^2 + \left(R_g(q_1, q_2) - \widehat{R_g}\right)^2, \quad (6\text{-}1)$$

where

q_0: left end point of Guinier interval,

$q_1 = q_0 + \Delta q$: intermediate point to split into sub-intervals,

Δq: sufficiently small constant value to efficiently evaluate the head slope,

q_2: right end point of Guinier interval,

such that $q_1 = q_0 + \Delta q \le q_2 \le \hat{Q}$,

$\hat{Q}$: approximate end point obtained as (4-2),

$R_g(p, q)$: R_g calculated linearly in a interval $[p, q]$ in the spline,

$\widehat{R_g}$: R_g obtained from Guinier · Porod fitting.

e Although the above objective function is bivariate, i.e. of variables q_0 and q_2, the main target of optimization is q_0.[4]

f Splitting into these sub-intervals makes the optimization more efficient because the deviation from linearity is more sharply evaluated for narrower intervals. (mainly in the left sub-interval)

g The real objective function implemented in the program is slightly more complex with two more factors to minimize, one of which is size factor of the interval, the other of which is the starting factor of the interval. See the formulation (4-3).

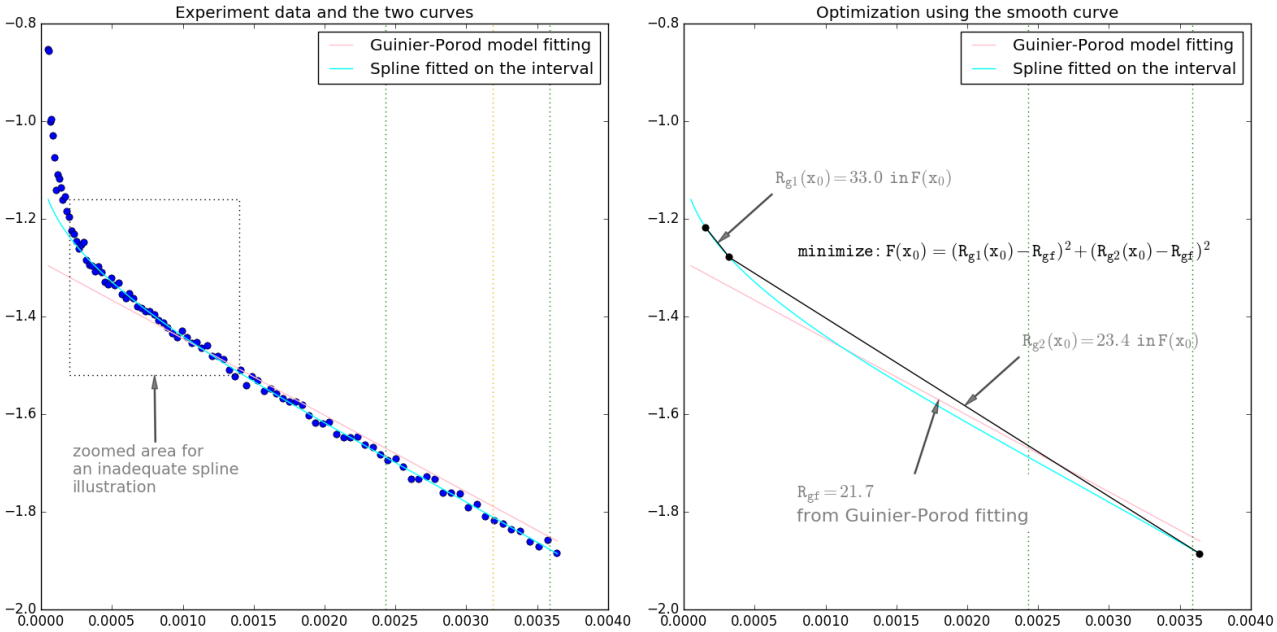
h The size factor is smaller for wider intervals, and the starting factor is smaller for earlier-starting intervals.

---

[3] The intercept is not important in this stage.

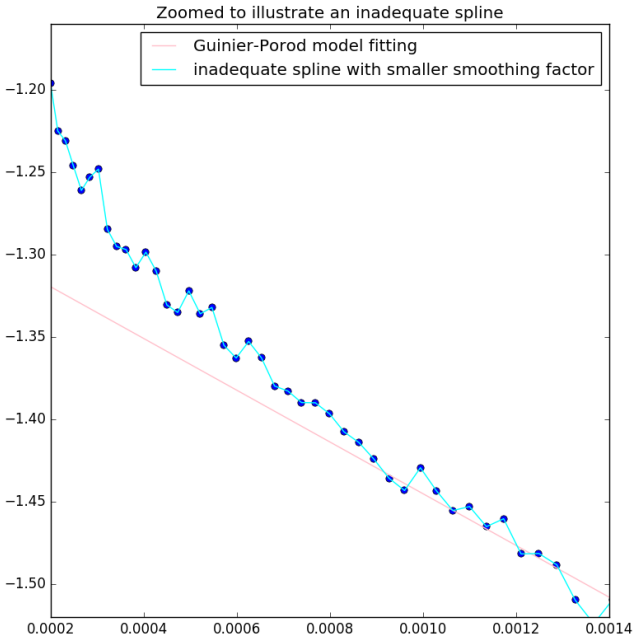
[4] In fact, it is observed that making it univariate by letting $q_2 = \hat{Q}$ (constant) would not change the optimization result significantly.

Fig. 6-1 Optimization using smooth curves



i It is essential that the smoothing factor for the spline be given sufficiently loose so that the objective function behaves moderately to correctly reflect the macroscopic trend of the data. See Fig. 6-2 for an inadequate spline example to understand the implication.

Fig. 6-2 Example of an inadequate spline



6.6 Estimation of the Rg

6.7 Estimation of the Errors

6.8 Calculation of the Quality Index

7 Implementation Details

Appendix

A Minimal History of SAXS Data Modeling

a In this appendix, we try to summarize minimal amount of, possibly biased, information to verify the adequacy of our solution scheme.

A.1 Guinier Law – André Guinier, 1939

a André Guinier pioneered in the small-angle techniques discovering the fundamental law expressed as follows.

$$I(Q) \approx G \exp\left(\frac{-Q^2 {R_g}^2}{3}\right) \quad \text{for low } Q \tag{A-1}$$

b In logarithmic scaling, it appears as an equivalent linear relation between $\mathrm{Ln}(I)$ and Q^2.

$$\ln(I) \approx \ln(I_0) - \frac{{R_g}^2}{3} Q^2 \quad \text{for low } Q \tag{A-2}$$

c The applicable range of Q of this approximation is later established as $Q \cdot Rg < 1.3$. See Appendix C for details.

d The Guinier Law is the most fundamental in the sense that it gives the size information (R_g) regardless of the shapes of objects.

A.2 Porod Law – Debye & Bueche, 1949; Günther Porod, 1951

a Debye & Bueche and Günther Porod independently made the next fundamental progress discovering another fundamental law for higher Q ranges.

$$I(Q) = \frac{D}{Q^4} \quad \text{for high } Q \tag{A-3}$$

b At this stage, the invariance suggested by the following (A-4) over the various shapes of objects had a significant importance.

$$\frac{D}{I_{total}} \propto \frac{S}{V}$$

$$I_{total} = \int_{q=0}^{\infty} q^2 I(q) dq \qquad \text{(A-4)}$$

c Give an example of application of the law.

T.B.D.

d However, from the viewpoint of verifying the adequacy of our program structure, the dependence (or variation) of the exponent from the different shapes clarified later is more important as described below.

A.3 Generalized Porod Law for fractal systems – Bale & Schmidt, 1984

a After the popularization by Mandelbrot, recognizing fractals in objects advanced in many fields including SAXS, and the Porod Law was generalized as below, allowing the exponent to vary according to the fractal dimension of the objects.

$$I(Q) = \frac{D}{Q^d} \quad \text{for larger } Q \qquad \text{(A-5)}$$

b This generalization clearly suggested the new role of the exponent (d) as a shape information indicator.

A.4 Unified Model – 1995, Greg Beaucage

a Guinier Law and generalized Porod Law had been used independently before Greg Beaucage introduced a unified model, concatenating the above two classical laws smoothly using the error function (erf).

$$I(Q) = G\exp\left(\frac{-Q^2{R_g}^2}{3}\right) + \frac{B}{Q^d}\left[\mathrm{erf}\left(\frac{QR_g}{6^{1/2}}\right)\right]^{3d} \tag{A-6}$$

b Observe the technical idea of using error function by its properties as follows.

$$\begin{aligned}\mathrm{erf}(x) &\to 0 \text{ as } x \to 0\\ \mathrm{erf}(x) &\to 1 \text{ as } x \to \infty\end{aligned} \tag{A-7}$$

c The formula, modified later by Boualem Hammouda in [2] so as to be applicable not only for *d=2*, is given below.

$$I(Q) = G\exp\left(\frac{-Q^2{R_g}^2}{3}\right) + \frac{C}{Q^d}\left[\mathrm{erf}\left(\frac{QR_g}{6^{1/2}}\right)\right]^{3d} \tag{A-8}$$

$$C = \frac{Gd}{{R_g}^d}\left[\frac{6d^2}{(2+d)(2+2d)}\right]^{\frac{d}{2}}\Gamma\left(\frac{d}{2}\right) \tag{A-9}$$

d This unification opened the possibility of global fitting to the data model.

A.5 Guinier-Porod Model – 2010, Boualem Hammouda

a Boualem Hammouda introduced Guinier-Porod model in 2010, which concatenates the classical two laws differently against the previous Unified (Beaucage) model.

b Instead of gradually transitioning by the error function (erf), this model separates the regions clearly imposing instead the continuity and smoothness conditions at the boundary, which is given in his paper (as explained in detail in the next Appendix B).

$$I(Q) = G \exp\left(\frac{-Q^2 {R_g}^2}{3}\right) \text{ for } Q \le Q_1, \qquad I(Q) = \frac{D}{Q^d} \text{ for } Q \ge Q_1 \tag{A-10}$$

$$Q_1 = \frac{1}{R_g}\left(\frac{3d}{2}\right)^{1/2}, \qquad D = G \exp\left(\frac{-{Q_1}^2 {R_g}^2}{3}\right) {Q_1}^d = G \exp\left(-\frac{d}{2}\right)\left(\frac{3d}{2}\right)^{d/2} \frac{1}{{R_g}^d} \tag{A-11}$$

c To understand the relation between the Guinier law application restriction ($QR_g < 1.3$) and this automatically determined boundary (Q_1), transform the Q_1 definition formula in (A-11) into the following.

$$Q_1 R_g = \left(\frac{3d}{2}\right)^{1/2} \tag{A-12}$$

d And observe the $Q_1 R_g$ values varying the Porod exponent (d) from 1 through 4.

Tab. A-1 $Q_1 R_g$ values corresponding to different Porod exponent[5]

Porod exponent	$Q_1 R_g$	Fractal Classification	Applicable samples
1	1.22	mass fractals	a stiff rod (or thin cylinder)
5/3	1.58		'fully swollen' chains (in a good solvent)
2	1.73		Gaussian polymer chains or two-dimensional structures (such as lamellae or platelets)
3	2.12	surface fractals	particles with very rough surfaces, 'collapsed' polymer chains (in a bad solvent)
4	2.45		particles with smooth surfaces

[5] This table is made by simply summarizing his 2010 paper, except the $Q_1 R_g$ columns added by the authors.

e The observation leads to an understanding that the Guinier law application restriction corresponds to particles with smaller values of exponent around one and the restriction might be relaxed up to $\mathrm{Q}R_g < 2.45$ for particles with larger values of exponent.[6]

f Jan Ilavsky calls for his user's attention in his Irena Manual that "However, there is a problem here. For systems, which do not adhere to Guinier-Porod model assumptions, cannot be modeled by the Guinier-Porod model at all. For example, these would be hierarchical fractal systems, particulate systems with broad size distribution, etc."

g Boualem Hammouda presented two more variations of the model in the above-mentioned paper. One is a generalized form, based on the works by other pioneers, for non-spherical scattering objects which adds an 's' parameter with "$3 - s$" suggesting the dimensionality (1D, 2D or 3D) of the objects. The other is a more complex three-region model, which adds two dimensionality parameters, s_1 and s_2 respectively to describe the first two generalized Guinier regions with the remaining third Porod region. See the paper for details.[7]

[6] In spite of this relaxation possibility, the program usually conforms to the constraint such that $\mathrm{Q}R_g < 1.3$. See appendix C for details. The constraint can be relaxed by an optional setting.

[7] The current program only utilizes the simplest form of the model for simplicity and speed.

B Derivation of the two-region Guinier-Porod Model

a It is a trivial chain of calculus, but we show the derivation in detail here since we believe it is important for understanding the implication of its fitting.

b The simplest two-region Guinier-Porod model can be thought of as a natural consequence of the two classical models in the following sense.

c It is derived simply from the following two equations requiring respectively the continuity (B-1) and smoothness (B-2), i.e. having the same slope, at the boundary Q_1.

$$I(Q) = G \exp\left(\frac{-Q^2 {R_g}^2}{3}\right) = \frac{D}{Q^d} \quad \text{at } Q = Q_1 \tag{B-1}$$

$$\frac{dI}{dQ} = G \exp\left(\frac{-Q^2 {R_g}^2}{3}\right)\left(\frac{-2Q{R_g}^2}{3}\right) = -DdQ^{-d-1} \quad \text{at } Q = Q_1 \tag{B-2}$$

d The Q_1 definition formula is easily derived by eliminating the shaded Guinier term from (B-1) and (B-2) as follows.

$$\frac{D}{Q^d}\left(\frac{-2Q{R_g}^2}{3}\right) = -DdQ^{-d-1} \quad \text{at } Q = Q_1 \tag{B-3}$$

$$Q^{-d+1}\left(\frac{-2{R_g}^2}{3}\right) = -dQ^{-d-1} \quad \text{at } Q = Q_1 \tag{B-4}$$

$$Q^2 = \frac{3d}{2{R_g}^2} \quad \text{at } Q = Q_1 \tag{B-5}$$

$$Q_1 = \frac{1}{R_g}\left(\frac{3d}{2}\right)^{1/2} \tag{B-6}$$

e The D definition formula is obtained from (B-1), substituting Q_1 by its shaded definition above and reducing as follows.

$$D = G \exp\left(\frac{-Q^2 {R_g}^2}{3}\right) Q^d \quad at\ Q = Q_1 \tag{B-7}$$

$$D = G \exp\left(\frac{-\left(\frac{1}{R_g}\left(\frac{3d}{2}\right)^{1/2}\right)^2 {R_g}^2}{3}\right)\left(\frac{1}{R_g}\left(\frac{3d}{2}\right)^{1/2}\right)^d = G \exp\left(-\frac{d}{2}\right)\left(\frac{3d}{2}\right)^{d/2}\frac{1}{{R_g}^d} \tag{B-8}$$

C Notes on the upper limit of Guinier Approximation

a Guinier Approximation is usually considered to be valid for low-q ranges such that $qR_g < 1.3$.

b However, according to the Guinier-Porod model, there is a possibility that the applicable range be extended as far as $qR_g < 2.45$ for some kind of particles (as with smooth surfaces).

c We will try here to clarify this gap.

d Guinier Approximation is derived theoretically by omitting higher-than-4th power terms from the following Maclaurin series.

$$\frac{\sin(qr)}{qr} = 1 - \frac{q^2r^2}{3!} + \frac{q^4r^4}{5!} - \cdots \tag{C-1}$$

$$I(q) = 4\pi \int_0^D \gamma(r) \frac{\sin(qr)}{qr} r^2 dr \tag{C-2}$$

e Skipping the details, Guinier Approximation corresponds to the shaded first two terms in the series.

f According to L.A. Feigin and D.I. Svergun 1987 (section 3.3.1), deviations caused by the omission can be estimated by the following formula, which is valid to an accuracy of terms proportional to q^6.

$$\Delta(q) \approx \Delta M \mu^4 \left(qR_g\right)^4$$

$$\text{where } \Delta M = \frac{3M_6M_2 - 5{M_4}^2}{360{M_2}^2}, \tag{C-3}$$

M_k : normalized kth moment of function $\gamma(t)$,

$\mu = D/R_g, D$: the largest dimension of in a particle

g Although ΔM is said to vary between 1×10^{-4} and 2×10^{-4}, we first ignore the variation of ΔM and see a possibility of relaxing (or tightening) the qR_g upper limit, from the reciprocal relationship of μ and qR_g when keeping the deviation $\Delta(q)$ to a constant level, as in Tab. C-1.

Tab. C-1 possibility of relaxing the q*Rg upper limit

particle shape	μ	qR_g level	remarks
solid sphere	2.46	1.30	setting this to the base level
infinitely thin disk	2.28	1.40	value to keep the deviation $\Delta(q)$ to the same level
infinitely long rod	3.46	0.92	same as above
ellipsoid of rotation with c/a=2	3.56	0.90	same as above

h In order to make the discussion more accurate, we have to get the values M_k for various shapes and the relationship between ΔM and μ. (T.B.D.)

i However, without further study, since $(2 \times 10^{-4})^{1/4} \approx 1.2$, it seems that the qR_g upper limit should be lower than $1.68 = 1.4 \times 1.2$ at the most and can be as low as $0.75 = 0.90/1.2$ at the least.

D Notes on Statistics

D.1 Weighted Least Squares

Tab. D-1 Summary of Weighted Least Squares

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix},\ \boldsymbol{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix},\ \boldsymbol{\beta} = [\beta_1 \quad \beta_2] \tag{D-1}$$

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{D-2}$$

$$\text{minimize: } (\boldsymbol{y} - \boldsymbol{X\beta})^T \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{X\beta}) \tag{D-3}$$

$$\mathbf{W} = \begin{bmatrix} \mathrm{w}_1 & & & \\ & \mathrm{w}_2 & & \\ & & \ddots & \\ & & & \mathrm{w}_N \end{bmatrix},\ \mathrm{w_i} = \frac{1}{{\sigma_\mathrm{i}}^2} \tag{D-4}$$

$$(\boldsymbol{X^T W X})\widehat{\boldsymbol{\beta}} = \boldsymbol{X^T W y} \tag{D-5}$$

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X^T W X})^{-1} \boldsymbol{X^T W y} \tag{D-6}$$

$$\boldsymbol{X^T W X} = \begin{bmatrix} \sum w_i & \sum \mathrm{w_i x_i} \\ \sum \mathrm{w_i x_i} & \sum \mathrm{w_i {x_i}^2} \end{bmatrix},\ \boldsymbol{X^T W y} = \begin{bmatrix} \sum w_i y_i \\ \sum w_i x_i y_i \end{bmatrix} \tag{D-7}$$

$$(\boldsymbol{X^T W X})^{-1} = \frac{1}{\Delta} \begin{bmatrix} \sum \mathrm{w_i {x_i}^2} & -\sum \mathrm{w_i x_i} \\ -\sum \mathrm{w_i x_i} & \sum w_i \end{bmatrix},\ \Delta = \sum w_i \sum \mathrm{w_i {x_i}^2} - \left(\sum \mathrm{w_i x_i}\right)^2 \tag{D-8}$$

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X^T W X})^{-1} \boldsymbol{X^T W y} = \frac{1}{\Delta} \begin{bmatrix} \sum \mathrm{w_i {x_i}^2} \sum w_i y_i - \sum \mathrm{w_i x_i} \sum w_i x_i y_i \\ -\sum \mathrm{w_i x_i} \sum w_i y_i + \sum w_i \sum w_i x_i y_i \end{bmatrix} \tag{D-9}$$

$$\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{e} = \begin{bmatrix} \mathrm{e}_1 \\ \mathrm{e}_2 \\ \vdots \\ \mathrm{e}_N \end{bmatrix} \tag{D-10}$$

$$\mathrm{S} = \left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)^T \boldsymbol{W} \left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right) = \sum \mathrm{w_i {e_i}^2} = \sum \frac{{\mathrm{e_i}}^2}{{\sigma_\mathrm{i}}^2} \tag{D-11}$$

$$\sigma^2 = \frac{\mathrm{S}}{(\mathrm{N} - 2)} \tag{D-12}$$

$$\mathrm{Var}\left(\widehat{\boldsymbol{\beta}}\right) = \sigma^2 (\boldsymbol{X^T W X})^{-1} \tag{D-13}$$

D.2 Propagation of Errors

E Python Modules and Tools

F References

[1] Hammouda, B. (2010a). J. Appl. Cryst.43, 716-719. A new Guinier-Porod model.

[2] Hammouda, B. (2010b). J. Appl. Cryst.43, 1474-1478. Analysis of the Beaucage model.

[3] Feigin, L. A. & Svergun, D. I. (1987). Structure Analysis by Small Angle X-ray and Neutron Scattering. New York: Plenum Press.