DISSERTATION

STATISTICAL INNOVATIONS FOR ESTIMATING SHAPE CHARACTERISTICS OF BIOLOGICAL MACROMOLECULES IN SOLUTION USING SMALL-ANGLE X-RAY SCATTERING DATA

Submitted by

Cody Alsaker

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2016

Doctoral Committee:

Advisor: F. Jay Breidt

Don Estep
Piotr Kokoszka
Karolin Luger

ABSTRACT

STATISTICAL INNOVATIONS FOR ESTIMATING SHAPE CHARACTERISTICS OF BIOLOGICAL MACROMOLECULES IN SOLUTION USING SMALL-ANGLE X-RAY SCATTERING DATA

Small-angle X-ray scattering (SAXS) is a technique that yields low-resolution images of biological macromolecules by exposing a solution containing the molecule to a powerful X-ray beam. The beam scatters when it interacts with the molecule. The intensity of the scattered beam is recorded on a detector plate at various scattering angles, and contains information on structural characteristics of the molecule in solution. In particular, the radius of gyration (R_g) for a molecule, which is a measure of the spread of its mass, can be estimated from the lowest scattering angles of SAXS data using a regression technique known as Guinier analysis. The analysis requires specification of a range or "window" of scattering angles over which the regression relationship holds. We have thus developed methodology and supporting asymptotic theory for selection of an optimal window, minimum mean square error estimation of the radius of gyration, and estimation of its variance. The theory and methodology are developed using a local polynomial model with autoregressive errors. Simulation studies confirm the quality of the asymptotic approximations and the superior performance of the proposed methodology relative to the accepted standard. We show that the algorithm is applicable to data acquired from proteins, nucleic acids and their complexes, and we demonstrate with examples that the algorithm improves the ability to test biological hypotheses.

The radius of gyration is a normalized second moment of the pairwise distance distribution $p(r)$, which describes the relative frequency of inter-atomic distances in the structure of the molecule. By extending the theory to fourth moments, we show that a new parameter ψ can be calculated theoretically from $p(r)$ and estimated from experimental SAXS data, using a method that extends Guinier's R_g estimation procedure. This new parameter yields an enhanced ability to use intensity data to distinguish between two molecules with different but similar R_g values. Analysis of existing structures in the protein data bank (PDB) shows that the theoretical ψ values relate closely to the aspect ratio of a molecular structure. The combined values for R_g and ψ acquired from experimental data provide estimates for the dimensions and associated uncertainties for a standard geometric shape, representing the particle in solution. We have chosen the cylinder as the standard shape and show that a simple, automated procedure gives a cylindrical estimate of a particle of interest. The cylindrical estimate in turn yields a good first approximation to the maximum inter-atomic distance in a molecule, D_{max}, an important parameter in shape reconstruction.

As with estimation of R_g, estimation of ψ requires specification of a window of angles over which to conduct the higher-order Guinier analysis. We again employ a local polynomial model with autoregressive errors to derive methodology and supporting asymptotic theory for selection of an optimal window, minimum mean square error estimation of the aspect ratio, and estimation of its variance.

Recent advances in SAXS data collection and more comprehensive data comparisons have resulted in a great need for automated scripts that analyze SAXS data. Our procedures to estimate R_g and ψ can be automated easily and can thus be used for large suites of SAXS data under various experimental conditions, in an objective and reproducible manner. The

new methods are applied to 357 SAXS intensity curves arising from a study on the wild type nucleosome core particle and its mutants and their behavior under different experimental conditions. The resulting $\widehat{R}_g^2$ values constitute a dataset which is then analyzed to account for the complex dependence structure induced by the experimental protocols. The analysis yields powerful scientific inferences and insight into better design of SAXS experiments.

Finally, we consider a measurement error problem relevant to the estimation of the radius of gyration. In a SAXS experiment, it is standard to obtain intensity curves at different concentrations of the molecule in solution. Concentration-by-angle interactions may be present in such data, and analysis is complicated by the fact that actual concentration levels are unknown, but are measured with some error. We therefore propose a model and estimation procedure that allows estimation of true concentration ratios and concentration-by-angle interactions, without requiring any information about concentration other than that contained in the SAXS data.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

OVERVIEW

Determining the structure of biological macromolecules (proteins, nucleic acids, and their complexes) is fundamental to determining their function. X-ray crystallography can provide high-resolution structural information, to the level of the atomic structure of the molecule. But X-ray crystallography requires crystallization of the molecule, and crystallization recipes are determined empirically and can be resource intensive. Further, many molecules resist crystallization (see [2]). As an alternative to high-resolution methods, small-angle X-ray scattering (SAXS) is an experimentally simple technique to acquire low-resolution information about the structure of biological macromolecules. SAXS is relatively inexpensive and fast and works much more generally than crystallography.

Figure 1.1 schematically depicts a SAXS experiment and the resulting output. The sample of the molecule in solution is exposed to a high-intensity X-ray beam, which scatters when interacting with the sample. The scattered X-ray is recorded by a two-dimensional detector plate, which measures the intensity of the scattered pattern at different angles. Scattering intensity at angles near zero is recorded near the center of the plate, and scattering intensity at progressively larger angles is measured along concentric circles of increasing radii. Angles very near zero are not recorded, as they correspond to the direct X-ray beam hitting a "beam stop" (often, a lead plug) in the center of the detector. The two-dimensional scattering intensity information is background-corrected (by subtracting an image of only solute, no molecules) and then condensed into a one-dimensional curve via "radial averaging;" that is, averaging the intensity on concentric annuli determined by a grid of angles. In Figure 1.1, averaging along the annulus depicted by the circle of radius $\|\boldsymbol{q}\|$ results in the average

intensity value plotted on a log scale in the right-hand-side figure, at the angle $s \propto \|\boldsymbol{q}\|$ indicated by the vertical reference line. The one-dimensional log-intensity curve contains shape and size properties for the molecule in solution. In this particular example, the log-intensity data correspond to the molecule myoglobin, with known atomic structure (protein data bank [3] entry 1WLA [4]) depicted in the upper right-hand corner of the log-intensity plot. For a novel molecule, this structure would be unknown, and SAXS data would be used to determine some low-resolution structural characteristics of the molecule in solution. The use of small angle scattering in structural biology is reviewed extensively elsewhere [5–7].

Recent advances in SAXS data collection and more comprehensive data comparisons have resulted in a great need for automated scripts that analyze SAXS data [8, 9]. We have thus developed a statistically-rigorous algorithm that automatically estimates the radius of gyration for a molecule, which is a measure of the spread of its mass, from the lowest scattering angles of SAXS data.

A useful summary of the atomic structure of a molecule is given by its pairwise distance distribution, which for a molecule with A atoms at coordinates $\{\boldsymbol{a}_i\}_{i=1}^{A}$ is

$$p(r) = \frac{\#\{(i,j) \in \{1,\ldots,A\}^2 : \|\boldsymbol{a}_i - \boldsymbol{a}_j\| = r\}}{A^2}, \quad 0 \le r \le D_{max},$$

with D_{max} the maximum pairwise distance. As A is typically on the order of 100–1000, we follow standard practice and use the continuous version of $p(r)$ in what follows.

The squared radius of gyration for a molecule is

$$R_g^2 = \frac{\int_0^{D_{max}} r^2 p(r)\, dr}{2\int_0^{D_{max}} p(r)\, dr} = \frac{2\pi \int_0^{D_{max}} r^2 p(r)\, dr}{\mathcal{I}(0)}, \tag{1}$$

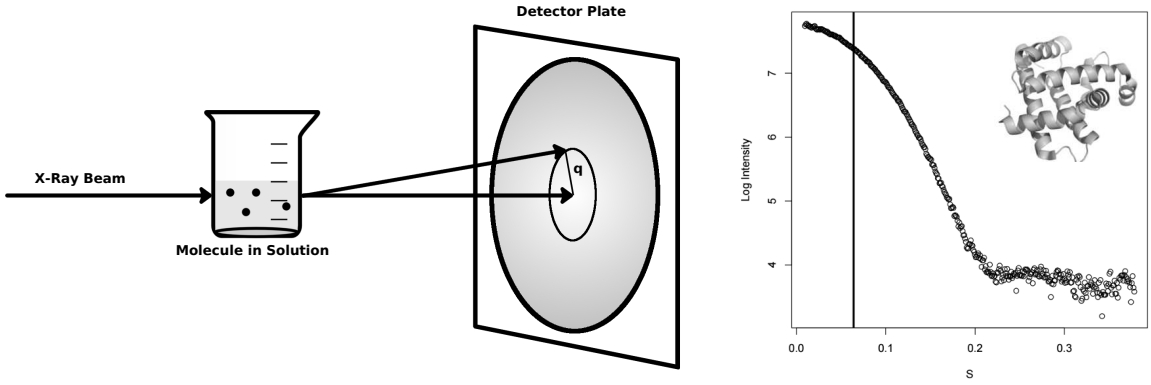


FIGURE 1.1. Schematic depiction of a SAXS experiment and resulting log-intensity data. The sample of the molecule in solution is exposed to a high-intensity X-ray beam, which scatters when interacting with the sample. The scattered pattern is recorded by a two-dimensional detector plate, which measures the intensity at different angles. In the example shown, the scattered beam intersects the detector at coordinate vector $\boldsymbol{q}$, with the origin at the center of the detector. The two-dimensional intensity data are reduced to one-dimensional data by first subtracting a reference image (not shown) and then computing an average intensity for each concentric annulus along a sequence of increasing angles. Averaging along the annulus depicted by the circle of radius $\|\boldsymbol{q}\|$ results in the average intensity value plotted on a log scale in the right-hand-side figure, at the angle $s \propto \|\boldsymbol{q}\|$ indicated by the vertical reference line. Log-intensity data in this example correspond to the molecule myoglobin, with known atomic structure depicted in the upper right-hand corner of the log-intensity plot.

where $\mathcal{I}(s)$ denotes the theoretical scattering intensity at scattering angle $s \geq 0$. Let $I(s)$ denote the corresponding empirical intensity from a SAXS experiment. From (1), the radius of gyration R_g is analogous to the standard deviation of a probability density, describing the spread of mass in a molecular model.

While $p(r)$ can be estimated using empirical intensity data from a SAXS experiment, such estimation requires modeling assumptions and regularization techniques to effect an inverse Fourier transformation (e.g., [10–13]). By contrast, R_g is an example of low-resolution structural information that can be estimated directly from SAXS data without modeling the

molecular structure. For the theoretical log-intensity, [14] derived the approximation

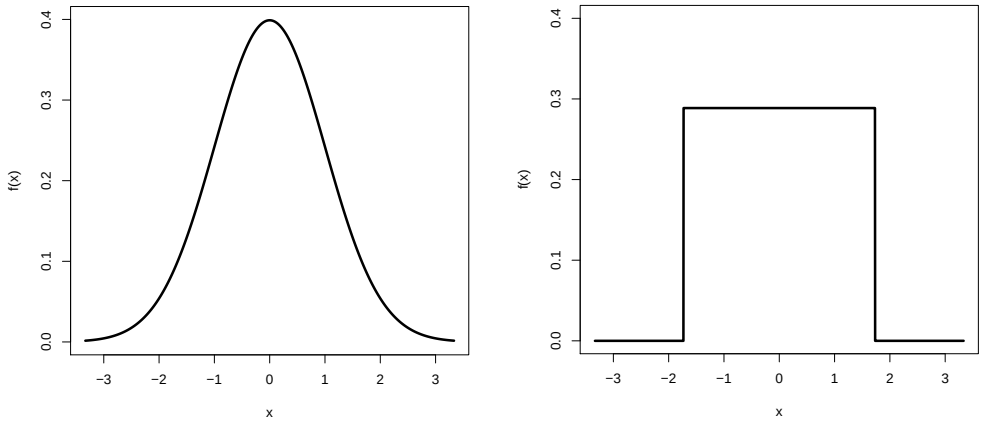
$$\ln \mathcal{I}(s) = \ln \mathcal{I}(0) - \frac{1}{3} R_g^2 s^2 + \mathrm{O}\left(s^4\right). \tag{2}$$

Let Y_i denote the empirical log-intensity at angle s_i from a SAXS experiment. Assuming Y_i is unbiased for $\ln \mathcal{I}(s_i)$ leads to the now-standard method of *Guinier analysis* [e.g., 15, p. 71] in which a data window of near-zero angles $s_1, \ldots, s_n$ is chosen such that the relationship between squared angle s_i^2 and empirical log-intensity $Y_i = \ln I(s_i)$ appears linear. The regression model

$$Y_i = \ln \mathcal{I}(0) - \frac{1}{3} R_g^2 s_i^2 + \varepsilon_i = \beta_0 + \beta_2 s_i^2 + \varepsilon_i \tag{3}$$

is then fitted, and $\widehat{R}_g^2 = -3\widehat{\beta}_2$ is used as the estimate of R_g^2. See [16] for an analogous problem, in which the memory parameter of a long-memory time series is estimated via regression of the log-periodogram on a function of the Fourier frequencies. Based on simulated data from idealized models of particles, Guinier recommended the now classical cutoff value s_n^{class}, obtained by iteratively adjusting the data window to achieve $s_n^{class} \widehat{R}_g < 1.3$ [17, p. 128]. Programs exist that automatically determine $\widehat{R}_g$, but these programs are constrained by the classical 1.3 cutoff value [18].

The radius of gyration is analogous to the second moment of a molecule, so the value of R_g^2 contains information regarding a molecule's shape and size. Therefore, it is natural to consider the fourth moment of a molecule in order to further distinguish subtle differences in molecules' shapes and sizes. This idea relates to the moments of probability distributions. Consider the two distributions in Figure 1.2. These two distributions both have the same first and second moment; however, they have very different shapes. Their fourth moment



(A) Normal (0,1) distribution. (B) Uniform $(-\sqrt{3}, \sqrt{3})$ distribution.

FIGURE 1.2. Normal (0,1) distribution (left) and Uniform $(-\sqrt{3}, \sqrt{3})$ distribution (right). Each distribution has mean zero and variance one. However, the fourth moment of the normal distribution is 3 and the fourth moment of the uniform distribution is 1.8.

values can help distinguish their shapes. The fourth moment for the normal distribution on the left is 3 and the fourth moment for the uniform distribution on the right is 1.8. Hence, we could use SAXS data in order to differentiate between two molecules with similar R_g values but different fourth moment values, provided we can estimate the higher-order moments. We show that classical Guinier analysis can be extended to higher-order Guinier analysis for estimation of higher-order moments from SAXS data.

Many SAXS experiments involve a suite of varying experimental factors (concentration level, exposure time, etc.) with multiple replicate data for each factor. These experiments can yield a large number of SAXS curves, which need to be analyzed in an efficient manner. If parameter estimates for all the replicate data under varying experimental conditions are determined, then scientific inference can be performed.

It turns out that the concentration level of a molecule in solution for a SAXS experiment is a particularly important factor. The concentration level is often assumed to not influence

the estimate of R_g, but this is not always the case. Frequently, a concentration by angle interaction is present. Another issue is that the prescribed nominal concentration level may not equal the actual concentration level of the molecule in solution. Therefore, the ability to estimate the concentration level using SAXS data is useful.

The contributions of this dissertation are summarized here briefly. First, a natural alternative to the classical $s_n^{class}\widehat{R}_g < 1.3$ rule-of-thumb is to use statistical methods to optimize the choice of cutoff value with respect to mean squared error (MSE), trading off the increased bias of a larger cutoff value (due to breakdown of the Guinier quadratic approximation (2)) with the decreased variance due to larger sample size. This minimum MSE approach requires estimation of the bias and variance, accounting for the fact that the Guinier approximation (2) holds only for small angles. In particular, estimating the bias requires allowance for higher-order terms in the Guinier approximation. We therefore develop an automated procedure to estimate R_g and its variance while accounting for the autocovariance structure of the empirical intensity curve (see [19]).

It turns out that outlying log-intensities may be present among the lowest scattering angles. These smallest angles are subject to the greatest intensity of the X-ray beam and are adjacent to the central beam stop, both of which may lead to unusual intensity values. It is currently standard practice for the operator to perform outlier detection and removal manually. We develop an automated statistical procedure to detect such outliers by adapting the standard DFBETAS criterion (e.g., [20], §10.4) in estimation of R_g^2 under model (3).

The next contribution relates to determining the fourth moment of a molecule. Hence, we define the new quantity

$$M^4 = \frac{\int_0^{D_{max}} r^4 p(r)\, dr}{2\pi \int_0^{D_{max}} p(r)\, dr}. \tag{4}$$

Just as R_g^2 is analogous to the second moment of a probability distribution, M^4 is the analog of the fourth moment of a distribution. Very roughly, in statistics the fourth moment of a distribution provides information about how much area is contained in the tails of the distribution. Likewise, M^4 provides information about how much mass is contained in the regions of a molecule farthest from the center of mass. For example, a rod-shaped molecule has a larger M^4 value than a spherical molecule, assuming both molecules have the same R_g value.

We wish to use these quantities to differentiate molecules based purely on shape, but both R_g^2 and M^4 are dependent on molecular size. Therefore, we define a new dimensionless ratio ψ given by

$$\psi = \frac{M^4}{R_g^4},$$

which contains information concerning molecular shape but is independent of size. Furthermore, ψ can be estimated from experimental SAXS data using an extension of Guinier analysis.

To relate the SAXS data to both R_g and the quantity M^4, we extend (2) to include an extra term, resulting in a higher-order (and more accurate for small s) approximation:

$$\ln \mathcal{I}(s) = \ln \mathcal{I}(0) - \frac{1}{3} R_g^2 s^2 + \left(\frac{1}{60} M^4 - \frac{1}{18} R_g^4 \right) s^4 + \mathrm{O}\left(s^6\right). \tag{5}$$

Using this equation, R_g and M^4 can be estimated directly from experimental SAXS data for a molecule; the ratio ψ can then be obtained easily. To determine the window of data to estimate R_g and M^4 in (5), we minimize the MSE of $\widehat{\psi}$ while accounting for the autocorrelation in the data. The parameter ψ provides a way to obtain a first approximation of the shape of a molecule in solution from the log intensity curve. Furthermore, the parameter ψ, in conjunction with R_g, additionally provides the means to compute the maximum pairwise distance D_{max}. Another potential benefit of both these innovations is the enhanced ability to use intensity data to distinguish between two molecules with different but similar R_g values.

Our procedures to estimate R_g and ψ are automatic and can thus be used for a suite of SAXS data under various experimental conditions in an objective and reproducible manner. The new methods are applied to 357 SAXS intensity curves arising from a study on the wild type nucleosome core particle and its mutants and their behavior under different experimental conditions. The resulting $\widehat{R}_g^2$ values constitute a dataset which is then analyzed to account for the complex dependence structure induced by the experimental protocols. The analysis yields powerful scientific inferences and insight into better design of SAXS experiments.

Finally, we propose a model and estimation procedure to determine concentration ratios of molecules in solution while accounting for a concentration by angle interaction effect. This model does not require any information about concentration other than that contained in the SAXS data. This model is validated with empirical data for which we have external measurements of concentrations.

The overall structure of this dissertation is as follows. Chapter 2 and Chapter 3 are each articles that have been submitted for publication. Therefore, there is some repeated information in each of these chapters. Both of these chapters pertain to determining an optimal

estimate of R_g and its variance. Chapter 2 contains the theoretical development for the estimation of R_g along with extensive simulations and applications of the estimation procedure. Chapter 3 is focused on the biological aspects of $\widehat{R}_g$, so this chapter does not contain the full theoretical development of the R_g estimation procedure. Chapter 4 describes the novel molecular parameter ψ that can be estimated from SAXS curves. The full theoretical development for the estimation of this parameter with applications to experimental data is given in Chapter 5. Chapter 6 develops a model that can estimate the concentration ratio for a set of molecules in solution and check for a concentration by angle interaction. Furthermore, applications to experimental data are also given. Finally, the Appendix contains simulations for the novel outlier detection procedure and instructions for implementing the R_g estimation program.

CHAPTER 2

MINIMUM MEAN SQUARED ERROR ESTIMATION OF THE RADIUS OF GYRATION IN SMALL-ANGLE X-RAY SCATTERING EXPERIMENTS

2.1. Introduction

2.1.1. Small-Angle X-Ray Scattering Experiments. Determining the structure of biological macromolecules (proteins, nucleic acids, and their complexes) is fundamental to determining their function. X-ray crystallography can provide high-resolution structural information, to the level of the atomic structure of the molecule. But X-ray crystallography requires crystallization of the molecule, and crystallization recipes are determined empirically and can be resource intensive. Further, many molecules resist crystallization (see [2]). As an alternative to high-resolution methods, small-angle X-ray scattering (SAXS) is an experimentally simple technique to acquire low-resolution information about the structure of biological macromolecules. SAXS is relatively inexpensive and fast and works much more generally than crystallography.

Figure 2.1 schematically depicts a SAXS experiment and the resulting output. The sample of the molecule in solution is exposed to a high-intensity X-ray beam, which scatters when interacting with the sample. The scattered X-ray is recorded by a two-dimensional detector plate, which measures the intensity of the scattered pattern at different angles. Scattering intensity at angles near zero is recorded near the center of the plate, and scattering intensity at progressively larger angles is measured along concentric circles of increasing radii. Angles very near zero are not recorded, as they correspond to the direct X-ray beam hitting a "beam stop" (often, a lead plug) in the center of the detector. The two-dimensional scattering

intensity information is background-corrected (by subtracting an image of only solute, no molecules) and then condensed into a one-dimensional curve via "radial averaging"; that is, averaging the intensity on concentric annuli determined by a grid of angles. In Figure 2.1, averaging along the annulus depicted by the circle of radius $\|\boldsymbol{q}\|$ results in the average intensity value plotted on a log scale in the right-hand-side figure, at the angle $s \propto \|\boldsymbol{q}\|$ indicated by the vertical reference line. The one-dimensional log-intensity curve contains shape and size properties for the molecule in solution. In this particular example, the log-intensity data correspond to the molecule myoglobin, with known atomic structure (protein data bank ([3]) entry 1WLA ([4])) depicted in the upper right-hand corner of the log-intensity plot. For a novel molecule, this structure would be unknown, and SAXS data would be used to determine some low-resolution structural characteristics of the molecule in solution. The use of small angle scattering in structural biology is reviewed extensively elsewhere ([5–7]).

Recent advances in SAXS data collection and more comprehensive data comparisons have resulted in a great need for automated scripts that analyze SAXS data ([8, 9]). We have thus developed a statistically-rigorous algorithm that automatically estimates the radius of gyration for a molecule, which is a measure of the spread of its mass, from the lowest scattering angles of SAXS data.

2.1.2. Guinier Analysis. A useful summary of the atomic structure of a molecule is given by its pairwise distance distribution, which for a molecule with A atoms at coordinates $\{\boldsymbol{a}_i\}_{i=1}^{A}$ is

$$p(r) = \frac{\#\{(i,j) \in \{1,\ldots,A\}^2 : \|\boldsymbol{a}_i - \boldsymbol{a}_j\| = r\}}{A^2}, \quad 0 \leq r \leq D_{max},$$

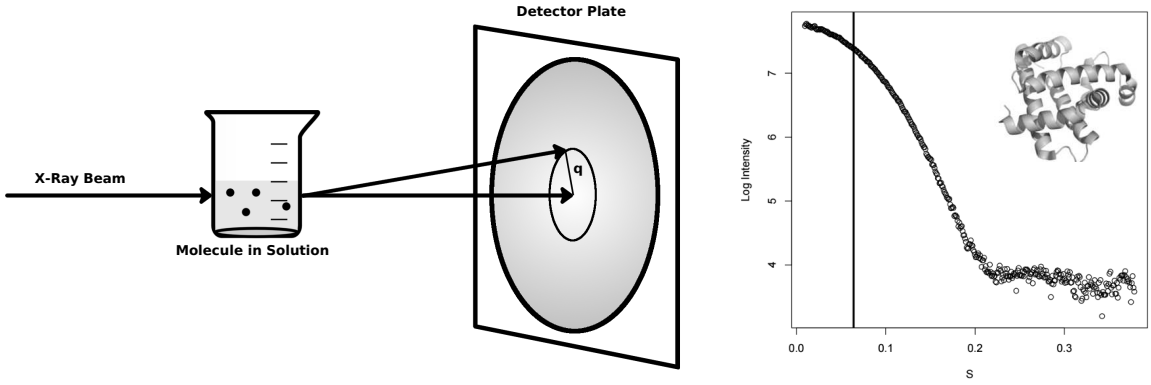


FIGURE 2.1. Schematic depiction of a SAXS experiment and resulting log-intensity data. The sample of the molecule in solution is exposed to a high-intensity X-ray beam, which scatters when interacting with the sample. The scattered pattern is recorded by a two-dimensional detector plate, which measures the intensity at different angles. In the example shown, the scattered beam intersects the detector at coordinate vector $\boldsymbol{q}$, with the origin at the center of the detector. The two-dimensional intensity data are reduced to one-dimensional data by first subtracting a reference image (not shown) and then computing an average intensity for each concentric annulus along a sequence of increasing angles. Averaging along the annulus depicted by the circle of radius $\|\boldsymbol{q}\|$ results in the average intensity value plotted on a log scale in the right-hand-side figure, at the angle $s \propto \|\boldsymbol{q}\|$ indicated by the vertical reference line. Log-intensity data in this example correspond to the molecule myoglobin, with known atomic structure depicted in the upper right-hand corner of the log-intensity plot.

with D_{max} the maximum pairwise distance. As A is typically on the order of 100–1000, we follow standard practice and use the continuous version of $p(r)$ in what follows.

The squared radius of gyration for a molecule is

$$R_g^2 = \frac{\int_0^{D_{max}} r^2 p(r)\, dr}{2\int_0^{D_{max}} p(r)\, dr} = \frac{2\pi \int_0^{D_{max}} r^2 p(r)\, dr}{\mathcal{I}(0)}, \tag{6}$$

where $\mathcal{I}(s)$ denotes the theoretical scattering intensity at scattering angle $s \geq 0$. Let $I(s)$ denote the corresponding empirical intensity from a SAXS experiment. From (6), the radius of gyration R_g is analogous to the standard deviation of a probability density, describing the spread of mass in a molecular model.

While $p(r)$ can be estimated using empirical intensity data from a SAXS experiment, such estimation requires modeling assumptions and regularization techniques to effect an inverse Fourier transformation (e.g., [10–13]). By constrast, R_g is an example of low-resolution structural information that can be estimated directly from SAXS data without modeling the molecular structure. For the theoretical log-intensity, [14] derived the approximation

$$\ln \mathcal{I}(s) = \ln \mathcal{I}(0) - \frac{1}{3} R_g^2 s^2 + \mathrm{O}\left(s^4\right); \tag{7}$$

see Remark 1 below for motivation.

Let Y_i denote the empirical log-intensity at angle s_i from a SAXS experiment. Assuming Y_i is unbiased for $\ln \mathcal{I}(s_i)$ leads to the now-standard method of *Guinier analysis* [e.g., 15, p. 71] in which a data window of near-zero angles $s_1, \ldots, s_n$ is chosen such that the relationship between squared angle s_i^2 and empirical log-intensity $Y_i = \ln I(s_i)$ appears linear. The regression model

$$Y_i = \ln \mathcal{I}(0) - \frac{1}{3} R_g^2 s_i^2 + \varepsilon_i = \beta_0 + \beta_2 s_i^2 + \varepsilon_i \tag{8}$$

is then fitted, and $\widehat{R}_g^2 = -3\widehat{\beta}_2$ is used as the estimate of R_g^2. See [16] for an analogous problem, in which the memory parameter of a long-memory time series is estimated via regression of the log-periodogram on a function of the Fourier frequencies.

Based on simulated data from idealized models of particles, Guinier recommended the now classical cutoff value s_n^{class}, obtained by iteratively adjusting the data window to achieve $s_n^{class} \widehat{R}_g < 1.3$ [17, p. 128]. A natural alternative to this physically-motivated approach is to use statistical methods to optimize the choice of cutoff value with respect to mean squared error (MSE), trading off the increased bias of a larger cutoff value (due to breakdown of

the Guinier quadratic approximation (7)) with the decreased variance due to larger sample size. This minimum MSE approach requires estimation of the bias and variance, accounting for the fact that the Guinier approximation (7) holds only for small angles. In particular, estimating the bias requires allowance for higher-order terms in the Guinier approximation.

In this chapter, we develop improved Guinier analysis methods by minimizing MSE of $\widehat{R}_g^2$ with respect to the cutoff angle s_n. We use an estimated generalized least squares (EGLS) version of the classical Guinier estimator from (8), based on the fitting of a p-th order autoregressive model, to account properly for the autocorrelation in SAXS data. To derive suitable bias and MSE approximations for the EGLS estimator, we extend the Guinier equation (7) to a fourth-degree polynomial in §2.2.2, and derive the asymptotic MSE and the optimal cutoff angle. We develop a plug-in implementation of the optimum cutoff in §2.2.3, in which the MSE approximation is minimized over an initial window determined via outlier removal, trend removal and changepoint detection. Quality of the asymptotic approximations is verified via simulation in §2.3. The proposed estimation method has much smaller MSE than the classical method across a wide range of realistic simulated conditions. Our main motivation for this work is fast and objective analysis for large suites of SAXS experimental data. In §2.4, the new methods are applied to 357 SAXS intensity curves arising from a study on the wild type nucleosome core particle and its mutants and their behavior under various experimental conditions. The resulting $\widehat{R}_g^2$ values constitute a dataset which is then analyzed using a split-split plot model to account for the complex dependence structure induced by the experimental protocols. The analysis yields powerful scientific inferences and insight into better design of SAXS experiments. A brief discussion follows in §2.5.

2.2. Theory and Methods

2.2.1. Generalized Least Squares Estimation of R_g^2. We first extend the quadratic working model (8) to allow for m independent replicates, with errors that have the same autocovariance structure in each replicate:

$$\varepsilon_{hi} = Y_{hi} - \mathrm{E}[Y_{hi}] = \phi_1 \varepsilon_{h,i-1} + \cdots + \phi_p \varepsilon_{h,i-p} + Z_{hi}. \tag{9}$$

Here, $\{Z_{hi}\}$ are independent and identically distributed with mean zero and variance σ^2, and $|1 - \phi_1 z - \cdots - \phi_p z^p| \neq 0$ for $|z| \leq 1$. Thus, the errors $\{\varepsilon_{hi}\}$ follow a causal pth order autoregressive process, AR(p), to capture the autocovariance structure of the empirical intensity curve (see [19]).

While classical Guinier analysis uses ordinary least squares, we use generalized least squares (GLS) to account for the dependence structure. Assuming the same window of angles for each of the m independent replicate SAXS log-intensity curves, the GLS estimator of the coefficients in the quadratic working model is

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\boldsymbol{\Gamma}^{-1}\mathbf{X}\right)^{-1} \mathbf{X}'\boldsymbol{\Gamma}^{-1} m^{-1} \sum_{h=1}^{m} \mathbf{Y}_h, \tag{10}$$

where $\mathbf{Y}_h = [Y_{h1}, \ldots, Y_{hn}]'$,

$$\mathbf{X}' = \begin{bmatrix} 1 & \cdots & 1 \\ s_1^2 & \cdots & s_n^2 \end{bmatrix},$$

and $\mathbf{\Gamma} = [\mathrm{Cov}(\varepsilon_{hi}, \varepsilon_{hj})]_{i,j=1}^{n}$ is an $n \times n$ covariance matrix corresponding to the AR(p) errors for each of the m replicates. Then the estimator of R_g^2 is

$$\widehat{R}_g^2 = \mathbf{e}' \left(\mathbf{X}' \mathbf{\Gamma}^{-1} \mathbf{X} \right)^{-1} \mathbf{X}' \mathbf{\Gamma}^{-1} m^{-1} \sum_{h=1}^{m} \mathbf{Y}_h, \tag{11}$$

where $\mathbf{e} = [0, -3]'$.

In what follows, we consider asymptotic approximations to the bias and variance of the GLS estimator, which is derived under the quadratic working model (8). In particular, determining the bias of the estimator requires a higher-order model for the mean structure.

2.2.2. Theoretical Results. We seek to minimize MSE of $\widehat{R}_g^2$ with respect to the cutoff angle s_n. A straightforward Taylor linearization argument shows that this is equivalent to minimizing MSE of $\widehat{R}_g$ with respect to s_n, since

$$\mathrm{MSE}\left(\widehat{R}_g\right) \simeq \frac{1}{4R_g^2} \mathrm{MSE}\left(\widehat{R}_g^2\right),$$

where R_g^2 does not depend on s_n. In this section, we present the main theorems regarding approximation and minimization of the mean squared error of $\widehat{R}_g^2$, beginning with the assumptions underlying these results. Lemmas and all proofs are given in the appendix.

2.2.2.1. *Assumptions.* Let Δ denote the spacing between angles and write $s_i = i\Delta$ for $i = 1, \ldots, N$. We consider an asymptotic formulation in which $N \to \infty$ with $\Delta \to 0$. Assume the following conditions:

(A1) The theoretical log-intensity satisfies

$$\ln \mathcal{I}(s) = \ln \mathcal{I}(0) - \frac{1}{3} R_g^2 s^2 + \frac{f^{(4)}(0)}{24} s^4 + \mathrm{O}\left(s^6\right) = f(s) + \mathrm{O}\left(s^6\right)$$

with $f^{(4)}(0) \neq 0$.

(A2) There exists $s_n \leq s_N$ such that $s_n \to 0$ and $s_n^5/\Delta \to \infty$ as $\Delta \to 0$ and $N \to \infty$.

(A3) For $s_i \leq s_N$, the empirical log-intensity of the hth replicate satisfies

$$Y_{hi} = \ln \mathcal{I}(0) - \frac{1}{3} R_g^2 s_i^2 + \frac{f^{(4)}(0)}{24} s_i^4 + \mathrm{O}\left(s_i^6\right) + \varepsilon_{hi},$$

where each $\{\varepsilon_{hi}\}$ is an independent copy of the same causal pth-order autoregressive process.

Remarks.

(1) The form of A1 arises because, under quite general conditions, the theoretical intensity satisfies

$$\mathcal{I}(s) = 4\pi \int_0^{D_{max}} p(r)\, \frac{\sin(sr)}{sr}\, dr$$

[e.g., 21], which can be expanded as

$$\begin{aligned}\mathcal{I}(s) &= 4\pi \int_0^{D_{max}} p(r)\, dr - 4\pi \frac{s^2}{3!} \int_0^{D_{max}} r^2 p(r)\, dr + 4\pi \frac{s^4}{5!} \int_0^{D_{max}} r^4 p(r)\, dr + \mathrm{O}\left(s^6\right) \\ &= \mathcal{I}(0) - \frac{1}{3} \mathcal{I}(0) R_g^2 s^2 + c_4 \mathcal{I}(0) s^4 + \mathrm{O}\left(s^6\right),\end{aligned}$$

where c_4 denotes a constant. Then in a neighborhood of zero,

$$\ln \mathcal{I}(s) = \ln \mathcal{I}(0) - \frac{1}{3} R_g^2 s^2 + \frac{f^{(4)}(0)}{24} s^4 + \mathrm{O}\left(s^6\right),$$

resulting in a quartic polynomial as specified in the assumption.

(2) Under A2–A3, the structure of the AR(p) errors remains fixed as the spacing between observations goes to zero. An alternative formulation would allow the dependence

to increase as the spacing decreases. Increasing dependence would occur if the $\{\varepsilon_{hi}\}$ are taken at a grid of points along a realization of a smooth continuous-time stochastic process. For an Ornstein-Uhlenbeck process, it can be shown that as $\Delta \to 0$, the variance of $\widehat{R}_g^2$ converges not to zero but to a positive constant. The proof is omitted. Related results for such "infill asymptotics" include [22], [23], and [24] among others. We show that our fixed-dependence asymptotic formulation leads to useful approximations in practice, and do not pursue the infill approach further in this paper.

2.2.2.2. *Mean Square Error Approximation and Optimal Cutoff.*

THEOREM 1. *Under A1–A3, the bias of* $\widehat{R}_g^2$ *is*

$$\mathrm{E}\left(\widehat{R}_g^2 - R_g^2\right) = -\frac{3}{28} f^{(4)}(0) s_n^2 + \mathrm{O}\left(s_n^4\right). \tag{12}$$

THEOREM 2. *Under A1–A3, the variance of* $\widehat{R}_g^2$ *is*

$$\mathrm{Var}\left(\widehat{R}_g^2\right) = \frac{405\sigma^2\Delta}{4ms_n^5\left(1 - \sum_{j=1}^p \phi_j\right)^2} + \mathrm{O}\left(\Delta^2 s_n^{-6}\right). \tag{13}$$

The factor $\sigma^2\left(1 - \sum_{j=1}^p \phi_j\right)^{-2}$ appearing in the asymptotic variance is 2π times the spectral density at frequency zero for an AR(p) process. It arises in other time series contexts; for example, as $\lim_{n\to\infty} n\mathrm{Var}(\bar{\varepsilon}_n)$ [e.g., 25, pp. 218–219], increasing or decreasing the variance of the sample mean due to positive or negative autocorrelation.

Immediate consequences of Theorem 1 and Theorem 2 are the approximate asymptotic mean squared error of $\widehat{R}_g^2$ and the asymptotically optimal cutoff angle.

THEOREM 3. *Under A1–A3, the approximate mean squared error of* $\widehat{R}_g^2$ *is*

$$MSE\left(\widehat{R}_g^2\right) = \frac{9}{784}\left\{f^{(4)}(0)\right\}^2 s_n^4 + \frac{405\sigma^2\Delta}{4ms_n^5\left(1-\sum_{j=1}^p \phi_j\right)^2} + \mathrm{O}\left(\Delta^2 s_n^{-6} + s_n^6\right), \tag{14}$$

which is minimized by

$$s_n^{opt} = \left[\frac{11025\sigma^2\Delta}{m\left\{f^{(4)}(0)\right\}^2\left(1-\sum_{j=1}^p \phi_j\right)^2}\right]^{1/9}. \tag{15}$$

2.2.3. IMPLEMENTATION. Use of the optimal cutoff angle (15) requires estimates of $f^{(4)}(0)$, the autoregressive order p, the coefficients $\{\phi_j\}_{j=1}^p$, and the white noise variance σ^2. We obtain estimates for each of these quantities by choosing an initial window of angles $\{s_i\}_{i=1}^N$ over which the quartic Guinier approximation in A3 is plausible, then obtaining estimates of the autoregressive model and the quartic coefficient over this initial window.

2.2.3.1. *Changepoint detection.* Our approach is to preprocess the data by first removing the trend. We difference four times

$$D_i = Y_i - 4Y_{i-1} + 6Y_{i-2} - 4Y_{i-3} + Y_{i-4}$$

to remove quite general smooth functions of s, including the fourth-degree polynomial trend assumed from (A1) to hold over the initial window. We then perform a statistical changepoint analysis (see, e.g., [26] for a recent review) on the differenced data $\{D_i\}$ to determine where the initial quartic model breaks. While there are many options for changepoint analysis, we are using a parametric method that maximizes a likelihood ratio test and is implemented in the `R` package `changepoint` ([27]). This method was chosen because it can detect an unknown number of changes in both the mean and variance of the data. With replicate

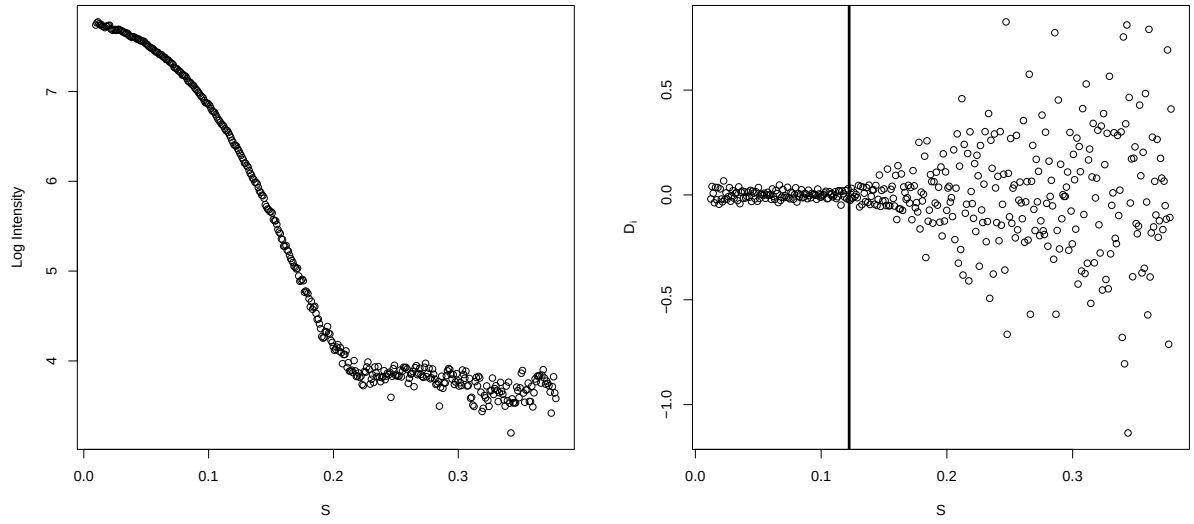


FIGURE 2.2. Left: Log intensities $\{Y_i\}$ from small-angle X-ray scattering versus scattering angle $\{s_i\}$ for the molecule myoglobin. Right: Log intensities differenced four times, with initial cutoff angle s_N selected via statistical changepoint analysis and marked with a vertical reference line.

intensity curves, we determine the changepoint for each curve and then take the minimum changepoint as the initial window for combined data.

Figure 2.2 (left) shows log-intensity versus angle for myoglobin and Figure 2.2 (right) shows those log-intensities differenced four times. Changepoint analysis on the fourth differences estimates an initial cutoff angle as $s_N = 0.122$, yielding $N = 120$ angles over which to minimize the MSE criterion.

2.2.3.2. *Outlier detection.* The left-hand endpoint s_1 is also of interest in a SAXS experiment, as outlying log-intensities may be present among the first few scattering angles. These smallest angles are subject to the greatest intensity of the X-ray beam and are adjacent to the central beam stop, both of which may lead to unusual intensity values. It is currently standard practice for the operator to perform outlier detection manually. We have developed an automated statistical procedure to detect such outliers, by adapting the standard

DFBETAS criterion (e.g., [20], §10.4) to estimation of R_g^2 under model (10). Specifically, we compute

$$\text{DFBETAS}\left(\widehat{R}^2_{g(-a)}\right) = \frac{\widehat{R}^2_g - \widehat{R}^2_{g(-a)}}{\text{SE}\left(\widehat{R}^2_g\right)}.$$

where $\widehat{R}^2_{g(-a)}$ deletes the first a observations and uses only the angles $s_{a+1}, s_{a+2}, \ldots, s_n$. Therefore, we can remove one outlying point at a time or groups of observations. We remove values if the absolute value of DFBETAS exceeds two, or if it exceeds a size-adjusted cutoff value of $2/\sqrt{\max\{n, n_{(-a)}\}}$, where $n_{(-a)}$ is the number of points used to calculate $\widehat{R}^2_{g(-a)}$. In simulations not described here, this outlier detection methodology works well, for both point outliers and groups of outliers. We assume henceforth that any initial outliers have been removed from the log-intensity curves.

2.2.3.3. *Plug-in estimation of the optimal cutoff.* We first describe the case without replicate intensity curves. Once the initial window of angles is finalized via changepoint detection and outlier elimination, we fit a cubic spline to the log-intensities over the window and obtain residuals. Over this initial window, we assume that a stationary autoregressive process is a reasonable approximation to the dependence structure; this stationarity assumption would break down over larger windows. We then use Yule-Walker estimation on the residuals from the cubic spline to fit successively higher-order autoregressive models, choosing the final order p by minimizing AIC and obtaining the Yule-Walker estimates $\{\widehat{\phi}_j\}_{j=1}^p$ and $\widehat{\sigma}^2$ [e.g., 25, §8.1]. We use the fitted autoregressive model to compute empirical generalized least squares estimates of the parameters in the model

$$Y_i = \beta_0 + \beta_2 s_i^2 + \beta_4 s_i^4 + \varepsilon_i \quad (i = 1, \ldots, N), \tag{16}$$

from which $\widehat{f}^{(4)}(0) = 24\widehat{\beta}_4$ is obtained, finalizing the set of estimates needed to plug in to (15).

For data with replicate intensity curves, we modify this procedure slightly, allowing for the fact that the initial angles for each replicate curve can vary due to the removal of outliers. Over the initial window, we determine Yule-Walker estimates $\{\widehat{\phi}_{hj}\}_{j=1}^{p_h}$ and $\widehat{\sigma}_h^2$ for each replicate curve, where $h = 1, \ldots, m$. To determine the plug-in value for (15), we average these estimates via

$$\frac{1}{m}\sum_{h=1}^{m}\frac{\widehat{\sigma}_h^2}{\left(1-\sum_{j=1}^{p_h}\widehat{\phi}_{hj}\right)^2}.$$

Furthermore, intensity data can exhibit small vertical shifts from replicate to replicate (due to variations in the X-ray source, exposure time, etc.), so we use the fitted autoregression for each replicate intensity curve to fit the model

$$Y_{hi} = \beta_{0h} + \beta_2 s_i^2 + \beta_4 s_i^4 + \varepsilon_{hi} \tag{17}$$

via empirical generalized least squares. From the fitted model, we obtain $\widehat{f}^{(4)}(0) = 24\widehat{\beta}_4$ and plug this value into (15).

2.3. Simulation Results

To evaluate our asymptotic theory and R_g estimation, we simulated artificial but realistic SAXS data as follows. First, we used the program `CRYSOL` ([28]) to compute the theoretical intensity curves for known molecular structures: DNA (a nucleic acid), glucose isomerase (a protein), and nucleosome core particle (a complex of protein and nucleic acid). From these theoretical curves, we determined R_g and chose nearby values of 20, 30, and 40, similar to the DNA, glucose isomerase, and nucleosome core particle, respectively. We also determined

$f^{(4)}(0)R_g^{-4}$ ratio values from polynomial fits to the theoretical curves and chose nearby values -0.05, 0.05, 0.10 that again are similar to DNA, glucose isomerase, and nucleosome core particle, respectively. Finally, we considered two different AR(2) models, each obtained from fits to real glucose isomerase data:

$$\begin{aligned} \text{model I:} \quad \varepsilon_I &= 0.26\varepsilon_{i-1} + 0.29\varepsilon_{i-2} + Z_i, \quad \{Z_i\} \text{ iid } \mathcal{N}(0, \sigma^2), \\ \text{model II:} \quad \varepsilon_I &= 0.10\varepsilon_{i-1} + 0.18\varepsilon_{i-2} + Z_i, \quad \{Z_i\} \text{ iid } \mathcal{N}(0, \sigma^2), \end{aligned} \tag{18}$$

with σ^2 chosen so that $\text{Var}(\varepsilon_i)R_g^{-2} = 0.0003$ under model I and $\text{Var}(\varepsilon_i)R_g^{-2} = 0.0004$ under model II. By crossing three R_g values, three $f^{(4)}(0)R_g^{-4}$ values, and two AR(2) models, we have 18 distinct experimental conditions to generate a wide range of realistic SAXS data.

For each of the 18 conditions, we simulated $m = 1$, 3, or 10 independent replicate SAXS intensity curves at $N = 99$ angles. We repeated the simulation for 1000 iterations under each of the 18 conditions at each of the three numbers of replicates.

For each simulated data set, we conducted changepoint detection as in §2.2.3.1 and used the methods of §2.2.3.3 to determine the plug-in estimate $\widehat{s}_n^{opt}$ of the optimal cutoff s_n^{opt} from (15). We also used the iterative Guinier approach to determine the classical cutoff s_n^{class} such that $s_n^{class}\widehat{R}_g < 1.3$. For replicated data, we first averaged the intensity curves across replicates to yield a single curve before applying the iterative Guinier approach. Averaging of replicate intensity curves is common in practice.

We compared these two empirical approaches to two theoretical standards: the asymptotic optimum given by equation (15) with *known* values of $f^{(4)}(0)$, $\{\phi_j\}_{j=1}^p$, and σ^2, and the empirical optimum, given by choosing s_n such that the empirical MSE (over the 1000 iterations) of $\widehat{R}_g^2$ is minimized.

In Table 2.1, we present root mean squared error (RMSE) ratios, with denominator RMSE corresponding to use of the asymptotic optimum. RMSE ratios for the empirical optimum are all close to one, showing that the asymptotic optimum is an excellent approximation to the empirical optimum under each of the conditions we consider. As expected, RMSE ratios for the estimated s_n^{opt} are larger than one, showing some loss of efficiency due to the need to plug in parameter estimates. The worst RMSE's for estimated s_n^{opt} are about 1.3 times those attained with the (infeasible) asymptotic optimum. Finally, the RMSE ratios for the classical cutoff s_n^{class} are, in nearly every case, much greater than those for the estimated s_n^{opt}, showing that the proposed method is far more efficient than the classical method across a wide range of realistic conditions.

Because the classical method effectively considers bias only, it tends to choose a small window, while the new optimal MSE method can choose a larger window, allowing some bias in return for a larger sample and reduced variance. These differences are most notable with $m = 1$ replicate. As the number of replicates increases, the optimal method chooses a smaller window. Therefore, the performance of the classical method generally shows some improvement relative to the optimum or estimated optimum as the number of replicates increases, as shown in Table 2.1. Nonetheless, the optimal method dominates the classical method in terms of RMSE even with larger numbers of replicates.

The spacing Δ and the corresponding number of angles depends on the resolution of the detector plate and other features of the equipment used in data collection. Therefore, we have conducted similar simulations (not shown here) for different values of Δ. For these simulations, we maintained comparable autoregressive structure of the data by interpolating

the original autocovariance function to reflect the new spacing, and we used this new autocovariance function to create autocorrelated data. For example, if the original autocovariance function was $\gamma_\Delta(k)$ and the spacing was cut in half, we used the interpolated autocovariance function

$$\gamma_{\Delta/2}(k) = \begin{cases} \frac{1}{2}\left\{\gamma_\Delta\left(\frac{k-1}{2}\right) + \gamma_\Delta\left(\frac{k+1}{2}\right)\right\}, & k \text{ odd;} \\ \gamma_\Delta(k/2), & k \text{ even} \end{cases}$$

to simulate twice as many data points as the original data. The R_g estimation method was applied to this new data set, and the new cutoff value was approximately double the original cutoff value, as expected. Furthermore, the RMSE values were similar to the original values. Thus, the method generalizes to the different resolutions that are common with SAXS data.

TABLE 2.1. Root mean squared error (RMSE) ratios for estimation of R_g, with RMSE's computed from 1000 simulated realizations of $m = 1$, 3, or 10 replicate SAXS log-intensity curves. Denominator RMSE corresponds to use of the asymptotic optimum cutoff angle s_n^{opt} from (15). Numerator RMSE corresponds to empirical optimum cutoff (angle yielding smallest simulation RMSE over 1000 simulated realizations), estimated asymptotic optimum $\widehat{s}_n^{opt}$, or classical cutoff s_n^{class}. Values for R_g of 20, 30, 40 and values for $f^{(4)}(0)R_g^{-4}$ of -0.05, 0.05, 0.10 correspond approximately to values for DNA, glucose isomerase, and nucleosome core particle, respectively. Second-order autoregressive models I and II are obtained from real glucose isomerase data and are given in (18).

				Empirical opt.			$\widehat{s}_n^{opt}$			s_n^{class}		
R_g	$\frac{f^{(4)}(0)}{R_g^4}$	AR	σ	1	3	10	1	3	10	1	3	10
20	-0.05	I	0.008	1.00	0.99	1.00	1.30	1.18	1.15	3.82	2.96	2.31
20	0.05	I	0.008	1.00	0.99	0.99	1.20	1.14	1.18	3.42	2.64	1.95
20	0.10	I	0.008	1.00	0.99	1.00	1.17	1.12	1.14	2.42	1.91	1.39
20	-0.05	II	0.012	0.99	1.00	1.00	1.21	1.25	1.14	4.58	3.64	2.75
20	0.05	II	0.012	0.98	1.00	0.99	1.33	1.15	1.15	4.03	3.16	2.34
20	0.10	II	0.012	0.99	0.98	0.99	1.26	1.07	1.09	2.83	2.27	1.66
30	-0.05	I	0.012	0.99	0.99	1.00	1.20	1.18	1.16	5.30	3.81	3.12
30	0.05	I	0.012	1.00	1.00	1.00	1.16	1.11	1.14	4.52	3.27	2.41
30	0.10	I	0.012	0.99	1.00	0.99	1.13	1.13	1.17	3.13	2.49	2.51
30	-0.05	II	0.019	1.00	1.00	1.00	1.26	1.16	1.24	7.37	5.04	3.69
30	0.05	II	0.019	0.98	1.00	0.98	1.18	1.09	1.07	6.14	4.30	3.27
30	0.10	II	0.019	0.96	0.98	1.00	1.08	1.10	1.18	4.42	3.10	2.74
40	-0.05	I	0.016	1.00	1.00	1.00	1.16	1.20	1.15	6.79	6.13	8.79
40	0.05	I	0.016	0.99	1.00	0.99	1.14	1.13	1.17	5.41	4.51	5.52
40	0.10	I	0.016	1.00	1.00	1.00	1.19	1.23	1.23	5.09	6.51	8.85
40	-0.05	II	0.025	0.98	1.00	0.98	1.11	1.21	1.14	8.66	6.74	7.49
40	0.05	II	0.025	0.99	1.00	0.98	1.11	1.11	1.19	6.94	5.88	5.80
40	0.10	II	0.025	0.98	0.99	1.00	1.12	1.22	1.27	6.11	5.91	8.36

2.4. Mixed Model Estimation for a Suite of SAXS Experiments

We applied the automated R_g estimation methods to a suite of 357 SAXS data sets for wild type nucleosome core particles (NCP's) and four mutations of NCP. For each molecule, [29] produced both a "salt" preparation (by adding 0.05 moles of potassium chloride per liter of solution) and a "no-salt" preparation, with two preparations of each for the wild type NCP and one preparation of each for the four remaining mutations. From each of the

12 preparations, samples at six different dilutions were formed; most (but not all) dilutions were replicated three times each for 18 samples per preparation, leading to fewer than 12×18 dilution replicates. Finally, each sample was exposed for both 0.5s and 1.0s. Three of the intensity curves were removed due to poor quality. The resulting suite of SAXS data sets consists of $n = 357$ intensity curves. We applied both the automated window selection method and preliminary outlier detection techniques to these data sets to obtain 357 $\widehat{R}_g$ values, one at each experimental setting: we emphasize that this is not trivial without the new semi-automated methods.

These estimates were used by [29] to conduct all 66 pairwise comparisons among the 12 preparations. Because the method yields smaller MSE's than the classical method, it is better able to distinguish among different R_g values, leading to a suggestion how the nucleosome changes shape in solution as a consequence of histone mutation.

We now extend the analysis by using restricted maximum likelihood to fit a linear mixed model of the form

$$\begin{aligned} \widehat{R}_g \;=\; & \texttt{mol} + \texttt{salt} + \texttt{mol*salt} + \texttt{dil} + \texttt{mol*dil} + \texttt{salt*dil} + \texttt{mol*salt*dil} \\ & + \texttt{exp\_time} + \texttt{mol*exp\_time} + \texttt{salt*exp\_time} + \texttt{mol*salt*exp\_time} \\ & + \texttt{dil*exp\_time} + \texttt{mol*dil*exp\_time} + \texttt{salt*dil*exp\_time} \\ & + \texttt{mol*salt*dil*exp\_time} \\ & + \texttt{prep} + \texttt{dil\_replicate} + \texttt{noise}, \end{aligned} \tag{19}$$

where the first 15 terms represent fixed effects of the given experimental factors and the final three terms represent zero-mean random effects, uncorrelated with one another. This linear

mixed model corresponds to a split-split plot analysis, in which SAXS data from the same preparation are correlated because they share `prep` values, with even greater correlation if they are from the same dilution replicate and share `dil_replicate` values. Coefficients of the 15 fixed effects and the variance components for the three random effects are estimated via maximum likelihood and restricted maximum likelihood, respectively, using standard statistical software.

A number of biological insights, extending those of [29], are apparent from the results of the analysis as shown in Table 2.2. For example, R_g is a measure of curvature in the log-intensity and should not change with increased concentration or exposure time unless the shapes of the curves change, for which there is extensive evidence. Further, it is clear that a better experimental design to detect differences among mutations would have started with more replicated `mol*salt` preparations [30].

We repeated the analysis in Table 2.2 but with a weighted approach using the estimated standard error of the $\widehat{R}_g$ values determined from each individual intensity curve. The results from this weighted fit are omitted since they were generally similar to the results in Table 2.2, with only one important difference. In the weighted analysis, the `mol*exp_time` interaction is significant and the `salt*exp_time` interaction is not significant, but in the unweighted analysis these results are reversed. The similarities between the weighted and unweighted analysis are not surprising, since the molecular structures of the NCP mutations and the experimental conditions are similar enough that we do not expect a lot of variation in the variance of $\widehat{R}_g$. Indeed, in this experiment the R_g values are estimated with considerable stability: the median estimated coefficient of variation for $\widehat{R}_g$ is 0.2% and the 95th percentile is 0.6%.

TABLE 2.2. Restricted maximum likelihood analysis for R_g estimates from a suite of SAXS experiments. Tests of main effects and interactions for molecule type (`mol`: five mutations of nucleosome core particle), salt level (`salt`: two levels), dilution (`dil`: six levels), and exposure time (`exp_time`: 0.5s and 1.0s) from fitting of a linear mixed model via restricted maximum likelihood, with random effects to account for correlations due to repeated exposures of the same dilution replicates, and due to forming dilution replicates from the same `mol*salt` preparation.

Effect	Num. DF	Den. DF	F Value	p-value
`mol`	4	2	30.84	0.0317
`salt`	1	2	110.99	0.089
`mol*salt`	4	2	3.76	0.2208
`dil`	5	125	17.41	< 0.0001
`mol*dil`	18	125	3.70	< 0.0001
`salt*dil`	5	125	8.47	< 0.0001
`mol*salt*dil`	16	125	2.20	0.0081
`exp_time`	1	125	9.21	0.0029
`mol*exp_time`	4	125	1.15	0.3384
`salt*exp_time`	1	125	4.34	0.0393
`mol*salt*exp_time`	4	125	2.07	0.0882
`dil*exp_time`	5	125	0.81	0.5414
`mol*dil*exp_time`	18	125	2.13	0.0083
`salt*dil*exp_time`	5	125	1.37	0.2387
`mol*salt*dil*exp_time`	16	125	1.58	0.0850

2.5. DISCUSSION

We have shown that a largely automatic procedure, developed from asymptotic theory but readily implemented with standard statistical tools, can be used to determine an optimal window of angles for estimation of the radius of gyration in small-angle X-ray scattering experiments. The fast and objective nature of this procedure makes it possible to process large suites of SAXS experiments, allowing the use of other statistical methods such as the split-split plot analysis described in §2.4. Use of such methods can in turn lead to better inference from SAXS data and better design of future SAXS experiments.

Software Availability. Example data sets and `R` code that implements the minimum MSE R_g estimation procedure are freely accessible at http://hdl.handle.net/10217/167285. The Appendix contains instructions for implementation of this code.

Acknowledgements. SAXS data presented in this work were collected at the Advanced Light Source (ALS), a national user facility operated by Lawrence Berkeley National Laboratory on behalf of the Department of Energy, Office of Basic Energy Sciences, through the Integrated Diffraction Analysis Technologies (IDAT) program, supported by DOE Office of Biological and Environmental Research. Additional support comes from the National Institutes of Health project MINOS (R01GM105404).

This work was supported by the Joint National Science Foundation/National Institute of General Medical Sciences Initiative to Support Research in the Area of Mathematical Biology [R01GM096192 to FJB]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute Of General Medical Sciences or the National Institutes of Health.

2.6. Appendix

2.6.1. Lemmas. In what follows, we use the Cholesky decomposition for the covariance matrix $\mathbf{\Gamma}$ of the autoregressive process,

$$\mathbf{T\Gamma T}' = \mathbf{D}, \tag{20}$$

where $\mathbf{T}'$ is an upper triangular matrix given by

$$\mathbf{T}' = \begin{bmatrix} 1 & -\phi_1 & -\phi_2 & -\phi_3 & \dots & -\phi_p & 0 & \dots & 0 \\ 0 & 1 & -\phi_1 & -\phi_2 & \dots & -\phi_{p-1} & -\phi_p & \dots & 0 \\ 0 & 0 & 1 & -\phi_1 & \dots & -\phi_{p-2} & \dots & & \vdots \\ \vdots & \vdots & \vdots & \ddots & & & & & \\ 0 & 0 & 0 & \dots & & & & & 1 \end{bmatrix} \tag{21}$$

and $\mathbf{D} = \text{diag}(\sigma^2, \dots, \sigma^2)$; see, for example, [25] §8.6 for further details.

LEMMA 1. *Using the Cholesky decomposition (20),*

$$n^{-1}\mathbf{X}'\mathbf{\Gamma}^{-1}\mathbf{X} = n^{-1}\mathbf{X}'\mathbf{T}'\mathbf{D}^{-1}\mathbf{T}\mathbf{X} = \frac{1}{\sigma^2}\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix},$$

where

$$a_{11} = n^{-1}\sum_{i=p+1}^{n}\left(1 - \sum_{j=1}^{p}\phi_j\right)^2 = \frac{n-p}{n}\left(1 - \sum_{j=1}^{p}\phi_j\right)^2$$

$$+ n^{-1}\left\{1 + (1-\phi_1)^2 + \cdots + (1-\phi_1 - \cdots - \phi_{p-1})^2\right\}$$

$$a_{12} = a_{21} = n^{-1}\left(1 - \sum_{j=1}^{p}\phi_j\right)\sum_{i=p+1}^{n}\left\{i^2\Delta^2 - \sum_{j=1}^{p}\phi_j\,(i-j)^2\,\Delta^2\right\}$$

$$+ n^{-1}\{s_1^2 + (1-\phi_1)\left(s_2^2 - \phi_1 s_1^2\right) + \dots$$

$$+ (1 - \phi_1 - \cdots - \phi_{p-1})\left(s_p^2 - \phi_1 s_{p-1}^2 - \cdots - \phi_{p-1}s_1^2\right)\}$$

$$a_{22} = n^{-1} \sum_{i=p+1}^{n} \left\{ i^2\Delta^2 - \sum_{j=1}^{p} \phi_j \left(i-j\right)^2 \Delta^2 \right\}^2$$

$$+ n^{-1}\{s_1^4 + \left(s_2 - \phi_1 s_1^2\right)^2 + \ldots$$

$$+ \left(s_p^2 - \phi_1 s_{p-1}^2 - \cdots - \phi_{p-1} s_1^2\right)^2\}.$$

PROOF. By (21), we have

$$\mathbf{X}'\mathbf{T}' = \begin{bmatrix} 1 & 1-\phi_1 & \ldots & 1-\sum_{j=1}^{p}\phi_j & \ldots & 1-\sum_{j=1}^{p}\phi_j \\ s_1^2 & s_2^2-\phi_1 s_1^2 & \ldots & s_{p+1}^2-\sum_{j=1}^{p}\phi_j s_{p+1-j}^2 & \ldots & s_n^2-\sum_{j=1}^{p}\phi_j s_{n-j}^2 \end{bmatrix}. \tag{22}$$

Substituting $s_i = i\Delta$, we can then write

$$n^{-1}\mathbf{X}'\mathbf{\Gamma}^{-1}\mathbf{X} = n^{-1}\mathbf{X}'\mathbf{T}'\mathbf{D}^{-1}\mathbf{T}\mathbf{X} = \frac{1}{\sigma^2}\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix},$$

where

$$a_{11} = n^{-1} \sum_{i=p+1}^{n} \left(1 - \sum_{j=1}^{p} \phi_j\right)^2 = \frac{n-p}{n}\left(1 - \sum_{j=1}^{p} \phi_j\right)^2$$

$$+ n^{-1}\left\{1 + \left(1-\phi_1\right)^2 + \cdots + \left(1-\phi_1-\cdots-\phi_{p-1}\right)^2\right\}$$

$$a_{12} = a_{21} = n^{-1}\left(1 - \sum_{j=1}^{p} \phi_j\right) \sum_{i=p+1}^{n} \left\{ i^2\Delta^2 - \sum_{j=1}^{p} \phi_j \left(i-j\right)^2 \Delta^2 \right\}$$

$$+ n^{-1}\{s_1^2 + \left(1-\phi_1\right)\left(s_2^2 - \phi_1 s_1^2\right) + \ldots$$

$$+ (1 - \phi_1 - \cdots - \phi_{p-1}) \left(s_p^2 - \phi_1 s_{p-1}^2 - \cdots - \phi_{p-1} s_1^2\right)\}$$

$$a_{22} = n^{-1} \sum_{i=p+1}^{n} \left\{ i^2\Delta^2 - \sum_{j=1}^{p} \phi_j \left(i-j\right)^2 \Delta^2 \right\}^2$$

$$+ n^{-1} \left\{ s_1^4 + \left(s_2 - \phi_1 s_1^2\right)^2 + \cdots + \left(s_p^2 - \phi_1 s_{p-1}^2 - \cdots - \phi_{p-1} s_1^2\right)^2 \right\},$$

proving the result. □

LEMMA 2. *Using the Cholesky decomposition (20),*

$$n^{-1}\mathbf{X}'\mathbf{\Gamma}^{-1} \begin{bmatrix} s_1^4 \\ \vdots \\ s_n^4 \end{bmatrix} = n^{-1}\mathbf{X}'\mathbf{T}'\mathbf{D}^{-1}\mathbf{T} \begin{bmatrix} s_1^4 \\ \vdots \\ s_n^4 \end{bmatrix} = \frac{1}{\sigma^2} \begin{bmatrix} b_1 \\ \\ b_2 \end{bmatrix},$$

where

$$b_1 = n^{-1} \left(1 - \sum_{j=1}^{p} \phi_j\right) \sum_{i=p+1}^{n} \left(i^4\Delta^4 - \sum_{j=1}^{p} \phi_j \left(i-j\right)^4 \Delta^4 \right)$$

$$+ n^{-1}\{s_1^4 + (1-\phi_1)\left(s_2^4 - \phi_1 s_1^4\right) + \ldots$$

$$+ (1 - \phi_1 - \cdots - \phi_{p-1}) \left(s_p^4 - \phi_1 s_{p-1}^4 - \cdots - \phi_{p-1} s_1^4\right)\}$$

$$b_2 = n^{-1} \sum_{i=p+1}^{n} \left(i^2\Delta^2 - \sum_{j=1}^{p} \phi_j \left(i-j\right)^2 \Delta^2 \right) \left(i^4\Delta^4 - \sum_{j=1}^{p} \phi_j \left(i-j\right)^4 \Delta^4 \right)$$

$$+ n^{-1}\{s_1^6 + \left(s_2^2 - \phi_1 s_1^2\right)\left(s_2^4 - \phi_1 s_1^4\right) + \ldots$$

$$+ \left(s_p^2 - \phi_1 s_{p-1}^2 - \cdots - \phi_{p-1} s_1^2\right) \left(s_p^4 - \phi_1 s_{p-1}^4 - \cdots - \phi_{p-1} s_1^4\right)\}.$$

PROOF. By (22), we can write

$$n^{-1}\mathbf{X'T'D^{-1}T} \begin{bmatrix} s_1^4 \\ \vdots \\ s_n^4 \end{bmatrix} = n^{-1}\mathbf{X'T'D^{-1}} \begin{bmatrix} s_1^4 \\ s_2^4 - s_1^4\phi_1 \\ \vdots \\ s_{p+1}^4 - \sum_{j=1}^p \phi_j s_{p+1-j}^4 \\ \vdots \\ s_n^4 - \sum_{j=1}^p \phi_j s_{n-j}^4 \end{bmatrix} = \frac{1}{\sigma^2} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix},$$

where

$$\begin{aligned} b_1 &= n^{-1} \left(1 - \sum_{j=1}^p \phi_j\right) \sum_{i=p+1}^n \left(s_i^4 - \sum_{j=1}^p \phi_j s_{i-j}^4\right) \\ &\quad + n^{-1}\{s_1^4 + (1-\phi_1)\left(s_2^4 - \phi_1 s_1^4\right) + \ldots \\ &\quad + (1 - \phi_1 - \cdots - \phi_{p-1})\left(s_p^4 - \phi_1 s_{p-1}^4 - \cdots - \phi_{p-1} s_1^4\right)\} \\ b_2 &= n^{-1} \sum_{i=p+1}^n \left(s_i^2 - \sum_{j=1}^p \phi_j s_{i-j}^2\right)\left(s_i^4 - \sum_{j=1}^p \phi_j s_{i-j}^4\right) \\ &\quad + n^{-1}\{s_1^6 + \left(s_2^2 - \phi_1 s_1^2\right)\left(s_2^4 - \phi_1 s_1^4\right) + \ldots \\ &\quad + \left(s_p^2 - \phi_1 s_{p-1}^2 - \cdots - \phi_{p-1} s_1^2\right)\left(s_p^4 - \phi_1 s_{p-1}^4 - \cdots - \phi_{p-1} s_1^4\right)\}. \end{aligned}$$

Substituting $s_i = i\Delta$ yields the result. □

LEMMA 3. *Under A2,*

(a) For k a non-negative integer,

$$\frac{\Delta^k}{n}\sum_{i=1}^{n} i^k = \frac{s_n^k}{k+1} + \mathrm{O}\left(\Delta s_n^{k-1}\right).$$

(b)

$$\frac{a_{12}}{\left(1-\sum_{j=1}^{p}\phi_j\right)} = \frac{\Delta^2}{n}\sum_{i=p+1}^{n}\left\{i^2 - \sum_{j=1}^{p}\phi_j(i-j)^2\right\} = \frac{s_n^2}{3}\left(1-\sum_{j=1}^{p}\phi_j\right) + \mathrm{O}\left(\Delta s_n\right).$$

(c)

$$a_{22} = \frac{\Delta^4}{n}\sum_{i=p+1}^{n}\left\{i^2 - \sum_{j=1}^{p}\phi_j(i-j)^2\right\}^2 = \frac{s_n^4}{5}\left(1-\sum_{j=1}^{p}\phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^3\right).$$

(d)

$$\frac{b_1}{\left(1-\sum_{j=1}^{p}\phi_j\right)} = \frac{\Delta^4}{n}\sum_{i=p+1}^{n}\left\{i^4 - \sum_{j=1}^{p}\phi_j(i-j)^4\right\} = \frac{s_n^4}{5}\left(1-\sum_{j=1}^{p}\phi_j\right) + \mathrm{O}\left(\Delta s_n^3\right).$$

(e)

$$b_2 = \frac{\Delta^6}{n}\sum_{i=p+1}^{n}\left\{i^2 - \sum_{j=1}^{p}\phi_j(i-j)^2\right\}\left\{i^4 - \sum_{j=1}^{p}\phi_j(i-j)^4\right\} = \frac{s_n^6}{7}\left(1-\sum_{j=1}^{p}\phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^5\right).$$

PROOF. In each summation on i, let k denote the highest power of i, and use the fact that for fixed $p \geq 0$, A2 implies

$$\frac{\Delta^k}{n}\sum_{i=p+1}^{n} i^k = \frac{\Delta^k n^{k+1}}{n(k+1)} + \mathrm{O}\left(\Delta^k n^{k-1}\right) = \frac{s_n^k}{k+1} + \mathrm{O}\left(\Delta s_n^{k-1}\right).$$

□

LEMMA 4. *Under A1–A3, we have*

$$\left(n^{-1}\mathbf{X}'\mathbf{\Gamma}^{-1}\mathbf{X}\right)^{-1} = \left(\frac{1}{\frac{4}{45}\left(1-\sum_{j=1}^{p}\phi_j\right)^4 s_n^4} + \mathrm{O}\left(s_n^{-4}\right)\right)$$
$$\times \begin{bmatrix} \frac{s_n^4}{5}\left(1-\sum_{j=1}^{p}\phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^3\right) & -\frac{s_n^2}{3}\left(1-\sum_{j=1}^{p}\phi_j\right)^2 + \mathrm{O}\left(\Delta s_n\right) \\ -\frac{s_n^2}{3}\left(1-\sum_{j=1}^{p}\phi_j\right)^2 + \mathrm{O}\left(\Delta s_n\right) & \left(1-\sum_{j=1}^{p}\phi_j\right)^2 \end{bmatrix}.$$

PROOF. By Lemma 1 and Lemma 3, we have

$$n^{-1}\mathbf{X}'\mathbf{\Gamma}^{-1}\mathbf{X} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix},$$

where

$$a_{11} = \left(1-\sum_{j=1}^{p}\phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^{-1}\right), \quad a_{12} = a_{21} = \frac{s_n^2}{3}\left(1-\sum_{j=1}^{p}\phi_j\right)^2 + \mathrm{O}\left(\Delta s_n\right),$$

and

$$a_{22} = \frac{s_n^4}{5}\left(1-\sum_{j=1}^{p}\phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^3\right).$$

Taking the inverse of $n\mathbf{X}'\mathbf{\Gamma}^{-1}\mathbf{X}$ yields

$$\left(n^{-1}\mathbf{X}'\mathbf{\Gamma}^{-1}\mathbf{X}\right)^{-1} = \frac{{s_n}^{-4}}{\frac{4}{45}\left(1-\sum_{j=1}^{p}\phi_j\right)^4 + \mathrm{O}\left(\Delta s_n^{-1}\right)}$$
$$\times \begin{bmatrix} \frac{s_n^4}{5}\left(1-\sum_{j=1}^{p}\phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^3\right) & -\frac{s_n^2}{3}\left(1-\sum_{j=1}^{p}\phi_j\right)^2 + \mathrm{O}\left(\Delta s_n\right) \\ -\frac{s_n^2}{3}\left(1-\sum_{j=1}^{p}\phi_j\right)^2 + \mathrm{O}\left(\Delta s_n\right) & \left(1-\sum_{j=1}^{p}\phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^{-1}\right) \end{bmatrix}.$$

By the assumed causality in A3, $1-\sum_{j=1}^{p}\phi_j \neq 0$, so we can write

$$\frac{{s_n}^{-4}}{\frac{4}{45}\left(1-\sum_{j=1}^{p}\phi_j\right)^4+\mathrm{O}\left(\Delta s_n^{-1}\right)}=\frac{{s_n}^{-4}}{\frac{4}{45}\left(1-\sum_{j=1}^{p}\phi_j\right)^4}+\mathrm{O}\left(\Delta s_n^{-5}\right).$$

□

2.6.2. Proof of Theorems.

Proof of Theorem 1. By A3, the first term of the bias of $\widehat{R}_g^2$ is

$$\mathrm{E}\left(\widehat{R}_g^2-R_g^2\right)=B_n$$

where

$$B_n=\frac{f^{(4)}(0)}{24}\mathbf{e}'\left(\mathbf{X}'\mathbf{\Gamma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{\Gamma}^{-1}\begin{bmatrix}s_1^4\\ \vdots\\ s_n^4\end{bmatrix}+\mathbf{e}'\left(\mathbf{X}'\mathbf{\Gamma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{\Gamma}^{-1}\begin{bmatrix}\mathrm{O}\left(s_1^6\right)\\ \vdots\\ \mathrm{O}\left(s_n^6\right)\end{bmatrix}$$

$$=\frac{f^{(4)}(0)}{24}\mathbf{e}'\left(\mathbf{X}'\mathbf{\Gamma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{\Gamma}^{-1}\begin{bmatrix}s_1^4\\ \vdots\\ s_n^4\end{bmatrix}+\mathbf{e}'\left(\mathbf{X}'\mathbf{\Gamma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{\Gamma}^{-1}\begin{bmatrix}\mathrm{O}\left(s_n^6\right)\\ \vdots\\ \mathrm{O}\left(s_n^6\right)\end{bmatrix}$$

$$=\frac{f^{(4)}(0)}{24}\mathbf{e}'\begin{bmatrix}\left(1-\sum_{j=1}^{p}\phi_j\right)^2+\mathrm{O}\left(\Delta s_n^{-1}\right) & \frac{s_n^2}{3}\left(1-\sum_{j=1}^{p}\phi_j\right)^2+\mathrm{O}\left(\Delta s_n\right)\\ \frac{s_n^2}{3}\left(1-\sum_{j=1}^{p}\phi_j\right)^2+\mathrm{O}\left(\Delta s_n\right) & \frac{s_n^4}{5}\left(1-\sum_{j=1}^{p}\phi_j\right)^2+\mathrm{O}\left(\Delta s_n^3\right)\end{bmatrix}^{-1}$$

$$\times\begin{bmatrix}\left(1-\sum_{j=1}^{p}\phi_j\right)^2\frac{s_n^4}{5}+\mathrm{O}\left(\Delta s_n^3\right)\\ \left(1-\sum_{j=1}^{p}\phi_j\right)^2\frac{s_n^6}{7}+\mathrm{O}\left(\Delta s_n^5\right)\end{bmatrix}+\mathrm{O}\left(s_n^4\right).$$

By Lemma 4,

$$B_n = \frac{f^{(4)}(0)}{24}\mathbf{e}'\left(\frac{1}{\frac{4}{45}\left(1-\sum_{j=1}^{p}\phi_j\right)^4 s_n^4} + \mathrm{O}\left(\Delta s_n^{-5}\right)\right)$$
$$\times\begin{bmatrix} \frac{s_n^4}{5}\left(1-\sum_{j=1}^{p}\phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^3\right) & -\frac{s_n^2}{3}\left(1-\sum_{j=1}^{p}\phi_j\right)^2 + \mathrm{O}\left(\Delta s_n\right) \\ -\frac{s_n^2}{3}\left(1-\sum_{j=1}^{p}\phi_j\right)^2 + \mathrm{O}\left(\Delta s_n\right) & \left(1-\sum_{j=1}^{p}\phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^{-1}\right) \end{bmatrix}$$
$$\times\begin{bmatrix} \left(1-\sum_{j=1}^{p}\phi_j\right)^2 \frac{s_n^4}{5} + \mathrm{O}\left(\Delta s_n^3\right) \\ \left(1-\sum_{j=1}^{p}\phi_j\right)^2 \frac{s_n^6}{7} + \mathrm{O}\left(\Delta s_n^5\right) \end{bmatrix} + \mathrm{O}\left(s_n^4\right).$$

Multiplying the row vector $\mathbf{e}'$ through yields

$$B_n = -\frac{f^{(4)}(0)}{8}\left(\frac{1}{\frac{4}{45}\left(1-\sum_{j=1}^{p}\phi_j\right)^4 s_n^4} + \mathrm{O}\left(\Delta s_n^{-5}\right)\right)$$
$$\times\begin{bmatrix} -\frac{s_n^2}{3}\left(1-\sum_{j=1}^{p}\phi_j\right)^2 + \mathrm{O}\left(\Delta s_n\right) & \left(1-\sum_{j=1}^{p}\phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^{-1}\right) \end{bmatrix}$$
$$\times\begin{bmatrix} \left(1-\sum_{j=1}^{p}\phi_j\right)^2 \frac{s_n^4}{5} + \mathrm{O}\left(\Delta s_n^3\right) \\ \left(1-\sum_{j=1}^{p}\phi_j\right)^2 \frac{s_n^6}{7} + \mathrm{O}\left(\Delta s_n^5\right) \end{bmatrix} + \mathrm{O}\left(s_n^4\right)$$
$$= -\frac{3}{24}f^{(4)}(0)\left(\frac{45}{4s_n^4} + \mathrm{O}\left(\Delta s_n^{-5}\right)\right)\left(\frac{8}{105}s_n^6 + \mathrm{O}\left(\Delta s_n^4\right)\right) + \mathrm{O}\left(s_n^4\right)$$
$$= -\frac{3}{28}f^{(4)}(0)s_n^2 + \mathrm{O}\left(s_n^4\right),$$

since $(\Delta s_n)/(s_n^4) = \Delta^{-2}n^{-3} \to 0$ by A2. Hence, the final bias expression of $\widehat{R}_g^2$ is

$$\mathrm{E}\left(\widehat{R}_g^2 - R_g^2\right) = -\frac{3}{28}f^{(4)}(0)s_n^2 + \mathrm{O}\left(s_n^4\right). \tag{23}$$

□

PROOF OF THEOREM 2. From (11), we have

$$\begin{aligned}
\mathrm{Var}\left(\widehat{R}_g^2\right) &= \mathrm{Var}\left\{\mathbf{e}'\left(\mathbf{X}'\boldsymbol{\Gamma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Gamma}^{-1}m^{-1}\sum_{h=1}^{m}\mathbf{Y}_h,\right\} \\
&= m^{-1}\mathbf{e}'\left(\mathbf{X}'\boldsymbol{\Gamma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathrm{Var}(\mathbf{Y}_1)\mathbf{X}\left(\mathbf{X}'\boldsymbol{\Gamma}^{-1}\mathbf{X}\right)^{-1}\mathbf{e} \\
&= m^{-1}\mathbf{e}'\left(\mathbf{X}'\boldsymbol{\Gamma}^{-1}\mathbf{X}\right)^{-1}\mathbf{e} \\
&= \frac{\sigma^2\mathbf{e}'}{mn}\begin{bmatrix} \left(1-\sum_{j=1}^{p}\phi_j\right)^2+\mathrm{O}\left(\Delta s_n^{-1}\right) & \frac{s_n^2}{3}\left(1-\sum_{j=1}^{p}\phi_j\right)^2+\mathrm{O}\left(\Delta s_n\right) \\ \frac{s_n^2}{3}\left(1-\sum_{j=1}^{p}\phi_j\right)^2+\mathrm{O}\left(\Delta s_n\right) & \frac{s_n^4}{5}\left(1-\sum_{j=1}^{p}\phi_j\right)^2+\mathrm{O}\left(\Delta s_n^3\right) \end{bmatrix}^{-1}\mathbf{e}.
\end{aligned}$$

Next, by Lemma 4 we have

$$\begin{aligned}
\mathrm{Var}\left(\widehat{R}_g^2\right) &= \frac{\sigma^2\mathbf{e}'}{mn}\left(\frac{1}{\frac{4}{45}\left(1-\sum_{j=1}^{p}\phi_j\right)^4 s_n^4}+\mathrm{O}\left(\Delta s_n^{-5}\right)\right) \\
&\quad\times\begin{bmatrix} \frac{s_n^4}{5}\left(1-\sum_{j=1}^{p}\phi_j\right)^2+\mathrm{O}\left(\Delta s_n^3\right) & -\frac{s_n^2}{3}\left(1-\sum_{j=1}^{p}\phi_j\right)^2+\mathrm{O}\left(\Delta s_n\right) \\ -\frac{s_n^2}{3}\left(1-\sum_{j=1}^{p}\phi_j\right)^2+\mathrm{O}\left(\Delta s_n\right) & \left(1-\sum_{j=1}^{p}\phi_j\right)^2+\mathrm{O}\left(\Delta s_n^{-1}\right) \end{bmatrix}\mathbf{e} \\
&= \frac{405\sigma^2}{4mns_n^4\left(1-\sum_{j=1}^{p}\phi_j\right)^2}+\mathrm{O}\left(\Delta^2 s_n^{-6}\right),
\end{aligned}$$

by Lemma 1 and Lemma 3. Writing $n^{-1}=\Delta s_n^{-1}$, we have

$$\mathrm{Var}\left(\widehat{R}_g^2\right) = \frac{405\sigma^2\Delta}{4ms_n^5\left(1-\sum_{j=1}^{p}\phi_j\right)^2}+\mathrm{O}\left(\Delta^2 s_n^{-6}\right).$$

This expression goes to zero as $\Delta\to 0$ and $s_n\to 0$ by A2. □

PROOF OF THEOREM 3. Combining (12) and (13), the mean squared error is given by

$$\begin{aligned}
\mathrm{MSE}\left(\widehat{R}_g^2\right) &= \left\{\mathrm{E}\left(\widehat{R}_g^2 - R_g^2\right)\right\}^2 + \mathrm{Var}\left(\widehat{R}_g^2\right) \\
&= \left\{-\frac{3}{28} f^{(4)}(0) s_n^2 + \mathrm{O}\left(s_n^4\right)\right\}^2 + \frac{405\sigma^2\Delta}{4ms_n^5\left(1 - \sum_{j=1}^p \phi_j\right)^2} + \mathrm{O}\left(\Delta^2 s_n^{-6}\right) \\
&= \frac{9}{784}\left\{f^{(4)}(0)\right\}^2 s_n^4 + \frac{405\sigma^2\Delta}{4ms_n^5\left(1 - \sum_{j=1}^p \phi_j\right)^2} + \mathrm{O}\left(\Delta^2 s_n^{-6} + s_n^6\right). \qquad (24)
\end{aligned}$$

Differentiating (24) with respect to s_n yields

$$\begin{aligned}
\frac{\partial}{\partial s_n}\left\{\mathrm{MSE}\left(\widehat{R}_g^2\right)\right\} &= \frac{\partial}{\partial s_n}\left[\frac{9}{784}\left\{f^{(4)}(0)\right\}^2 s_n^4 + \frac{405\sigma^2\Delta}{4ms_n^5\left(1 - \sum_{j=1}^p \phi_j\right)^2}\right] \\
&= \frac{9}{196}\left\{f^{(4)}(0)\right\}^2 s_n^3 - \frac{2025\sigma^2\Delta}{4ms_n^6\left(1 - \sum_{j=1}^p \phi_j\right)^2}. \qquad (25)
\end{aligned}$$

Setting (25) equal to zero and solving for s_n yields the optimal cutoff as given in (15). □

CHAPTER 3

ESTIMATING THE RADIUS OF GYRATION FOR BIOLOGICAL MACROMOLECULES

3.1. Introduction

Small-angle scattering is a low-resolution solution-based biophysical characterization technique that provides information about the shape and size of molecules and complexes in solution. The shape is approached with molecular envelopes [31]. However, scattering can provide information that is different from information available from crystallography by probing dynamic molecular behavior in solution. Questions about the influence of complexation, substrate binding, the buffer (pH, presence or absence of specific ions), or mutations on the global molecular properties can in theory be addressed. In practice each of these parameters will generally only cause small changes in the global behavior of a molecule, that is, changes in parameters measured are often small. It is therefore important to accurately and precisely quantify small changes in these parameters and compare these changes with the noise inherent in the experiment to ascertain that they relevant. This quantification can be accomplished by a correct statistical treatment of the data and can be improved by inclusion of replicate experimental data sets. This work provides a new algorithm to derive an optimized value for the radius of gyration R_g from scattering data and further improve this value by enabling simultaneous consideration of experimental replicates. A balanced approach is to increase precision of R_g by considering more data points in Guinier analysis, without compromising accuracy.

In 1939 André Guinier published a seminal paper [14] that describes the theory, instrumental development and data interpretation of small-angle X-ray scattering (SAXS) data. In this work, he demonstrated how SAXS can be applied in various disciplines, such as material science, colloid chemistry and structural biology. Over time, Guinier's work became easily accessible and the method of small angle scattering became widely used [17]. Although this technique was established decades ago, it has found a renewed interest among structural biologists with the publication of new methods to reconstruct molecular envelopes from scattering data [32, 12]. SAXS studies on biological systems can reveal important new insight, particularly for samples that are not amenable to traditional structural analysis methods such as X-ray crystallography. Recently published examples are intrinsically disordered proteins (for example antitoxin PaaA2, [33]), nucleic acids (for example riboswitches reviewed in [34]), and protein-nucleic acid complexes (for example DNA-methyltransferase complex, [35]). The use of small angle scattering in structural biology is reviewed extensively elsewhere [5, 6].

The radius of gyration is a parameter that can be derived from SAXS data without any assumptions about the sample. It represents the square root of the average squared distance of each electron from the center of the molecule. Thus, it depends both on size (number of atoms) and shape (distribution of atoms) of a particle. It can be used to probe the change of size or shape, for example in the case of a formation of a complex.

The example of the riboswitches reviewed in Zhang *et al.* elegantly shows that biological information can be derived from R_g values, and how important the precision of R_g is when comparing data from riboswitches, with and without Mg^{+2} and with and without ligands. It is clear that riboswitches change their shape in response to these biologically important

conditions and the changes can be expressed in R_g values. The comparisons are significantly aided by increased precision.

We have developed a new algorithm that analyses the intensity of a small-angle scattering experiment, determines the linear range of the experimental data points in Guinier analysis, and optimizes and reports the precision of R_g.

This algorithm enables the simultaneous consideration of experimental replicates, a method well known to improve the estimated value of parameters derived from experiments, in this case R_g. It enables a more objective interpretation of scattering data by providing a statistical basis for the choice of one cutoff point in the Guinier analysis. We demonstrate advantages in practice by applying this algorithm to experimental SAXS data for nucleosome core particles.

3.2. Materials and Methods

3.2.1. Classical Guinier analysis for estimation of R_g. Let $\mathcal{I}(s)$ denote the theoretical scattering intensity at scattering angle s and let $I(s)$ denote the corresponding empirical intensity from a SAXS experiment. Guinier [14] derived the theoretical approximation

$$\ln \mathcal{I}(s) \simeq \ln \mathcal{I}(0) - \frac{1}{3} R_g^2 s^2, \tag{26}$$

valid for small values of s. This quadratic relationship near the origin is used as the basis for an estimation method: choose a data window of near-zero angles $s_1, \ldots, s_n$ such that the relationship between squared angle s_i^2 and empirical log-intensity $\ln I(s_i)$ appears linear, fit

the linear regression model

$$\ln I(s_i) = \ln \mathcal{I}(0) - \frac{1}{3} R_g^2 s_i^2 + \varepsilon_i = \beta_0 + \beta_2 s_i^2 + \varepsilon_i, \tag{27}$$

and use $\widehat{R}_g^2 = -3\widehat{\beta}_2$ as the estimate of R_g^2. An important issue is choosing the size of the cutoff value s_n.

Guinier showed by analysis of simulated data derived from idealized models of particles that the optimal cutoff point depends on the shape of the particle under consideration. From this work follows a rule of thumb for choosing a data window: using an iterative method, accept scattering data only up to the "classical" cutoff s_n^{class} that provides $s_n^{class} \widehat{R}_g = 1.3$ [17, p. 128]. This method is now commonly applied [31, p. 71].

3.2.2. Mean square error of the Guinier estimator. The key innovations of this paper are (a) choosing the data window not by the classical 1.3 rule-of-thumb but by optimizing the window with respect to the mean squared error (MSE) of the estimator and (b) deriving rigorous variance estimators for $\widehat{R}_g^2$ and $\widehat{R}_g$. The theoretical MSE is defined as

$$\mathrm{MSE}_n\left(\widehat{R}_g^2\right) = \mathrm{Var}_n\left(\widehat{R}_g^2\right) + \left\{\mathrm{Bias}_n\left(\widehat{R}_g^2\right)\right\}^2, \tag{28}$$

where $\mathrm{Bias}_n\left(\widehat{R}_g^2\right) = \mathrm{E}\left[\widehat{R}_g^2\right] - R_g^2$. As n increases, the variance $\mathrm{Var}_n\left(\widehat{R}_g^2\right)$ decreases due to increased sample size, but the bias $\mathrm{Bias}_n\left(\widehat{R}_g^2\right)$ increases as the Guinier quadratic approximation (26), valid only near the origin, begins to break down. As n decreases, the variance increases and the bias decreases. Figure 3.1 illustrates this behavior using log intensity versus angle for the nucleosome core particle [1]. Panel A shows the quadratic fit of the data using the cutoff value s_n^{class}, resulting in $n = 18$. This curve fits the data well, so that $\widehat{R}_g^2$

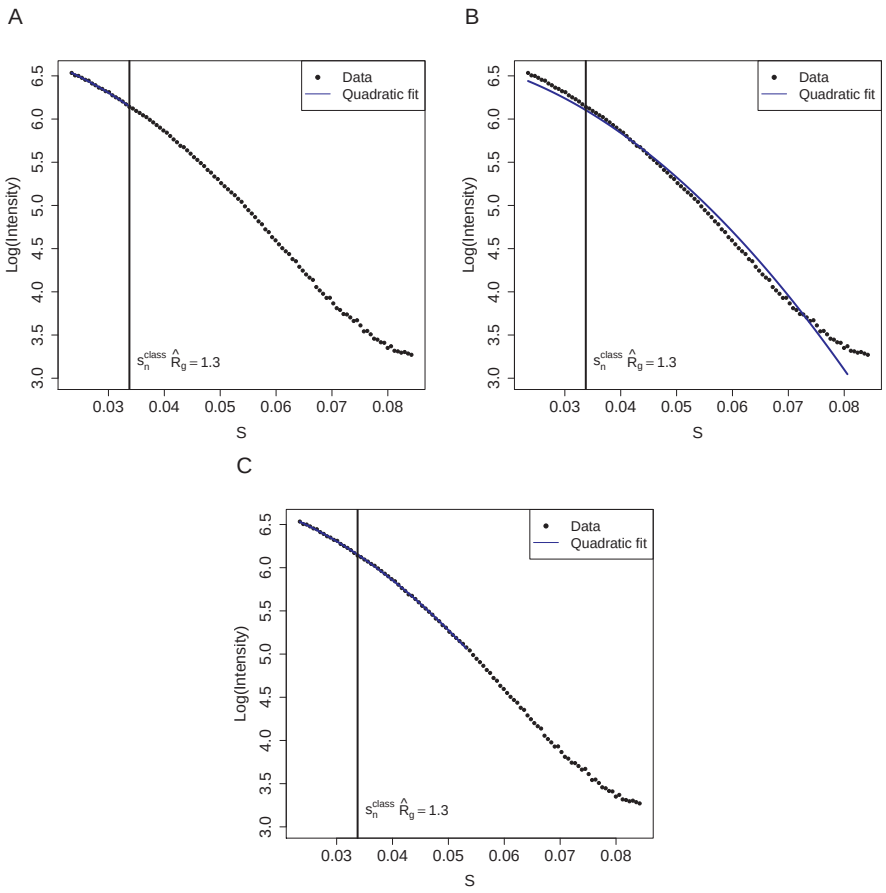


FIGURE 3.1. The number of experimental data points influences the precision and accuracy of $\widehat{R}_g^2$. (**A**) Classical cutoff with $s_n\widehat{R}_g \approx 1.3$ provides $\widehat{R}_g = 44.32$ and $\widehat{\text{Var}}(\widehat{R}_g) = 1.162$. (**B**) Choice of $n = 95$ provides $\widehat{R}_g = 41.10$ and $\widehat{\text{Var}}(\widehat{R}_g) = 0.004$. (**C**) Optimal cutoff s_n that minimizes estimated mean squared error, $\text{MSE}\left(\widehat{R}_g^2\right)$, provides $\widehat{R}_g = 43.75$ and $\widehat{\text{Var}}(\widehat{R}_g) = 0.028$.

will have low bias; however, the curve remains quadratic well past the classical cutoff value. Panel B shows a cutoff value chosen to yield $n = 95$; clearly the quadratic approximation fails for s_n so far from the origin. Panel C of Figure 3.1 shows a cutoff value chosen according to our MSE optimization: it uses much more data ($n = 63$) than the classical cutoff, for greater precision/lower variance, while maintaining a high-quality quadratic approximation, for high accuracy/low bias.

The MSE optimization relies on estimation of the MSE, which in turn relies on estimation of the variance and estimation of the bias in (28). The approach to MSE estimation and optimization will explicitly handle replicate data in an optimal way, without resorting to ad hoc devices like averaging the replicate intensity curves prior to analysis. Let m denote the number of replicate intensity curves used in the analysis. The results hold for general m, including the special case of no replication, $m = 1$. We assume a common angle spacing Δ across replicates, as is common in practice.

3.2.3. Accounting for correlation in Guinier estimation of R_g^2. To better estimate R_g^2 and its variance, we properly account for the correlation structure within a single replicate of the log-intensity data; see [19]. We begin by choosing an initial large data window, by performing a statistical changepoint analysis [36] on the third differences of the log-intensity data. We fit a cubic spline [37] to the log-intensity data over the initial window, obtain residuals $\{r_i\}$ from this fit, and model the residuals as a pth order autoregressive process

$$r_i = \phi_1 r_{i-1} + \cdots + \phi_p r_{i-p} + e_i,$$

where p is selected with Akaike's information criterion (AIC) [38], and $\{e_i\}$ are uncorrelated, with $\mathrm{E}[e_i] = 0$ and $\mathrm{Var}(e_i) = \sigma^2$. We then obtain the parameter estimates $\widehat{\phi}_1, \ldots, \widehat{\phi}_p$ and $\widehat{\sigma}^2$ via Yule-Walker estimation [39]. Using this estimated autoregressive model, we fit the regression (27) via generalized least squares to obtain $\widehat{R}_g^2$.

3.2.4. Estimation of the variance of the Guinier estimator. For m replicates and a given cutoff angle s_n, assuming n data values in each replicate, the variance estimator

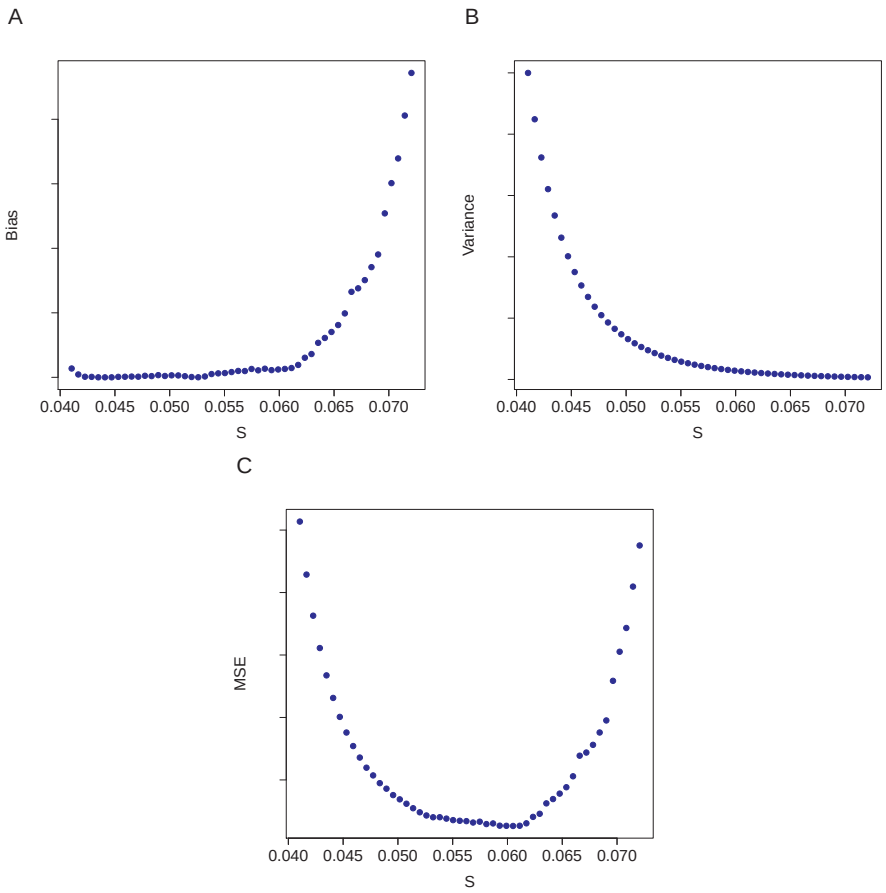


FIGURE 3.2. Minimization of the mean squared error criterion for the nucleosome core particle leads to optimized cutoff value for estimation of R_g in Figure 3.1. (**A**) Estimated bias of $\widehat{R}_g^2$ determined using formula given in (31). (**B**) Estimated variance of $\widehat{R}_g^2$ calculated using formula (29). (**C**) Estimated MSE($\widehat{R}_g^2$).

for $\widehat{R}_g^2$ is

$$\widehat{\mathrm{Var}}_n\left(\widehat{R}_g^2\right) = \frac{405\widehat{\sigma}^2\Delta}{4ms_n^5\left(1-\sum_{j=1}^{p}\widehat{\phi}_j\right)^2}, \tag{29}$$

where Δ is the known spacing between consecutive angles. If the data are uncorrelated within replicates, $\widehat{\phi}_j$ values are all near zero; in practice, the factor $\left(1-\sum_{j=1}^{p}\widehat{\phi}_j\right)^{-2}$ is greater than one, reflecting the smaller information content of positively-correlated observations. The estimated variance decreases with more replicates m and/or a larger cutoff s_n.

Panel B of Figure 3.2 shows the behavior of (29) for a single replicate of experimental SAXS data for the nucleosome core particle over a range of s_n values.

It is also of interest to estimate the variance and standard deviation = (variance)$^{1/2}$ of $\widehat{R}_g$, in order to produce appropriate confidence intervals. A standard delta-method argument [40] shows that an appropriate variance estimator for $\widehat{R}_g$ is

$$\widehat{\mathrm{Var}}_n\left(\widehat{R}_g\right) = \frac{\widehat{\mathrm{Var}}_n\left(\widehat{R}_g^2\right)}{4\widehat{R}_g^2}. \tag{30}$$

3.2.5. Estimation of the bias of the Guinier estimator. The bias estimator begins by extending (27) to an additional term to account for the breakdown of the quadratic approximation (26) as s_n increases and fitting the expanded model

$$\ln I(s_i) = \beta_0 + \beta_2 s_i^2 + \beta_4 s_i^4 + \varepsilon_i$$

via generalized least squares, using the estimated autoregressive model to account for the correlation structure in the data. If the quadratic approximation (26) is good, $\widehat{\beta}_4$ is near zero, and the Guinier estimator is nearly unbiased. The resulting bias estimator is then

$$\widehat{\mathrm{Bias}}_n\left(\widehat{R}_g^2\right) = -\frac{18}{7}\widehat{\beta}_4 s_n^2; \tag{31}$$

As the cutoff s_n increases, the bias increases. The number of replicates, m, does not affect the bias.

Figure 3.2 A shows the behavior of (31), applied to experimental data for the nucleosome core particle. Overall the bias of $\widehat{R}_g^2$ increases as n increases. The increase is gradual at first

but becomes increasingly large. This behavior is readily seen in Figure 3.1. The curve is roughly quadratic at first, but is clearly non-quadratic for larger n values.

3.2.6. Optimal data window selection for the Guinier estimator. We choose s_n to minimize the estimated MSE, obtained by adding (29) and the squared value of (31). Straightforward calculus then shows that the resulting optimal s_n is

$$s_n^{opt} = \left\{ \frac{11025\widehat{\sigma}^2\Delta}{576m\widehat{\beta}_4^2\left(1-\sum_{j=1}^{p}\widehat{\phi}_j\right)^2} \right\}^{1/9}. \tag{32}$$

If the quadratic approximation (26) is good, $\widehat{\beta}_4$ is near zero, and the optimal window is large. The optimal cutoff is small if the number of replicates is large.

The last panel of Figure 3.2 demonstrates determination of the optimum window. This panel is the estimated mean squared error of $\widehat{R}_g^2$ for different values of s_n, and the minimum value of the plot is estimated by (32).

3.2.7. Metrics for evaluation. The performance of the minimum-MSE s_n^{opt} cutoff relative to the classical s_n^{class} cutoff was evaluated through application to nine simulated data scenarios (three molecules, each with one, three, or 15 replicates) and to a data set of two independent wild type preparations and four mutants of the nucleosome core particle, each with 0 mM KCl and 50 mM KCl added. Further detail on the data sets is provided below.

For each of the nine simulated scenarios, the true value of R_g is known; see details below. For both s_n^{class} and s_n^{opt}, we compute the percent relative bias of the estimator $\widehat{R}_g$,

$$\begin{aligned} \%\text{RB} &= \frac{\text{Average}\left(\text{estimator}-\text{truth}\right)}{\text{truth}} \times 100\%, \\ &= \frac{\sum_{i=1}^{M}\left(\widehat{R}_{g(i)}-R_g\right)/M}{R_g} \times 100\%, \end{aligned}$$

where the average is computed over $M = 1000$ simulated iterations. We also compute the root mean squared error (RMSE),

$$\begin{aligned} \text{RMSE} &= \left[\text{Average}\left\{(\text{estimator} - \text{truth})^2\right\}\right]^{1/2} \\ &= \left\{\sum_{i=1}^{M}\left(\widehat{R}_{g(i)} - R_g\right)^2 / M\right\}^{1/2} \end{aligned}$$

for both cutoff values, and summarize with the RMSE ratio

$$\text{RMSE}\left(s_n^{class}\right) / \text{RMSE}\left(s_n^{opt}\right).$$

An RMSE ratio greater than one favors the proposed method.

The true theoretical variance $\text{Var}\left(\widehat{R}_g\right)$ is well-approximated by the empirical variance of the $\widehat{R}_g$ estimates over the $M = 1000$ simulated iterations,

$$S_M^2 = \frac{1}{M-1}\left\{\sum_{i=1}^{M}\widehat{R}_{g(i)}^2 - \frac{1}{M}\left(\sum_{i=1}^{M}\widehat{R}_{g(i)}\right)^2\right\}. \tag{33}$$

We evaluate the variance estimator (30) by computing its value $\widehat{V}_i$ for each iteration i, then averaging over all M iterations,

$$\bar{V}_M = \frac{1}{M}\sum_{i=1}^{M}\widehat{V}_i, \tag{34}$$

and comparing to S_M^2. If the variance estimator is approximately unbiased (approximately correct on average), then $\bar{V}_M \simeq S_M^2$.

The nucleosome mutant data set is used to compare the classical and optimal cutoffs in their respective abilities to discriminate among R_g values in an experimental setting. Let a

denote one of the 12 scenarios (one of the wild type samples, or one of four mutants, either without or with salt added) and let $b \neq a$ denote another scenario; there are 66 such pairs, as shown in Table 3.4. For each such pair, we compute

$$t_{ab} = \frac{\widehat{R}_g(a) - \widehat{R}_g(b)}{\sqrt{\widehat{\mathrm{Var}}_n\left(\widehat{R}_g(a)\right) + \widehat{\mathrm{Var}}_n\left(\widehat{R}_g(b)\right)}}, \tag{35}$$

using (30) in the denominator, and using either the classical cutoff s_n^{class} or the optimal s_n^{opt}. If $|t_{ab}| > 1.96$, we declare a statistically significant difference, then compare results between classical and optimal methods.

3.2.8. Simulation data sets. Simulation studies were conducted to determine the performance of the algorithm compared to a classical procedure when the true R_g value is known. Nine simulation scenarios, consisting of three different molecules at three different replication levels (one, three, or 15), are considered. The three molecules (myoglobin (PDB entry 1WLA) [41], DNA (PDB entry 1BNA) [42], and the nucleosome core particle (PDB entry 1AOI) [1]) have known atomic structures, and so R_g can be computed exactly from their atomic coordinates. The molecules were selected for testing purposes based on their varied nature, size, shape, and R_g values.

For each molecule a theoretical log-intensity curve was calculated from the crystal structure using CRYSOL [28]. One simulation scenario involves generating 1000, 3000, or 15000 log-intensity curves, depending on the number of replicates. Each such simulated replicate is obtained by adding randomly-generated noise to the theoretical log-intensity curve. The randomly-generated noise is independent across replicates, but correlated within replicates,

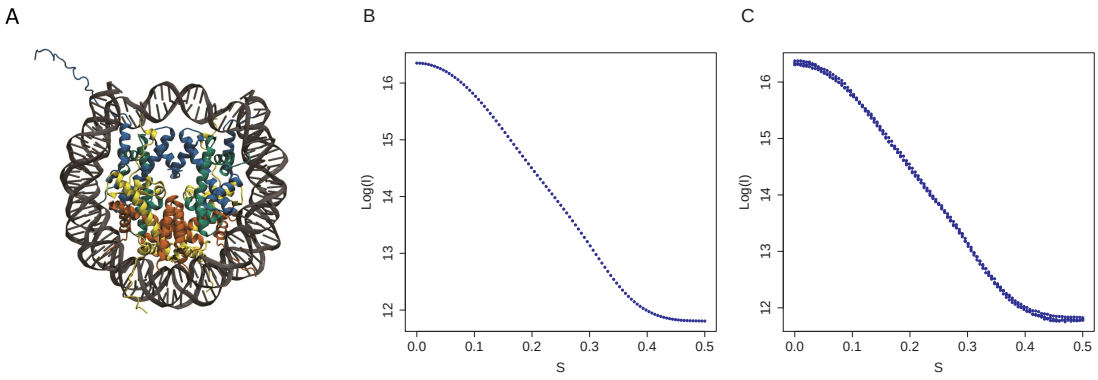


FIGURE 3.3. Illustration of one iteration of the simulation process, generating three simulated replicate intensity curves for the nucleosome; the process is repeated 1000 times to obtain the (Nucleosome, 3 replicates) cell of Table 3.2. (**A**) View of the canonical nucleosome from the crystal structure [1], with the DNA shown in gray, the histones in color. The H3 histone is represented in blue. (**B**) Theoretical log intensity curve derived from the crystal structure. (**C**) One simulated iteration of three simulated replicate log intensity curves, each formed by adding simulated noise to the theoretical log intensity curve.

as observed in real data [19]. Figure 3.3 illustrates this process for one iteration of the nucleosome with three replicates.

3.2.9. EXPERIMENTAL PROCEDURE. To assess amino acid sequence-dependent conformational variability in the nucleosome, [29] conducted an experiment using SAXS data for both wild type and mutant nucleosomes. Nucleosomes are intricate complexes of histones and DNA (Figure 3.3A). The histones contain important modification sites, that, when acetylated or phosphorylated, influence chromatin behavior. By making mutants in the H3 histone (blue in Figure 3.3A) [29] tested the hypothesis that modifications in that area will change the stability and shape of the nucleosome. The sites were chosen to mimic methylation and phosphorylation sites. Nucleosome core particle samples were prepared by previously established methods [43]. Data collection and data processing methods were used that follow previous work [44].

Data sets used for radius of gyration calculations were collected at various sample dilutions and exposure times and the data here reported were consistently derived from $1/16^{\text{th}}$ dilution (with a typical nucleosome concentration of 0.1 mg/ml) and 1 second exposure. The data set contains three replicate curves for each of the two wild types WT, WT2, and the four H3 mutants Y41E, I51A, T45E, and R42A. The samples WT and WT2 are two independent repeats of the full experiment for the same molecule. For all preparations, wild type and mutants, samples either have no salt added to the buffer, or have 50 mM KCl added, leading to 12 scenarios with three replicate intensity curves each. We use these data to compare the ability of the classical method and the optimized method to discriminate among mutant varieties on the basis of $\widehat{R}_g$.

3.3. Results and Discussion

3.3.1. Proposed method dominates classical method in simulation experiments. Tables 3.1 and 3.2 compare the proposed method to the classical method under nine different simulated scenarios, three different molecules at three different replication levels (one, three, or 15) each. In most of the simulated scenarios, our method has a slightly larger percent relative bias than the classical method, but its estimates of R_g are far more precise. Our method outperforms the classical method with respect to root mean squared error as shown in Table 3.2.

Table 3.3 demonstrates the quality of the variance estimator (30). For each simulation, the empirical variance of the 1000 $\widehat{R}_g$ estimates, S_M^2, is approximately equal to the average estimated variances, $\widehat{V}$. Thus the variance estimator (30) is an adequate approximation of the variance of $\widehat{R}_g$.

TABLE 3.1. The classical and optimized algorithms both result in small relative bias. The bias tends to be slightly larger for the proposed algorithm. Three examples of different nature (protein, DNA and protein-DNA complex) were investigated. The bias estimate for either method is relative to the true R_g value (see text). Ideally the bias should be close to zero. Increasing bias values go hand-in-hand with decreasing accuracy for R_g.

Replicates	Method	Percent relative bias		
		Nucleosome	DNA	Myoglobin
1	Classical	-2.44	-0.25	0.36
	Optimized	0.04	-1.66	0.88
3	Classical	-2.33	-0.28	0.35
	Optimized	0.03	-1.24	0.80
15	Classical	-2.25	-0.22	0.28
	Optimized	0.01	-1.05	0.65

TABLE 3.2. Root mean squared error (RMSE) ratio of the classical method to the proposed method indicates that the proposed method outperforms the classical method in simulation studies. The ratio is calculated via RMSE(classical) / RMSE(optimized), with values greater than one favoring the proposed method. The calculations are based on 1000 simulated samples for each combination of molecule and number of replicates ($m = 1$, 3, or 15).

Replicates	RMSE ratio		
	Nucleosome	DNA	Myoglobin
1	14.78	4.05	2.81
3	16.05	3.19	1.95
15	16.79	1.95	1.30

3.3.2. PROPOSED METHOD BETTER DISCRIMINATES AMONG MUTANTS OF THE NUCLEOSOME CORE PARTICLE. By using the optimized method to estimate R_g and its variance, we are able to use SAXS to successfully test the hypothesis that amino acid sequence-dependent conformational variability exists in the nucleosome. With samples of two wild types and four mutants to which either no salt or 50 mM salt were added, there are 12 scenarios and a total of 66 possible pairwise comparisons; see Table 3.4. Among these, at the conventional comparison of two standard deviations (which provides an approximately 95% confidence

TABLE 3.3. The proposed variance estimator (30) is nearly unbiased for the true theoretical variance $\text{Var}\left(\widehat{R}_g\right)$. The theoretical variance is well-approximated by S_M^2 in (33), the empirical variance of the $\widehat{R}_g$ estimates over the $M = 1000$ simulated iterations. The comparison was done for three models of different nature, with one, three or 15 replicates. In each simulated scenario, the average variance estimate, $\bar{V}_M$ from (34), is close to S_M^2.

Replicates		Nucleosome	DNA	Myoglobin
1	S_M^2	0.1492	0.2592	0.1873
	$\bar{V}_M$	0.1518	0.2469	0.1786
3	S_M^2	0.0848	0.1802	0.1202
	$\bar{V}_M$	0.0670	0.1457	0.1037
15	S_M^2	0.0370	0.0963	0.0733
	$\bar{V}_M$	0.0333	0.0957	0.0644

interval), 11 had no statistically significant difference detected by either classical Guinier analysis or the optimized procedure (white background in Table 3.4) and 30 had a significant comparison under both methods (yellow background in Table 3.4). One case led to a significant difference using the classical analysis but not using the optimized procedure (green background), while the remaining 24 cases had a non-significant difference using the classical analysis and a significant difference using the optimized procedure (blue background). In this practical example, the optimized method is superior in its ability to distinguish among different R_g values, leading to a suggestion how the nucleosome changes shape in solution as a consequence of histone mutation.

TABLE 3.4. Pairwise comparisons of R_g values for wild type (WT) and H3 mutant nucleosomes shows superiority of new point and interval estimation method over classical Guinier analysis. (**A**) Wild type and mutants without extra salt in the buffer. The background color of the table entries signify if the pair-wise R_g comparisons are not significant (no color), significant for the new algorithm but not for the old (blue), or significant for both algorithms (yellow). The values in the table are the result of a t-test as defined in equation 35. In each field in this table, the top value is derived from the classical Guinier analysis, the bottom value from the new algorithm. A value greater than 1.96 indicates a statistically significant difference. (**B**) As (A) but samples to which 50 mM KCl was added to the buffer. (**C**) As (A), cross-comparison with 50 mM KCl data in columns and 0 mM KCl data in rows. The green background indicates that a significant difference was detected by conventional algorithm, but not with the optimized algorithm.

A: 0 mM KCl added

WT	WT2	Y41E	I51A	T45E	R42A	
0 0	0.4 0.6	1.2 3.8	2.5 5.1	1.6 6.8	0.6 5.8	WT
	0 0	1.8 3.2	3.5 4.0	2.5 5.7	1.4 4.8	WT2
		0 0	1.2 0.5	0.1 0.9	1.1 0.1	Y41E
			0 0	1.9 4.6	4.9 2.0	I51A
				0 0	2.6 2.2	T45E

B: 50 mM KCl added

WT	WT2	Y41E	I51A	T45E	R42A	
0 0	1.9 1.2	1.9 2.3	3.6 9.5	4.8 11.7	3.4 7.6	WT
	0 0	1.1 1.8	3.1 10.6	6.7 16.7	3.5 9.7	WT2
		0 0	0.1 4.6	0.5 4.6	0.3 2.2	Y41E
			0 0	1.1 1.1	1.1 4.6	H3I
				0 0	3.7 6.2	T45E

C: Cross-comparison of 0 mM (→) and 50 mM KCl (↓)

WT	WT2	Y41E	I51A	T45E	R42A	
0.9 2.7	1.1 2.2	1.4 0.4	3.2 7.1	4.8 8.5	3.0 4.4	WT
0.6 3.2	2.0 2.9	1.7 0.0	4.0 6.4	6.2 7.4	4.1 3.5	WT2
1.9 6.2	0.4 6.4	0.9 2.4	2.2 2.9	3.9 2.6	1.8 0.8	Y41E
3.0 8.5	2.5 12.5	0.5 2.4	1.7 4.6	4.8 7.4	1.1 0.5	I51A
2.3 10.3	0.8 16.0	0.9 3.4	2.7 3.2	6.4 4.4	3.0 3.5	T45E
1.6 9.2	1.3 13.5	1.3 2.9	4.1 3.9	9.9 5.7	6.2 1.8	R42A

3.4. Conclusions

The radius of gyration of biological macromolecules gives insight into the size and shape of the molecule, and can aid in testing hypotheses about the molecular shape. We have developed a semi-automated, statistically sound procedure that estimates the radius of gyration for a molecule and gives a reliable variance estimate. Simulation results show that this estimate of R_g has favorable mean squared error properties compared to the classical method. Furthermore, this method is shown to be more powerful when experimental replicates are present. Results for the procedure applied to experimental data show an improved ability to differentiate R_g values over the classical Guinier method. Therefore implementation of this new procedure will yield more precise estimates of R_g and its variance, enabling improved experimental hypothesis testing.

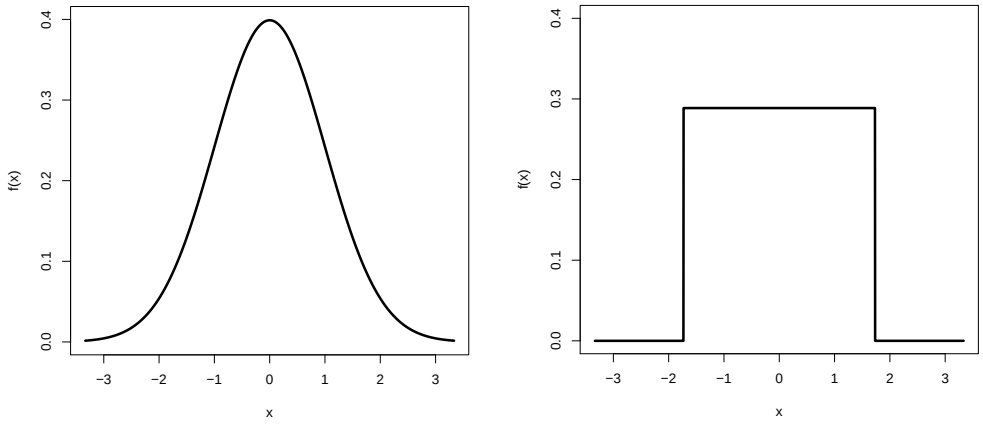
CHAPTER 4

A NEW SHAPE CHARACTERISTIC BASED ON HIGHER-ORDER MOMENTS

4.1. Introduction

Small-angle X-ray scattering (SAXS) is a technique that gives low-resolution information about a molecule in solution. In particular, the radius of gyration (R_g) of a molecule can be estimated using the log intensity curve from a SAXS experiment. Guinier [14] derived an approximation to estimate R_g^2, which is analogous to the second moment of a molecule. The value of R_g^2 contains information regarding a molecule's shape and size. Moreover, two molecules of similar atomic weight but different shape can be distinguished using their R_g^2 values. We extend this idea to estimate the fourth moment, denoted by M^4, of a molecule in order to distinguish subtle differences in molecules' shapes and sizes.

This idea relates to the moments of probability distributions. Consider the two distributions in Figure 4.1. These two distributions both have the same first and second moment; however, they have very different shapes. Their fourth moment values can help distinguish their shapes. The fourth moment for the normal distribution on the left is 3 and the fourth moment for the uniform distribution on the right is 1.8. Our goal is to extend this idea to SAXS data in order to differentiate between two molecules with similar R_g values but different fourth moment values.



(A) Normal (0,1) distribution. (B) Uniform $(-\sqrt{3}, \sqrt{3})$ distribution.

FIGURE 4.1. Normal (0,1) distribution (left) and Uniform $(-\sqrt{3}, \sqrt{3})$ distribution (right). Each distribution has mean zero and variance one. However, the fourth moment of the normal distribution is 3 and the fourth moment of the uniform distribution is 1.8.

4.2. Definition of ψ

In essence, R_g describes the mass spread present in a molecule. Since the variance (also known as the second moment) of a probability distribution is also a measure of spread, R_g^2 can be thought of as an analog to variance.

We define R_g using the function $p(r)$, which is the distribution of the distances r between all pairwise points of a molecule:

$$R_g^2 = \frac{\int_0^{D_{max}} r^2 p(r)\, dr}{2\pi \int_0^{D_{max}} p(r)\, dr}, \tag{36}$$

where D_{max} is the maximum pairwise distance in the molecule. We further define the new quantity

$$M^4 = \frac{\int_0^{D_{max}} r^4 p(r)\, dr}{2\pi \int_0^{D_{max}} p(r)\, dr}. \tag{37}$$

Just as R_g^2 is analogous to the second moment of a probability distribution, M^4 is the analog of the fourth moment of a distribution. Very roughly, the fourth moment provides information about how much area is contained in the tails of the distribution. Likewise, M^4 provides information about how much mass is contained in the regions of a molecule farthest from the center of mass. For example, a rod-shaped molecule has a larger M^4 value than a spherical molecule, assuming both molecules have the same R_g value.

We wish to use these quantities to differentiate molecules based purely on shape, but both R_g and M^4 are dependent on molecular size. Therefore, we define a new dimensionless ratio ψ given by

$$\psi = \frac{M^4}{R_g^4},$$

which contains information concerning molecular shape but is independent of size. Furthermore, ψ can be estimated from experimental SAXS data using an extension of Guinier analysis.

4.3. Calculating ψ for Geometric Shapes

We first determine the ψ value for cylinders of varying height/radius ratios, and then we use this information to estimate the shape of a molecule using its ψ value. The quantity ψ for a cylinder depends solely on the cylinder's height/radius ratio, and this relationship is illustrated by the curve in Figure 4.2. The plot shows ψ versus increasing height/radius ratios of cylinders. The ψ value for a cylinder with a fixed height/radius ratio is independent of its size. Furthermore, as the height/radius ratio increases ψ will converge to 4.8 since this is the ψ value for a line (infinitely long cylinder on finite radius) of any length. A cylinder of height zero is a circle, and its ψ value is 10/3, which is independent of the circle's radius.

The minimum value of the curve in Figure 4.2 has a ψ value of 2.89 and a height/radius ratio of approximately 1.7.

Given a ψ value we wish to estimate a cylinder's height/radius ratio. However, for values of ψ in the range 2.8 to 3.3 there are two possible height/radius ratios as can be seen in Figure 4.2. Therefore, we have an identifiability problem for values in this range.

The plot in Figure 4.2 shows the ψ versus aspect ratio (H/R) curve for ellipsoids and rectangles. The black curve represents varying cylinders, the dashed red curve represents varying ellipsoids, and the dotted blue curve represents varying rectangles. For the larger values on the curve, ψ is larger for ellipsoids and smaller for rectangles compared to cylinders. The ellipsoid curve touches the line 20/7 at height/radius = 2 since this is the exact ψ value for a sphere. The ψ value for the rectangle curve, the dotted blue line, converges to 3.4 as its height goes to zero. This is the ψ value for a square. The ψ value for the ellipsoid curve converges to 10/3 as its height goes to zero since, like a cylinder, it converges to a circle. For the rest of this paper we will focus on ψ values for cylinders, but a different shape could be used instead if desired.

4.4. Estimating the Height/Radius Ratio for a Molecule Given its Atomic Structure

We use principal component analysis (PCA) to determine the aspect ratio of the best fitting cylinder for a molecule, given its atomic structure [45]. Let $\mathbf{X}$ be the $n \times 3$ matrix containing the 3-dimensional coordinates for a molecule with n atoms, and let $\mathbf{S}$ be the 3×3 covariance matrix for $\mathbf{X}$. From principal component analysis theory, the eigenvalues and eigenvectors of $\mathbf{S}$ contain the spread and direction of most variability in $\mathbf{X}$. Let λ_1, λ_2, and λ_3 be the largest, second-largest, and smallest eigenvalues of $\mathbf{S}$, respectively, and let

$\mathbf{v_1}, \mathbf{v_2}$, and $\mathbf{v_3}$ be the corresponding mutually orthogonal eigenvectors. Then, the direction of largest variability in $\mathbf{X}$ is $\mathbf{v_1}$ with relative length $\sqrt{\lambda_1}$, and the direction of second-largest variability in $\mathbf{X}$ is $\mathbf{v_2}$ with relative length $\sqrt{\lambda_2}$.

For molecules with $\frac{20}{7} < \psi \leq \frac{10}{3}$, the method yields two good-fitting cylinders. A similar ambiguity occurs when estimating the height/radius (H/R) ratio using PCA. There are two alternate approximations that may be used to find the H/R ratio of a molecule:

$$\frac{H}{R} \approx 2\sqrt{\frac{\lambda_1}{\lambda_2}}, \quad \text{and} \quad \frac{H}{R} \approx 2\sqrt{\frac{\lambda_3}{\lambda_1}}.$$

In the first approximation, it is assumed that the direction of greatest variability corresponds to the cylinder's height, so the resulting cylinder has a height greater than its diameter. In the second approximation, the opposite is true: the direction of least variability corresponds to height, resulting in a cylinder with height smaller than diameter. We therefore select the approximation that best represents the actual shape of the molecule.

4.5. ψ Plot for Molecules

For a molecule of arbitrary shape with known $p(r)$ function, its height/radius ratio can be approximated using principal component analysis, and ψ can be computed using the previously discussed methods. Figure 4.2 contains a plot of ψ versus height/radius ratio for 3,430 molecules; the atomic structures for these molecules were obtained from the database at `PDB.org`. The black curve is again ψ versus height/radius ratio for cylinders of various dimensions, which one can see provides a good fit for the molecular data, with coefficient of determination 0.91. Therefore, on average, cylinders yield an adequate fit for the molecules.

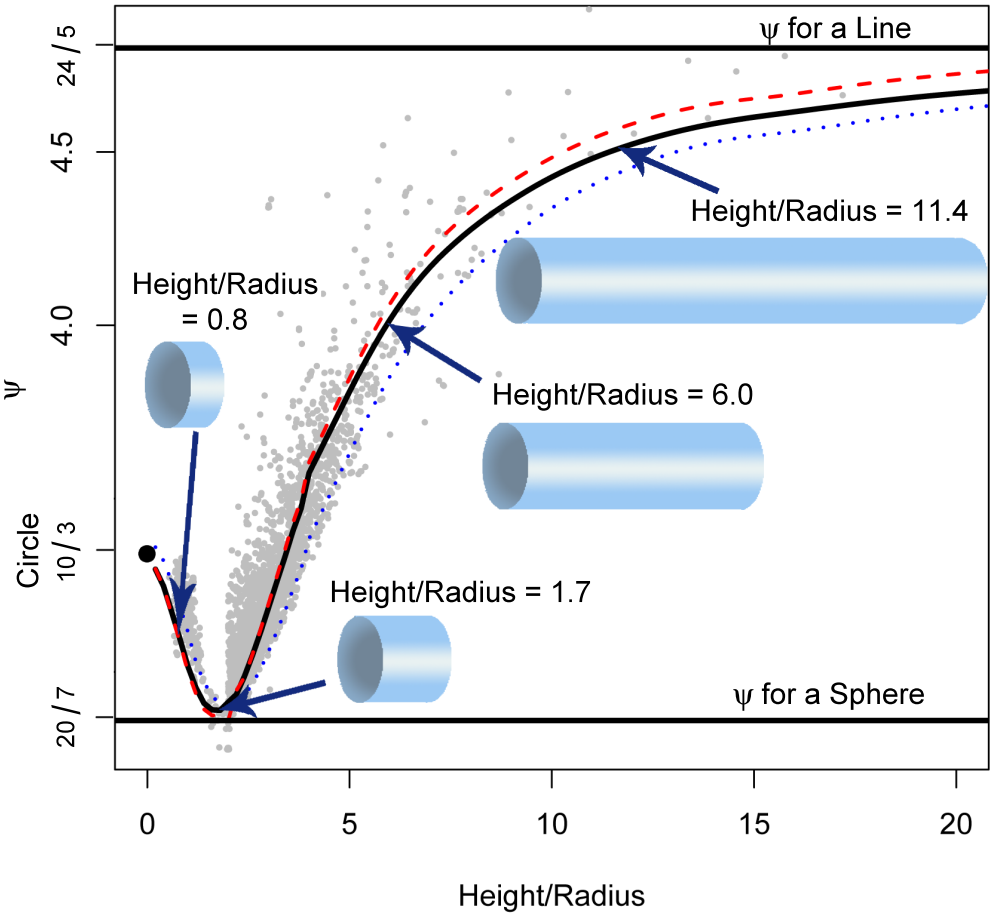


FIGURE 4.2. Plot of ϕ versus aspect ratio for three shapes along with examples of cylinders that fit different height/radius ratios. The gray points are the estimated height/radius ratio calculated using principal component analysis versus the exact ψ value determined from the atomic structure of theoretical molecules. The black represents varying cylinders, the dashed red curve represents varying ellipsoids, and the dotted blue curve represents varying rectangles. Selected cylinders of different height/radius ratios are also given on the plot.

From the atomic structure of the molecule, we calculate ψ, and using this value with Figure 4.2 we determine a height/radius ratio for the cylinder corresponding to the molecule. Additionally, from the molecule's atomic structure we estimate the height/radius ratio of a cylinder using principal component analysis. A plot of H/R estimated from principal component analysis versus H/R estimated via cylinder fitting along with the identity line is

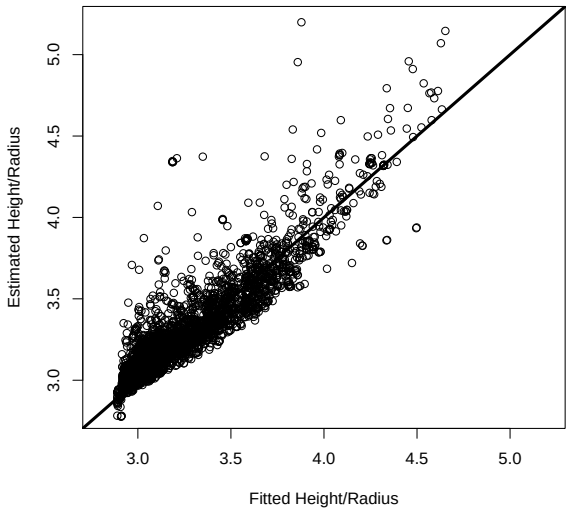


FIGURE 4.3. Plot of H/R estimated from principal component analysis versus H/R estimated via cylinder fitting, along with the identity line.

shown in Figure 4.3. This plot illustrates the high degree of correlation between these two methods of fitting a cylinder to the molecule.

Using the atomic structure of the molecules in Figure 4.2, we calculate the true D_{max} value for each molecule. Furthermore, from the atomic structure we calculate ψ and R_g and using (59) along with Figure 4.2 we determine a good-fitting cylinder for each molecule. Then, we estimate D_{max} for each molecule from this cylinder. A plot of true D_{max} versus estimated D_{max} from this cylinder fitting process along with the identity line is given in Figure 4.4. The correlation for these two values is 0.98. Therefore, the fitted cylinder give a good estimate of the D_{max} value for a molecule.

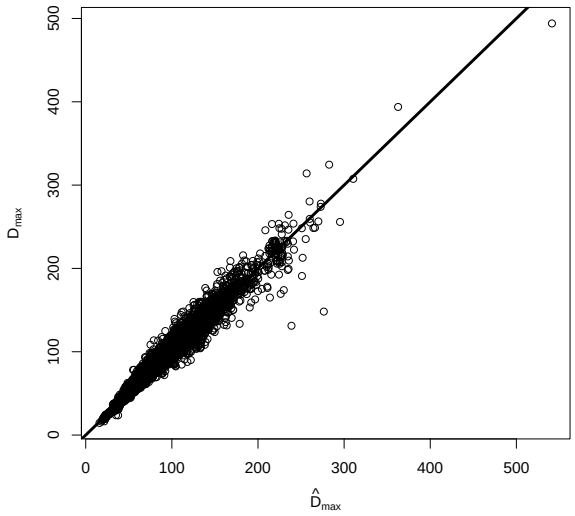


FIGURE 4.4. Plot of theoretical D_{max} for molecules calculated from their atomic structure versus their D_{max} value estimate using the cylinder fitting.

4.6. DISCUSSION

We develop a new molecular shape parameter, ψ, that gives the aspect ratio of a molecule. Previous work has been able to estimate a molecule's radius of gyration; however, as shown show in several examples two molecules can have similar radius of gyration values but vastly different shape. Some of this different shape can be accounted for comparing the two molecules' fourth moment values. The fourth moment and the radius of gyration is related to the parameter ψ. This new parameter can be used to determine a low-resolution shape and size of the molecule in solution.

CHAPTER 5

STATISTICAL INFERENCE FOR THE ASPECT RATIO VIA HIGHER-ORDER GUINIER ANALYSIS

5.1. Introduction

Guinier derived an equation relating R_g and the scattering curve:

$$\ln \mathcal{I}(s) = \ln \mathcal{I}(0) - \frac{1}{3} R_g^2 s^2 + \mathrm{O}\left(s^4\right). \tag{38}$$

This equation can be used along with quadratic regression to estimate R_g from a molecule's experimental SAXS intensity curve. To relate the SAXS data to both R_g and the quantity M^4, we extend (38) to include an extra term, resulting in a higher-order (and more accurate for small s) approximation:

$$\ln \mathcal{I}(s) = \ln \mathcal{I}(0) - \frac{1}{3} R_g^2 s^2 + \left(\frac{1}{60} M^4 - \frac{1}{18} R_g^4\right) s^4 + \mathrm{O}\left(s^6\right). \tag{39}$$

Using this equation, R_g and M^4 can be estimated directly from experimental SAXS data for a molecule; the ratio ψ can then be obtained easily.

5.2. Estimating R_g^2 and M^4

To estimate R_g^2 and M^4, consider the model written as

$$\ln I(s_i) = \beta_0 + \beta_2 s_i^2 + \beta_4 s_i^4 + \varepsilon_i, \tag{40}$$

where $i = 1, \ldots, n$. We can estimate R_g^2 and M^4 using the relation

$$\begin{aligned} \ln \mathcal{I}(s_i) &= \ln \mathcal{I}(0) - \frac{1}{3} R_g^2 s^2 + \frac{1}{60} M^4 s^4 - \frac{1}{18} R_g^4 s^4 + \mathrm{O}\left(s^6\right) \\ &= \ln \mathcal{I}(0) - \frac{1}{3} R_g^2 s^2 + \left(\frac{1}{60} M^4 - \frac{1}{18} R_g^4 \right) s^4 + \mathrm{O}\left(s^6\right). \end{aligned} \tag{41}$$

Therefore, we need to determine the relationship between the β_i's and R_g^2 and M^4. Equating the coefficients in the formulas (40) and (41) yields the approximations for R_g^2 and M^4. To estimate R_g^2 and M^4, we fit (40) and use the approximations given by

$$\widehat{R}_g^2 = -3\widehat{\beta}_2 \quad \text{and} \tag{42}$$

$$\widehat{M}^4 = 60\widehat{\beta}_4 + 30\widehat{\beta}_2^2. \tag{43}$$

5.3. Minimizing the MSE of $\widehat{\psi}$ With AR(p) Errors

First, we calculate the MSE of $\widehat{\beta}_2$ and $\widehat{\beta}_4$. Let $f(\cdot)$ be the true log intensity curve. The empirical log intensity curve values are $Y_1, \ldots, Y_n$ measured at the angles $s_1, \ldots, s_n$. Furthermore, let

$$s_i = i\Delta,$$

where Δ is the spacing between points. We consider an asymptotic formulation in which $n \to \infty$ with $\Delta \to 0$, and the goal is to minimize MSE($\widehat{\psi}$).

Thus, the model for this problem is given by

$$Y_i = f(s_i) + \varepsilon_i,$$

where $\mathrm{E}[\varepsilon_i] = 0$ and $i = 1, \ldots, n$. Therefore, we have $\mathrm{E}[Y_i] = f(s_i)$ and the errors $\{\varepsilon_i\}$ follow a causal pth order autoregressive process, AR(p), with covariance matrix $\mathbf{\Gamma}$. We fit the model

$$Y_i = \beta_0 + \beta_2 s_i^2 + \beta_4 s_i^4 + \varepsilon_i$$

to the observed data using generalized least squares. The estimator of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{\Gamma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{\Gamma}^{-1}\mathbf{Y},$$

where $\mathbf{Y} = [Y_1, \ldots, Y_n]'$ and

$$\mathbf{X} = \begin{bmatrix} 1 & s_1^2 & s_1^4 \\ \vdots & \vdots & \vdots \\ 1 & s_n^2 & s_n^4 \end{bmatrix}$$

is an $n \times 3$ matrix. Then, the estimator of β_2 is

$$\widehat{\beta}_2 = \mathbf{e_2}'\left(\mathbf{X}'\mathbf{\Gamma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{\Gamma}^{-1}\mathbf{Y},$$

where $\mathbf{e_2}$ is the 3×1 vector given by $\mathbf{e_2} = [0, 1, 0]'$ and the estimator of β_4 is

$$\widehat{\beta}_4 = \mathbf{e_4}'\left(\mathbf{X}'\mathbf{\Gamma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{\Gamma}^{-1}\mathbf{Y},$$

where $\mathbf{e_4}$ is the 3×1 vector given by $\mathbf{e_4} = [0, 0, 1]'$. Next let $\mathbf{M} = [f(s_1), f(s_2), \ldots, f(s_n)]'$. Thus, the expected value of $\widehat{\boldsymbol{\beta}}$ is given by

$$\mathrm{E}\widehat{\boldsymbol{\beta}} = \mathrm{E}\left\{\left(\mathbf{X}'\mathbf{\Gamma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{\Gamma}^{-1}\mathbf{Y}\right\}$$

$$= \left(\mathbf{X}'\mathbf{\Gamma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{\Gamma}^{-1}\mathrm{E}(\mathbf{Y})$$

$$= \left(\mathbf{X}'\mathbf{\Gamma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{\Gamma}^{-1}\mathbf{M}. \tag{44}$$

Now, we employ the Taylor series of $f(s_i)$ to rewrite (44). Since the log intensity curve can be written as a polynomial containing only even powers of s, its Taylor series centered at 0 is given by

$$f(s_i) = f(0) + \frac{f''(0)}{2}s_i^2 + \frac{f^{(4)}(0)}{24}s_i^4 + \frac{f^{(6)}(0)}{720}s_i^6 + \mathrm{O}\left(s^8\right).$$

Hence, $\mathbf{M}$ can be written as

$$\mathbf{M} = \mathbf{X}\begin{bmatrix} f(0) \\ \frac{1}{2}f''(0) \\ \frac{1}{24}f^{(4)}(0) \end{bmatrix} + \frac{f^{(6)}(0)}{720}\begin{bmatrix} s_1^6 \\ \vdots \\ s_n^6 \end{bmatrix} + \mathrm{O}\left(s^8\right).$$

Therefore, the first term in this expansion of $\mathrm{E}\widehat{\boldsymbol{\beta}}$ is

$$\mathbf{e}_i'\left(\mathbf{X}'\mathbf{\Gamma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{\Gamma}^{-1}\mathbf{X}\begin{bmatrix} f(0) \\ \frac{1}{2}f''(0) \\ \frac{1}{24}f^{(4)}(0) \end{bmatrix} = \mathbf{e}_i'\begin{bmatrix} f(0) \\ \frac{1}{2}f''(0) \\ \frac{1}{24}f^{(4)}(0) \end{bmatrix} = \beta_i$$

for $i = 2, 4$. Thus, the bias of $\widehat{\beta}_i$ is given by

$$\mathrm{E}\left(\widehat{\beta}_i - \beta_i\right) = \frac{f^{(6)}(0)}{720}\mathbf{e}_i'\left(\mathbf{X}'\mathbf{\Gamma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{\Gamma}^{-1}\begin{bmatrix} s_1^6 \\ \vdots \\ s_n^6 \end{bmatrix} + \mathrm{O}\left(s^8\right).$$

Therefore, if $f(s) = \beta_0 + \beta_2 s^2 + \beta_4 s^4$ for some constants $\beta_0, \beta_2, \beta_4 \in \mathbb{R}$, then $\widehat{\beta}_i$ is unbiased. In order to compute the leading term in the bias of $\widehat{\beta}_i$ we use the following lemmas.

The first lemma is the Cholesky decomposition of the covariance matrix $\mathbf{\Gamma}$ for the autoregressive error process (see [25] §8.6 for more details). This lemma is used to determine the inverse of $\mathbf{\Gamma}$ in future computations.

Consider the Cholesky decomposition for the matrix $\mathbf{\Gamma}$,

$$\mathbf{T}\mathbf{\Gamma}\mathbf{T}' = \mathbf{D}, \tag{45}$$

where $\mathbf{T}'$ is an upper triangular matrix given by

$$\mathbf{T}' = \begin{bmatrix} 1 & -\phi_1 & -\phi_2 & -\phi_3 & \dots & -\phi_p & 0 & \dots & 0 \\ 0 & 1 & -\phi_1 & -\phi_2 & \dots & -\phi_{p-1} & -\phi_p & \dots & 0 \\ 0 & 0 & 1 & -\phi_1 & \dots & -\phi_{p-2} & \dots & & \vdots \\ \vdots & \vdots & \vdots & \ddots & & & & & \\ 0 & 0 & 0 & \dots & & & & & 1 \end{bmatrix}$$

and $\mathbf{D} = \mathrm{diag}(\sigma^2, \dots, \sigma^2)$.

LEMMA 5. *Using the Cholesky decomposition (45), we can write*

$$n^{-1}\mathbf{X}'\mathbf{\Gamma}^{-1}\mathbf{X} = n^{-1}\mathbf{X}'\mathbf{T}'\mathbf{D}^{-1}\mathbf{T}\mathbf{X} = \frac{1}{\sigma^2}\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix},$$

where

$$a_{11} = \frac{n-p}{n}\left(1 - \sum_{j=1}^{p}\phi_j\right)^2 + n^{-1}\left\{1 + (1-\phi_1)^2 + \cdots + (1-\phi_1 - \cdots - \phi_{p-1})^2\right\}$$

$$a_{12} = a_{21} = n^{-1}\left(1 - \sum_{j=1}^{p}\phi_j\right)\sum_{i=p+1}^{n}\left\{i^2\Delta^2 - \sum_{j=1}^{p}\phi_j\,(i-j)^2\,\Delta^2\right\}$$

$$+ n^{-1}\{s_1^2 + (1-\phi_1)\left(s_2^2 - \phi_1 s_1^2\right) + \cdots +$$

$$(1-\phi_1 - \cdots - \phi_{p-1})\left(s_p^2 - \phi_1 s_{p-1}^2 - \cdots - \phi_{p-1}s_1^2\right)\}$$

$$a_{13} = a_{31} = n^{-1}\left(1 - \sum_{j=1}^{p}\phi_j\right)\sum_{i=p+1}^{n}\left\{i^4\Delta^4 - \sum_{j=1}^{p}\phi_j\,(i-j)^4\,\Delta^4\right\}$$

$$+ n^{-1}\{s_1^4 + (1-\phi_1)\left(s_2^4 - \phi_1 s_1^4\right) + \ldots$$

$$+ (1-\phi_1 - \cdots - \phi_{p-1})\left(s_p^4 - \phi_1 s_{p-1}^4 - \cdots - \phi_{p-1}s_1^4\right)\}$$

$$a_{22} = n^{-1}\sum_{i=p+1}^{n}\left\{i^2\Delta^2 - \sum_{j=1}^{p}\phi_j\,(i-j)^2\,\Delta^2\right\}^2$$

$$+ n^{-1}\left\{s_1^4 + \left(s_2^2 - \phi_1 s_1^2\right)^2 + \cdots + \left(s_p^2 - \phi_1 s_{p-1}^2 - \cdots - \phi_{p-1}s_1^2\right)^2\right\}$$

$$a_{23} = a_{32} = n^{-1} \sum_{i=p+1}^{n} \left\{ i^2\Delta^2 - \sum_{j=1}^{p} \phi_j (i-j)^2 \Delta^2 \right\} \left\{ i^4\Delta^4 - \sum_{j=1}^{p} \phi_j (i-j)^4 \Delta^4 \right\}$$

$$+ n^{-1}\{s_1^6 + \left(s_2^2 - \phi_1 s_1^2\right)\left(s_2^4 - \phi_1 s_1^4\right) + \dots$$

$$+ \left(s_p^2 - \phi_1 s_{p-1}^2 - \cdots - \phi_{p-1} s_1^2\right)\left(s_p^4 - \phi_1 s_{p-1}^4 - \cdots - \phi_{p-1} s_1^4\right)\}$$

$$a_{33} = n^{-1} \sum_{i=p+1}^{n} \left\{ i^4\Delta^4 - \sum_{j=1}^{p} \phi_j (i-j)^4 \Delta^4 \right\}^2$$

$$+ n^{-1} \left\{ s_1^8 + \left(s_2^4 - \phi_1 s_1^4\right)^2 + \cdots + \left(s_p^4 - \phi_1 s_{p-1}^4 - \cdots - \phi_{p-1} s_1^4\right)^2 \right\}.$$

PROOF. Applying the Cholesky decomposition (45) of the matrix $\mathbf{\Gamma}$, we have

$$\mathbf{X}'\mathbf{\Gamma}^{-1}\mathbf{X} = \mathbf{X}'\mathbf{T}'\mathbf{D}^{-1}\mathbf{T}\mathbf{X}.$$

First, we have

$$\mathbf{X}'\mathbf{T}' = \begin{bmatrix} 1 & 1-\phi_1 & \dots & 1-\sum_{j=1}^{p}\phi_j & \dots & 1-\sum_{j=1}^{p}\phi_j \\ s_1^2 & s_2^2-\phi_1^2 & \dots & s_{p+1}^2-\sum_{j=1}^{p}\phi_j s_{p+1-j}^2 & \dots & s_n^2-\sum_{j=1}^{p}\phi_j s_{n-j}^2 \\ s_1^4 & s_2^4-\phi_1^2 & \dots & s_{p+1}^4-\sum_{j=1}^{p}\phi_j s_{p+1-j}^4 & \dots & s_n^4-\sum_{j=1}^{p}\phi_j s_{n-j}^4 \end{bmatrix}. \tag{46}$$

Substituting $s_i = i\Delta$, we can write $n^{-1}\mathbf{X}'\mathbf{T}'\mathbf{D}^{-1}\mathbf{T}\mathbf{X}$ as

$$\frac{1}{\sigma^2} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix},$$

where

$$a_{11} = \frac{n-p}{n}\left(1-\sum_{j=1}^{p}\phi_j\right)^2 + n^{-1}\left\{1+(1-\phi_1)^2+\cdots+(1-\phi_1-\cdots-\phi_{p-1})^2\right\}$$

$$a_{12} = a_{21} = n^{-1}\left(1-\sum_{j=1}^{p}\phi_j\right)\sum_{i=p+1}^{n}\left\{i^2\Delta^2-\sum_{j=1}^{p}\phi_j(i-j)^2\Delta^2\right\}$$

$$+ n^{-1}\{s_1^2+(1-\phi_1)\left(s_2^2-\phi_1 s_1^2\right)+\ldots$$

$$+ (1-\phi_1-\cdots-\phi_{p-1})\left(s_p^2-\phi_1 s_{p-1}^2-\cdots-\phi_{p-1}s_1^2\right)\}$$

$$a_{13} = a_{31} = n^{-1}\left(1-\sum_{j=1}^{p}\phi_j\right)\sum_{i=p+1}^{n}\left\{i^4\Delta^4-\sum_{j=1}^{p}\phi_j(i-j)^4\Delta^4\right\}$$

$$+ n^{-1}\{s_1^4+(1-\phi_1)\left(s_2^4-\phi_1 s_1^4\right)+\ldots$$

$$+ (1-\phi_1-\cdots-\phi_{p-1})\left(s_p^4-\phi_1 s_{p-1}^4-\cdots-\phi_{p-1}s_1^4\right)\}$$

$$a_{22} = n^{-1}\sum_{i=p+1}^{n}\left\{i^2\Delta^2-\sum_{j=1}^{p}\phi_j(i-j)^2\Delta^2\right\}^2$$

$$+ n^{-1}\left\{s_1^4+\left(s_2^2-\phi_1 s_1^2\right)^2+\cdots+\left(s_p^2-\phi_1 s_{p-1}^2-\cdots-\phi_{p-1}s_1^2\right)^2\right\}$$

$$a_{23} = a_{32} = n^{-1}\sum_{i=p+1}^{n}\left\{i^2\Delta^2-\sum_{j=1}^{p}\phi_j(i-j)^2\Delta^2\right\}\left\{i^4\Delta^4-\sum_{j=1}^{p}\phi_j(i-j)^4\Delta^4\right\}$$

$$+ n^{-1}\{s_1^6+\left(s_2^2-\phi_1 s_1^2\right)\left(s_2^4-\phi_1 s_1^4\right)+\ldots$$

$$+ \left(s_p^2-\phi_1 s_{p-1}^2-\cdots-\phi_{p-1}s_1^2\right)\left(s_p^4-\phi_1 s_{p-1}^4-\cdots-\phi_{p-1}s_1^4\right)\}$$

$$a_{33} = n^{-1} \sum_{i=p+1}^{n} \left\{ i^4\Delta^4 - \sum_{j=1}^{p} \phi_j \left(i-j\right)^4 \Delta^4 \right\}^2$$

$$+ n^{-1} \left\{ s_1^8 + \left(s_2^4 - \phi_1 s_1^4\right)^2 + \cdots + \left(s_p^4 - \phi_1 s_{p-1}^4 - \cdots - \phi_{p-1} s_1^4\right)^2 \right\},$$

proving the result. □

LEMMA 6. *Using the Cholesky decomposition (45),*

$$n^{-1}\mathbf{X}'\mathbf{\Gamma}^{-1} \begin{bmatrix} s_1^6 \\ \vdots \\ s_n^6 \end{bmatrix} = n^{-1}\mathbf{X}'\mathbf{T}'\mathbf{D}^{-1}\mathbf{T} \begin{bmatrix} s_1^6 \\ \vdots \\ s_n^6 \end{bmatrix} = \frac{1}{\sigma^2} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix},$$

where

$$b_1 = n^{-1} \left(1 - \sum_{j=1}^{p} \phi_j\right) \sum_{i=p+1}^{n} \left(i^6\Delta^6 - \sum_{j=1}^{p} \phi_j \left(i-j\right)^6 \Delta^6 \right)$$

$$+ n^{-1}\{ s_1^6 + (1-\phi_1)\left(s_2^6 - \phi_1 s_1^6\right) + \ldots$$

$$+ (1 - \phi_1 - \cdots - \phi_{p-1})\left(s_p^6 - \phi_1 s_{p-1}^6 - \cdots - \phi_{p-1} s_1^6\right)\}$$

$$b_2 = n^{-1} \sum_{i=p+1}^{n} \left(i^2\Delta^2 - \sum_{j=1}^{p} \phi_j \left(i-j\right)^2 \Delta^2 \right) \left(i^6\Delta^6 - \sum_{j=1}^{p} \phi_j \left(i-j\right)^6 \Delta^6 \right)$$

$$+ n^{-1}\{ s_1^8 + \left(s_2^2 - \phi_1 s_1^2\right)\left(s_2^6 - \phi_1 s_1^6\right) + \ldots$$

$$+ \left(s_p^2 - \phi_1 s_{p-1}^2 - \cdots - \phi_{p-1} s_1^2\right)\left(s_p^6 - \phi_1 s_{p-1}^6 - \cdots - \phi_{p-1} s_1^6\right)\}$$

$$b_3 = n^{-1} \sum_{i=p+1}^{n} \left(i^4\Delta^4 - \sum_{j=1}^{p} \phi_j \left(i-j\right)^4 \Delta^4 \right) \left(i^6\Delta^6 - \sum_{j=1}^{p} \phi_j \left(i-j\right)^6 \Delta^6 \right)$$

$$+ n^{-1}\{s_1^{10} + \left(s_2^4 - \phi_1 s_1^4\right)\left(s_2^6 - \phi_1 s_1^6\right) + \ldots$$

$$+ \left(s_p^4 - \phi_1 s_{p-1}^4 - \cdots - \phi_{p-1} s_1^4\right)\left(s_p^6 - \phi_1 s_{p-1}^6 - \cdots - \phi_{p-1} s_1^6\right)\}.$$

PROOF. By (46), we have

$$n^{-1}\mathbf{X'T'D^{-1}T} \begin{bmatrix} s_1^6 \\ \vdots \\ s_n^6 \end{bmatrix} = n^{-1}\mathbf{X'T'D^{-1}} \begin{bmatrix} s_1^6 \\ s_2^6 - s_1^6\phi_1 \\ \vdots \\ s_{p+1}^6 - \sum_{j=1}^{p} \phi_j s_{p+1-j}^6 \\ \vdots \\ s_n^6 - \sum_{j=1}^{p} \phi_j s_{n-j}^6 \end{bmatrix} = \frac{1}{\sigma^2} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix},$$

where

$$b_1 = n^{-1}\left(1 - \sum_{j=1}^{p} \phi_j\right) \sum_{i=p+1}^{n} \left(i^6\Delta^6 - \sum_{j=1}^{p} \phi_j \left(i-j\right)^6 \Delta^6 \right)$$

$$+ n^{-1}\{s_1^6 + (1-\phi_1)\left(s_2^6 - \phi_1 s_1^6\right) + \ldots$$

$$+ (1 - \phi_1 - \cdots - \phi_{p-1})\left(s_p^6 - \phi_1 s_{p-1}^6 - \cdots - \phi_{p-1} s_1^6\right)\}$$

$$b_2 = n^{-1} \sum_{i=p+1}^{n} \left(i^2\Delta^2 - \sum_{j=1}^{p} \phi_j \left(i-j\right)^2 \Delta^2 \right) \left(i^6\Delta^6 - \sum_{j=1}^{p} \phi_j \left(i-j\right)^6 \Delta^6 \right)$$

$$+ n^{-1}\{s_1^8 + \left(s_2^2 - \phi_1 s_1^2\right)\left(s_2^6 - \phi_1 s_1^6\right) + \ldots$$

$$+ \left(s_p^2 - \phi_1 s_{p-1}^2 - \cdots - \phi_{p-1} s_1^2\right)\left(s_p^6 - \phi_1 s_{p-1}^6 - \cdots - \phi_{p-1} s_1^6\right)\}$$

$$b_3 = n^{-1} \sum_{i=p+1}^{n} \left(i^4\Delta^4 - \sum_{j=1}^{p} \phi_j \left(i-j\right)^4 \Delta^4\right)\left(i^6\Delta^6 - \sum_{j=1}^{p} \phi_j \left(i-j\right)^6 \Delta^6\right)$$

$$+ n^{-1}\{s_1^{10} + \left(s_2^4 - \phi_1 s_1^4\right)\left(s_2^6 - \phi_1 s_1^6\right) + \ldots$$

$$+ \left(s_p^4 - \phi_1 s_{p-1}^4 - \cdots - \phi_{p-1} s_1^4\right)\left(s_p^6 - \phi_1 s_{p-1}^6 - \cdots - \phi_{p-1} s_1^6\right)\},$$

by substituting $s_i = i\Delta$, proving the result. □

LEMMA 7. *(a) For k a non-negative integer,*

$$\frac{\Delta^k}{n} \sum_{i=1}^{n} i^k = \frac{s_n^k}{k+1} + \mathrm{O}\left(\Delta s_n^{k-1}\right).$$

(b)

$$\frac{a_{12}}{\left(1 - \sum_{j=1}^{p} \phi_j\right)} = \frac{\Delta^2}{n} \sum_{i=p+1}^{n} \left\{i^2 - \sum_{j=1}^{p} \phi_j (i-j)^2\right\} + \mathrm{O}\left(\Delta s_n\right). = \frac{s_n^2}{3}\left(1 - \sum_{j=1}^{p} \phi_j\right) + \mathrm{O}\left(\Delta s_n\right).$$

(c)

$$\frac{a_{13}}{\left(1 - \sum_{j=1}^{p} \phi_j\right)} = \frac{\Delta^2}{n} \sum_{i=p+1}^{n} \left\{i^4 - \sum_{j=1}^{p} \phi_j (i-j)^4\right\} + \mathrm{O}\left(\Delta s_n^3\right) = \frac{s_n^4}{5}\left(1 - \sum_{j=1}^{p} \phi_j\right) + \mathrm{O}\left(\Delta s_n^3\right).$$

(d)

$$a_{22} = \frac{\Delta^4}{n} \sum_{i=p+1}^{n} \left\{i^2 - \sum_{j=1}^{p} \phi_j (i-j)^2\right\}^2 + \mathrm{O}\left(\Delta s_n^3\right) = \frac{s_n^4}{5}\left(1 - \sum_{j=1}^{p} \phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^3\right).$$

(e)

$$\frac{a_{23}}{\left(1-\sum_{j=1}^{p}\phi_j\right)} = \frac{\Delta^2}{n}\sum_{i=p+1}^{n}\left\{i^2-\sum_{j=1}^{p}\phi_j(i-j)^2\right\}\left\{i^4-\sum_{j=1}^{p}\phi_j(i-j)^4\right\}+\mathrm{O}\left(\Delta s_n^5\right)$$

$$= \frac{s_n^6}{7}\left(1-\sum_{j=1}^{p}\phi_j\right)+\mathrm{O}\left(\Delta s_n^5\right).$$

(f)

$$a_{33} = \frac{\Delta^4}{n}\sum_{i=p+1}^{n}\left\{i^4-\sum_{j=1}^{p}\phi_j(i-j)^4\right\}^2+\mathrm{O}\left(\Delta s_n^7\right) = \frac{s_n^8}{9}\left(1-\sum_{j=1}^{p}\phi_j\right)^2+\mathrm{O}\left(\Delta s_n^7\right).$$

(g)

$$\frac{b_1}{\left(1-\sum_{j=1}^{p}\phi_j\right)} = \frac{\Delta^6}{n}\sum_{i=p+1}^{n}\left\{i^6-\sum_{j=1}^{p}\phi_j(i-j)^6\right\}+\mathrm{O}\left(\Delta s_n^5\right) = \frac{s_n^6}{7}\left(1-\sum_{j=1}^{p}\phi_j\right)+\mathrm{O}\left(\Delta s_n^5\right).$$

(h)

$$b_2 = \frac{\Delta^8}{n}\sum_{i=p+1}^{n}\left\{i^2-\sum_{j=1}^{p}\phi_j(i-j)^2\right\}\left\{i^6-\sum_{j=1}^{p}\phi_j(i-j)^6\right\}+\mathrm{O}\left(\Delta s_n^7\right)$$

$$= \frac{s_n^8}{9}\left(1-\sum_{j=1}^{p}\phi_j\right)^2+\mathrm{O}\left(\Delta s_n^7\right).$$

(i)

$$b_3 = \frac{\Delta^{10}}{n}\sum_{i=p+1}^{n}\left\{i^4-\sum_{j=1}^{p}\phi_j(i-j)^4\right\}\left\{i^6-\sum_{j=1}^{p}\phi_j(i-j)^6\right\}+\mathrm{O}\left(\Delta s_n^9\right)$$

$$= \frac{s_n^{10}}{11}\left(1-\sum_{j=1}^{p}\phi_j\right)^2+\mathrm{O}\left(\Delta s_n^9\right).$$

PROOF. In each summation on i, let k denote the highest power of i, and use the fact that for fixed $p \geq 0$, we have

$$\frac{\Delta^k}{n} \sum_{i=p+1}^{n} i^k = \frac{\Delta^k n^{k+1}}{n(k+1)} + \mathrm{O}\left(\Delta^k n^{k-1}\right) = \frac{s_n^k}{k+1} + \mathrm{O}\left(\Delta s_n^{k-1}\right).$$

□

LEMMA 8. *We can write*

$$\left(n^{-1}\mathbf{X}'\mathbf{\Gamma}^{-1}\mathbf{X}\right)^{-1} = \left\{ \frac{{s_n}^{-12}}{\frac{256}{496125}\left(1 - \sum_{j=1}^{p} \phi_j\right)^4} + \mathrm{O}\left(\Delta s_n^{-13}\right) \right\} \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix},$$

where

$$c_{11} = \frac{4s_n^{12}}{2205}\left(1 - \sum_{j=1}^{p} \phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^{11}\right)$$

$$c_{12} = c_{21} = -\frac{8s_n^{10}}{945}\left(1 - \sum_{j=1}^{p} \phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^{9}\right)$$

$$c_{13} = c_{31} = \frac{4s_n^{8}}{525}\left(1 - \sum_{j=1}^{p} \phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^{7}\right)$$

$$c_{22} = \frac{16s_n^{8}}{225}\left(1 - \sum_{j=1}^{p} \phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^{7}\right)$$

$$c_{23} = c_{32} = -\frac{8s_n^{6}}{105}\left(1 - \sum_{j=1}^{p} \phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^{5}\right)$$

$$c_{33} = \frac{4s_n^{4}}{45}\left(1 - \sum_{j=1}^{p} \phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^{3}\right).$$

PROOF. By Lemma 5 and Lemma 7, we have

$$n^{-1}\mathbf{X}'\mathbf{\Gamma}^{-1}\mathbf{X} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix},$$

where

$$a_{11} = \left(1 - \sum_{j=1}^{p} \phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^{-1}\right)$$

$$a_{12} = a_{21} = \frac{s_n^2}{3}\left(1 - \sum_{j=1}^{p} \phi_j\right)^2 + \mathrm{O}\left(\Delta s_n\right)$$

$$a_{13} = a_{31} = \frac{s_n^4}{5}\left(1 - \sum_{j=1}^{p} \phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^3\right)$$

$$a_{22} = \frac{s_n^4}{5}\left(1 - \sum_{j=1}^{p} \phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^3\right)$$

$$a_{23} = a_{32} = \frac{s_n^6}{7}\left(1 - \sum_{j=1}^{p} \phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^5\right)$$

$$a_{33} = \frac{s_n^8}{9}\left(1 - \sum_{j=1}^{p} \phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^7\right).$$

Taking the inverse of $n^{-1}\mathbf{X}'\mathbf{\Gamma}^{-1}\mathbf{X}$ yields

$$\left(n^{-1}\mathbf{X}'\mathbf{\Gamma}^{-1}\mathbf{X}\right)^{-1} = \frac{{s_n}^{-12}}{\frac{256}{496125}\left(1 - \sum_{j=1}^{p} \phi_j\right)^4 + \mathrm{O}\left(\Delta s_n^{-1}\right)} \frac{1}{\sigma^2} \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix},$$

where

$$c_{11} = \frac{4s_n^{12}}{2205}\left(1-\sum_{j=1}^{p}\phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^{11}\right)$$

$$c_{12} = c_{21} = -\frac{8s_n^{10}}{945}\left(1-\sum_{j=1}^{p}\phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^{9}\right)$$

$$c_{13} = c_{31} = \frac{4s_n^{8}}{525}\left(1-\sum_{j=1}^{p}\phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^{7}\right)$$

$$c_{22} = \frac{16s_n^{8}}{225}\left(1-\sum_{j=1}^{p}\phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^{7}\right)$$

$$c_{23} = c_{32} = -\frac{8s_n^{6}}{105}\left(1-\sum_{j=1}^{p}\phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^{5}\right)$$

$$c_{33} = \frac{4s_n^{4}}{45}\left(1-\sum_{j=1}^{p}\phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^{3}\right).$$

Assuming causality in the autoregressive process, $1-\sum_{j=1}^{p}\phi_j \neq 0$, we can write

$$\frac{{s_n}^{-12}}{\frac{256}{496125}\left(1-\sum_{j=1}^{p}\phi_j\right)^4 + \mathrm{O}\left(\Delta s_n^{-1}\right)} = \frac{{s_n}^{-12}}{\frac{256}{496125}\left(1-\sum_{j=1}^{p}\phi_j\right)^4} + \mathrm{O}\left(\Delta s_n^{-13}\right).$$

□

5.3.1. MSE OF $\widehat{\beta}_2$. The first term of the bias of $\widehat{\beta}_2$ is given by

$$\mathrm{E}\left(\widehat{\beta}_2 - \beta_2\right) = \frac{f^{(6)}(0)}{720}\mathbf{e}_2'\left(\mathbf{X}'\Gamma^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Gamma^{-1}\begin{bmatrix} s_1^6 \\ \vdots \\ s_n^6 \end{bmatrix}$$

$$+\frac{f^{(6)}(0)}{720}\mathbf{e}_2'\left(\mathbf{X}'\Gamma^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Gamma^{-1}\begin{bmatrix}\mathrm{O}\left(s_1^8\right)\\ \vdots\\ \mathrm{O}\left(s_n^8\right)\end{bmatrix}$$

$$=\frac{f^{(6)}(0)}{720}\mathbf{e}_2'\left(\mathbf{X}'\Gamma^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Gamma^{-1}\begin{bmatrix}s_1^6\\ \vdots\\ s_n^6\end{bmatrix}$$

$$+\frac{f^{(6)}(0)}{720}\mathbf{e}_2'\left(\mathbf{X}'\Gamma^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Gamma^{-1}\begin{bmatrix}\mathrm{O}\left(s_n^8\right)\\ \vdots\\ \mathrm{O}\left(s_n^8\right)\end{bmatrix}.$$

From Lemma 8, we write

$$\mathrm{E}\left(\widehat{\beta}_2-\beta_2\right)=\frac{f^{(6)}(0)}{720}\mathbf{e}_2'\left\{\frac{{s_n}^{-12}}{\frac{256}{496125}\left(1-\sum_{j=1}^p\phi_j\right)^4}+\mathrm{O}\left(\Delta s_n^{-13}\right)\right\}\begin{bmatrix}c_{11}&c_{12}&c_{13}\\ c_{21}&c_{22}&c_{23}\\ c_{31}&c_{32}&c_{33}\end{bmatrix}$$

$$\times\begin{bmatrix}\frac{s_n^6}{7}\left(1-\sum_{j=1}^p\phi_j\right)^2+\mathrm{O}\left(\Delta s_n^5\right)\\ \frac{s_n^8}{9}\left(1-\sum_{j=1}^p\phi_j\right)^2+\mathrm{O}\left(\Delta s_n^7\right)\\ \frac{s_n^{10}}{11}\left(1-\sum_{j=1}^p\phi_j\right)^2+\mathrm{O}\left(\Delta s_n^9\right)\end{bmatrix}+\mathrm{O}\left(s_n^6\right),$$

where

$$c_{11} = \frac{4s_n^{12}}{2205}\left(1 - \sum_{j=1}^{p}\phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^{11}\right)$$

$$c_{12} = c_{21} = -\frac{8s_n^{10}}{945}\left(1 - \sum_{j=1}^{p}\phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^{9}\right)$$

$$c_{13} = c_{31} = \frac{4s_n^{8}}{525}\left(1 - \sum_{j=1}^{p}\phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^{7}\right)$$

$$c_{22} = \frac{16s_n^{8}}{225}\left(1 - \sum_{j=1}^{p}\phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^{7}\right)$$

$$c_{23} = c_{32} = -\frac{8s_n^{6}}{105}\left(1 - \sum_{j=1}^{p}\phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^{5}\right)$$

$$c_{33} = \frac{4s_n^{4}}{45}\left(1 - \sum_{j=1}^{p}\phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^{3}\right).$$

Thus, we have the final expression for the bias:

$$\begin{aligned}\mathrm{E}\left(\widehat{\beta}_2 - \beta_2\right) &= \frac{f^{(6)}(0)}{720}\left\{\frac{{s_n}^{-12}}{\frac{256}{496125}\left(1 - \sum_{j=1}^{p}\phi_j\right)^4} + \mathrm{O}\left(\Delta s_n^{-13}\right)\right\}\left\{-\frac{256}{1091475}s_n^{16} + \mathrm{O}\left(\Delta s_n^{15}\right)\right\} \\ &\quad \times \left(1 - \sum_{j=1}^{p}\phi_j\right)^4 + \mathrm{O}\left(s_n^6\right) \\ &= -\frac{1}{1584}f^{(6)}(0)s_n^4 + \mathrm{O}\left(s_n^6\right). \qquad (47)\end{aligned}$$

Next, we calculate the variance of $\widehat{\beta}_2$. Therefore, consider

$$\mathrm{Var}\left(\widehat{\beta}_2\right) = \mathrm{Var}\left\{\mathbf{e}_2'\left(\mathbf{X}'\Gamma^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Gamma^{-1}\mathbf{Y}\right\}$$

$$= \mathbf{e}_2' \left(\mathbf{X}'\Gamma^{-1}\mathbf{X}\right)^{-1} \mathbf{X}'\Gamma^{-1}\mathrm{Var}(\mathbf{Y})\Gamma^{-1}\mathbf{X} \left(\mathbf{X}'\Gamma^{-1}\mathbf{X}\right)^{-1} \mathbf{e}_2$$

$$= \mathbf{e}_2' \left(\mathbf{X}'\Gamma^{-1}\mathbf{X}\right)^{-1} \mathbf{e}_2$$

$$= \frac{\sigma^2}{n}\mathbf{e}_2' \left\{ \frac{{s_n}^{-12}}{\frac{256}{496125}\left(1-\sum_{j=1}^{p}\phi_j\right)^4} + \mathrm{O}\left(\Delta s_n^{-13}\right) \right\} \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix} \mathbf{e}_2,$$

where

$$c_{11} = \frac{4s_n^{12}}{2205}\left(1-\sum_{j=1}^{p}\phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^{11}\right)$$

$$c_{12} = c_{21} = -\frac{8s_n^{10}}{945}\left(1-\sum_{j=1}^{p}\phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^{9}\right)$$

$$c_{13} = c_{31} = \frac{4s_n^{8}}{525}\left(1-\sum_{j=1}^{p}\phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^{7}\right)$$

$$c_{22} = \frac{16s_n^{8}}{225}\left(1-\sum_{j=1}^{p}\phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^{7}\right)$$

$$c_{23} = c_{32} = -\frac{8s_n^{6}}{105}\left(1-\sum_{j=1}^{p}\phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^{5}\right)$$

$$c_{33} = \frac{4s_n^{4}}{45}\left(1-\sum_{j=1}^{p}\phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^{3}\right),$$

Finally, we have

$$\mathrm{Var}\left(\widehat{\beta}_2\right) = \frac{2205\sigma^2}{16ns_n^4\left(1-\sum_{j=1}^{p}\phi_j\right)^2} + \mathrm{O}\left(\Delta^2 s_n^{-6}\right).$$

Then since $n^{-1} = \Delta s_n^{-1}$, the variance of $\widehat{\beta}_2$ can be written as

$$\text{Var}\left(\widehat{\beta}_2\right) = \frac{2205\sigma^2\Delta}{16s_n^5\left(1-\sum_{j=1}^{p}\phi_j\right)^2} + \text{O}\left(\Delta^2 s_n^{-6}\right). \tag{48}$$

Combining (47) and (48), the mean squared error of $\widehat{\beta}_2$ is given by

$$\begin{aligned}
\text{MSE}\left(\widehat{\beta}_2\right) &= \left\{\text{E}\left(\widehat{\beta}_2 - \beta_2\right)\right\}^2 + \text{Var}\left(\widehat{\beta}_2\right) \\
&= \left\{-\frac{1}{1584}f^{(6)}(0)s_n^4 + \text{O}\left(s_n^6\right)\right\}^2 + \frac{2205\sigma^2\Delta}{16s_n^5\left(1-\sum_{j=1}^{p}\phi_j\right)^2} + \text{O}\left(\Delta^2 s_n^{-6}\right) \\
&= \frac{1}{2509056}\left\{f^{(6)}(0)\right\}^2 s_n^8 + \frac{2205\sigma^2\Delta}{16s_n^5\left(1-\sum_{j=1}^{p}\phi_j\right)^2} + \text{O}\left(s_n^{10} + \Delta^2 s_n^{-6}\right). \qquad (49)
\end{aligned}$$

Now, we minimize MSE $\left(\widehat{\beta}_2\right)$ with respect to s_n. Therefore, differentiating (49) with respect to s_n yields

$$\begin{aligned}
\frac{\partial}{\partial s_n}\left\{\text{MSE}\left(\widehat{\beta}_2\right)\right\} &= \frac{\partial}{\partial s_n}\left[\frac{1}{2509056}\left\{f^{(6)}(0)\right\}^2 s_n^8 + \frac{2205\sigma^2\Delta}{16s_n^5\left(1-\sum_{j=1}^{p}\phi_j\right)^2}\right] \\
&= \frac{1}{313632}\left\{f^{(6)}(0)\right\}^2 s_n^7 - \frac{11025\sigma^2\Delta}{16s_n^6\left(1-\sum_{j=1}^{p}\phi_j\right)^2}. \qquad (50)
\end{aligned}$$

Setting (50) equal to zero and solving for s_n yields the minimum of MSE $\left(\widehat{\beta}_2\right)$ and is given by

$$s_n = \left[\frac{216112050\sigma^2\Delta}{\left\{f^{(6)}(0)\right\}^2\left(1-\sum_{j=1}^{p}\phi_j\right)^2}\right]^{1/13}. \tag{51}$$

5.3.2. MSE OF $\widehat{\beta}_4$. The first term of the bias of $\widehat{\beta}_4$ is given by

$$\begin{aligned}
\mathrm{E}\left(\widehat{\beta}_4 - \beta_4\right) &= \frac{f^{(6)}(0)}{720}\mathbf{e}_4\left(\mathbf{X}'\Gamma^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Gamma^{-1}\begin{bmatrix} s_1^6 \\ \vdots \\ s_n^6 \end{bmatrix} \\
&\quad + \frac{f^{(6)}(0)}{720}\mathbf{e}_4\left(\mathbf{X}'\Gamma^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Gamma^{-1}\begin{bmatrix} \mathrm{O}\left(s_1^8\right) \\ \vdots \\ \mathrm{O}\left(s_n^8\right) \end{bmatrix} \\
&= \frac{f^{(6)}(0)}{720}\mathbf{e}_4'\left(\mathbf{X}'\Gamma^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Gamma^{-1}\begin{bmatrix} s_1^6 \\ \vdots \\ s_n^6 \end{bmatrix} \\
&\quad + \frac{f^{(6)}(0)}{720}\mathbf{e}_4'\left(\mathbf{X}'\Gamma^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Gamma^{-1}\begin{bmatrix} \mathrm{O}\left(s_n^8\right) \\ \vdots \\ \mathrm{O}\left(s_n^8\right) \end{bmatrix}.
\end{aligned}$$

Using Lemma 8, we have

$$\mathrm{E}\left(\widehat{\beta}_4 - \beta_4\right) = \frac{f^{(6)}(0)}{720}\mathbf{e}_4'\left\{\frac{{s_n}^{-12}}{\frac{256}{496125}\left(1-\sum_{j=1}^{p}\phi_j\right)^4} + \mathrm{O}\left(\Delta s_n^{-13}\right)\right\}\begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix}$$

$$\times \begin{bmatrix} \frac{s_n^6}{7}\left(1-\sum_{j=1}^{p}\phi_j\right)^2+\mathrm{O}\left(\Delta s_n^5\right) \\ \frac{s_n^8}{9}\left(1-\sum_{j=1}^{p}\phi_j\right)^2+\mathrm{O}\left(\Delta s_n^7\right) \\ \frac{s_n^{10}}{11}\left(1-\sum_{j=1}^{p}\phi_j\right)^2+\mathrm{O}\left(\Delta s_n^9\right) \end{bmatrix} + \mathrm{O}\left(s_n^4\right),$$

where

$$c_{11} = \frac{4s_n^{12}}{2205}\left(1-\sum_{j=1}^{p}\phi_j\right)^2+\mathrm{O}\left(\Delta s_n^{11}\right)$$

$$c_{12} = c_{21} = -\frac{8s_n^{10}}{945}\left(1-\sum_{j=1}^{p}\phi_j\right)^2+\mathrm{O}\left(\Delta s_n^{9}\right)$$

$$c_{13} = c_{31} = \frac{4s_n^{8}}{525}\left(1-\sum_{j=1}^{p}\phi_j\right)^2+\mathrm{O}\left(\Delta s_n^{7}\right)$$

$$c_{22} = \frac{16s_n^{8}}{225}\left(1-\sum_{j=1}^{p}\phi_j\right)^2+\mathrm{O}\left(\Delta s_n^{7}\right)$$

$$c_{23} = c_{32} = -\frac{8s_n^{6}}{105}\left(1-\sum_{j=1}^{p}\phi_j\right)^2+\mathrm{O}\left(\Delta s_n^{5}\right)$$

$$c_{33} = \frac{4s_n^{4}}{45}\left(1-\sum_{j=1}^{p}\phi_j\right)^2+\mathrm{O}\left(\Delta s_n^{3}\right).$$

Thus, the final bias expression of $\widehat{\beta}_4$ is given by

$$\mathrm{E}\left(\widehat{\beta}_4-\beta_4\right) = \frac{f^{(6)}(0)}{720}\left\{\frac{{s_n}^{-12}}{\frac{256}{496125}\left(1-\sum_{j=1}^{p}\phi_j\right)^4}+\mathrm{O}\left(\Delta s_n^{-13}\right)\right\}\left\{\frac{256}{363825}s_n^{14}+\mathrm{O}\left(\Delta s_n^{13}\right)\right\}$$
$$\times\left(1-\sum_{j=1}^{p}\phi_j\right)^4+\mathrm{O}\left(s_n^4\right)$$

$$= \frac{1}{528} f^{(6)}(0) s_n^2 + \mathrm{O}\left(s_n^4\right). \tag{52}$$

Now, we calculate the variance of $\widehat{\beta}_4$ given by

$$\begin{aligned}
\mathrm{Var}\left(\widehat{\beta}_4\right) &= \mathrm{Var}\left\{\mathbf{e}_4' \left(\mathbf{X}'\Gamma^{-1}\mathbf{X}\right)^{-1} \mathbf{X}'\Gamma^{-1}\mathbf{Y}\right\} \\
&= \mathbf{e}_4' \left(\mathbf{X}'\Gamma^{-1}\mathbf{X}\right)^{-1} \mathbf{X}'\Gamma^{-1}\mathrm{Var}(\mathbf{Y})\Gamma^{-1}\mathbf{X} \left(\mathbf{X}'\Gamma^{-1}\mathbf{X}\right)^{-1} \mathbf{e}_4 \\
&= \mathbf{e}_4' \left(\mathbf{X}'\Gamma^{-1}\mathbf{X}\right)^{-1} \mathbf{e}_4 \\
&= \frac{\sigma^2}{n} \mathbf{e}_4' \left\{ \frac{{s_n}^{-12}}{\frac{256}{496125}\left(1 - \sum_{j=1}^{p} \phi_j\right)^4} + \mathrm{O}\left(\Delta s_n^{-13}\right) \right\} \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix} \mathbf{e}_4,
\end{aligned}$$

where

$$c_{11} = \frac{4 s_n^{12}}{2205} \left(1 - \sum_{j=1}^{p} \phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^{11}\right)$$

$$c_{12} = c_{21} = -\frac{8 s_n^{10}}{945} \left(1 - \sum_{j=1}^{p} \phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^{9}\right)$$

$$c_{13} = c_{31} = \frac{4 s_n^{8}}{525} \left(1 - \sum_{j=1}^{p} \phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^{7}\right)$$

$$c_{22} = \frac{16 s_n^{8}}{225} \left(1 - \sum_{j=1}^{p} \phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^{7}\right)$$

$$c_{23} = c_{32} = -\frac{8 s_n^{6}}{105} \left(1 - \sum_{j=1}^{p} \phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^{5}\right)$$

$$c_{33} = \frac{4s_n^4}{45}\left(1 - \sum_{j=1}^{p}\phi_j\right)^2 + \mathrm{O}\left(\Delta s_n^3\right).$$

Finally, we have

$$\mathrm{Var}\left(\widehat{\beta}_4\right) = \frac{11025\sigma^2}{64ns_n^8\left(1 - \sum_{j=1}^{p}\phi_j\right)^2} + \mathrm{O}\left(\Delta^2 s_n^{-10}\right).$$

Then since $n^{-1} = \Delta s_n^{-1}$, the variance of $\widehat{\beta}_4$ can be written as

$$\mathrm{Var}\left(\widehat{\beta}_4\right) = \frac{11025\sigma^2\Delta}{64s_n^9\left(1 - \sum_{j=1}^{p}\phi_j\right)^2} + \mathrm{O}\left(\Delta^2 s_n^{-10}\right). \tag{53}$$

Combining (52) and (53), the mean squared error of $\widehat{\beta}_4$ is given by

$$\begin{aligned}
\mathrm{MSE}\left(\widehat{\beta}_4\right) &= \left\{\mathrm{E}\left(\widehat{\beta}_4 - \beta_4\right)\right\}^2 + \mathrm{Var}\left(\widehat{\beta}_4\right) \\
&= \left\{\frac{1}{528}f^{(6)}(0)s_n^2 + \mathrm{O}\left(s_n^4\right)\right\}^2 + \frac{11025\sigma^2\Delta}{64s_n^9\left(1 - \sum_{j=1}^{p}\phi_j\right)^2} + \mathrm{O}\left(\Delta^2 s_n^{-10}\right) \\
&= \frac{1}{278784}\left\{f^{(6)}(0)\right\}^2 s_n^4 + \frac{11025\sigma^2\Delta}{64s_n^9\left(1 - \sum_{j=1}^{p}\phi_j\right)^2} + \mathrm{O}\left(s_n^6 + \Delta^2 s_n^{-10}\right). \qquad (54)
\end{aligned}$$

Next, we minimize $\mathrm{MSE}\left(\widehat{\beta}_4\right)$ with respect to s_n. Therefore, differentiating (54) with respect to s_n yields

$$\begin{aligned}
\frac{\partial}{\partial s_n}\left\{\mathrm{MSE}\left(\widehat{\beta}_4\right)\right\} &= \frac{\partial}{\partial s_n}\left[\frac{1}{278784}\left\{f^{(6)}(0)\right\}^2 s_n^4 + \frac{11025\sigma^2\Delta}{64s_n^9\left(1 - \sum_{j=1}^{p}\phi_j\right)^2}\right] \\
&= \frac{1}{69696}\left\{f^{(6)}(0)\right\}^2 s_n^3 - \frac{99225\sigma^2\Delta}{64s_n^{10}\left(1 - \sum_{j=1}^{p}\phi_j\right)^2}. \qquad (55)
\end{aligned}$$

Setting (55) equal to zero and solving for s_n yields the minimum of MSE $\left(\widehat{\beta}_4\right)$ and is given by

$$s_n = \left[\frac{108056025\sigma^2\Delta}{\left\{f^{(6)}(0)\right\}^2\left(1-\sum_{j=1}^{p}\phi_j\right)^2}\right]^{1/13}. \tag{56}$$

5.3.3. Minimizing the MSE of $\widehat{\psi}$. We minimize the mean squared error of $\widehat{\psi}$. Recall

$$\begin{aligned}\psi &= \frac{M^4}{R_g^4}\\ &= \frac{60\beta_4+30\beta_2^2}{(-3\beta_2)^2}\\ &= \frac{20}{3}\frac{\beta_4}{\beta_2^2}+\frac{10}{3}.\end{aligned}$$

Hence, minimizing the MSE$\left(\widehat{\psi}\right)$ is equivalent to minimizing the MSE$\left(\widehat{\beta}_4/\widehat{\beta}_2^2\right)$. Therefore, consider the approximation

$$\frac{\widehat{\beta}_4}{\widehat{\beta}_2^2} \approx \frac{\beta_4}{\beta_2^2}+\frac{1}{\beta_2^2}\left(\widehat{\beta}_4-\beta_4\right)-\frac{2\beta_4}{\beta_2^3}\left(\widehat{\beta}_2-\beta_2\right).$$

Thus, we can write

$$\begin{aligned}\mathrm{Var}\left(\frac{\widehat{\beta}_4}{\widehat{\beta}_2^2}\right) &\approx \frac{1}{\beta_2^4}\mathrm{Var}\left(\widehat{\beta}_4\right)+\frac{4\beta_4^2}{\beta_2^6}\mathrm{Var}\left(\widehat{\beta}_2\right)-\frac{4\beta_4}{\beta_2^5}\mathrm{Cov}\left(\widehat{\beta}_4,\widehat{\beta}_2\right)\\ &= \frac{1}{\beta_2^4}\left(\frac{11025\sigma^2\Delta}{64s_n^9\left(1-\sum_{j=1}^{p}\phi_j\right)^2}\right)+\frac{4\beta_4^2}{\beta_2^6}\left(\frac{2205\sigma^2\Delta}{16s_n^5\left(1-\sum_{j=1}^{p}\phi_j\right)^2}\right)\\ &\quad -\frac{4\beta_4}{\beta_2^5}\left(\frac{-4725\sigma^2\Delta}{32s_n^7\left(1-\sum_{j=1}^{p}\phi_j\right)^2}\right)+\mathrm{O}\left(\Delta^2 s_n^{-10}\right).\end{aligned}$$

Furthermore, the bias of $\widehat{\beta}_4/\widehat{\beta}_2^2$ becomes

$$\begin{aligned}
\mathrm{E}\left(\frac{\widehat{\beta}_4}{\widehat{\beta}_2^2}-\frac{\beta_4}{\beta_2^2}\right) &= \frac{1}{\beta_2^2}\left(\mathrm{E}\widehat{\beta}_4-\beta_4\right)-\frac{2\beta_4}{\beta_2^3}\left(\mathrm{E}\widehat{\beta}_2-\beta_2\right) \\
&= \frac{1}{\beta_2^2}\left\{\frac{1}{528}f^{(6)}(0)s_n^2\right\}-\frac{2\beta_4}{\beta_2^3}\left\{-\frac{1}{1584}f^{(6)}(0)s_n^4\right\}+\mathrm{O}\left(s_n^6\right).
\end{aligned}$$

Therefore, the $\mathrm{MSE}\left(\widehat{\beta}_4/\widehat{\beta}_2^2\right)$ is given by

$$\begin{aligned}
\mathrm{MSE}\left(\frac{\widehat{\beta}_4}{\widehat{\beta}_2^2}\right) =& \mathrm{Var}\left(\frac{\widehat{\beta}_4}{\widehat{\beta}_2^2}\right)+\left\{\mathrm{Bias}\left(\frac{\widehat{\beta}_4}{\widehat{\beta}_2^2}\right)\right\}^2 \\
\approx& \frac{1}{\beta_2^4}\left(\frac{11025\sigma^2\Delta}{64s_n^9\left(1-\sum_{j=1}^p\phi_j\right)^2}\right)+\frac{4\beta_4^2}{\beta_2^6}\left(\frac{2205\sigma^2\Delta}{16s_n^5\left(1-\sum_{j=1}^p\phi_j\right)^2}\right) \\
&+\frac{4\beta_4}{\beta_2^5}\left(\frac{4725\sigma^2\Delta}{32s_n^7\left(1-\sum_{j=1}^p\phi_j\right)^2}\right)+\mathrm{O}\left(\Delta^2 s_n^{-10}\right) \\
&+\left\{\frac{1}{\beta_2^2}\left(\frac{1}{528}f^{(6)}(0)s_n^2\right)+\frac{2\beta_4}{\beta_2^3}\left(\frac{1}{1584}f^{(6)}(0)s_n^4\right)+\mathrm{O}\left(s_n^6\right)\right\}^2 \\
=& \frac{1}{\beta_2^4}\left(\frac{11025\sigma^2\Delta}{64s_n^9\left(1-\sum_{j=1}^p\phi_j\right)^2}\right)+\frac{4\beta_4^2}{\beta_2^6}\left(\frac{2205\sigma^2\Delta}{16s_n^5\left(1-\sum_{j=1}^p\phi_j\right)^2}\right) \\
&+\frac{4\beta_4}{\beta_2^5}\left(\frac{4725\sigma^2\Delta}{32s_n^7\left(1-\sum_{j=1}^p\phi_j\right)^2}\right)+\frac{1}{278784\beta_2^4}\left\{f^{(6)}(0)\right\}^2 s_n^4 \\
&+\frac{\beta_4^2}{209748\beta_2^5}\left\{f^{(6)}(0)\right\}^2 s_n^6+\frac{\beta_4^2}{627264\beta_2^6}\left\{f^{(6)}(0)\right\}^2 s_n^8+\mathrm{O}\left(\Delta^2 s_n^{-10}+s_n^{10}\right).
\end{aligned}$$

We minimize MSE $\left(\widehat{\beta}_4/\widehat{\beta}_2^2\right)$ with respect to s_n. Therefore, differentiating with respect to s_n yields

$$\begin{aligned}\frac{\partial}{\partial s_n}\left\{\text{MSE}\left(\frac{\widehat{\beta}_4}{\widehat{\beta}_2^2}\right)\right\} = &-\frac{99225\sigma^2\Delta}{64s_n^{10}\left(1-\sum_{j=1}^p\phi_j\right)^2\beta_2^4} - \frac{11025\sigma^2\Delta\beta_4^2}{4s_n^6\left(1-\sum_{j=1}^p\phi_j\right)^2\beta_2^6} \\ &- \frac{33075\sigma^2\Delta\beta_4}{8s_n^8\left(1-\sum_{j=1}^p\phi_j\right)^2\beta_2^5} + \frac{1}{69696\beta_2^4}\left\{f^{(6)}(0)\right\}^2 s_n^3 \\ &+ \frac{\beta_4^2}{34958\beta_2^5}\left\{f^{(6)}(0)\right\}^2 s_n^5 + \frac{\beta_4^2}{78408\beta_2^6}\left\{f^{(6)}(0)\right\}^2 s_n^7. \qquad (57)\end{aligned}$$

Setting (57) equal to zero yields

$$\begin{aligned}&-\frac{99225\sigma^2\Delta}{54\left(1-\sum_{j=1}^p\phi_j\right)^2} - \frac{33075\sigma^2\Delta\beta_4}{8\beta_2\left(1-\sum_{j=1}^p\phi_j\right)^2}s_n^2 - \frac{11025\sigma^2\Delta\beta_4^2}{4\beta_2^2\left(1-\sum_{j=1}^p\phi_j\right)^2}s_n^4 + \\ &\frac{1}{69696}\left\{f^{(6)}(0)\right\}^2 s_n^{13} + \frac{\beta_4^2}{34958\beta_2}\left\{f^{(6)}(0)\right\}^2 s_n^{15} + \frac{\beta_4^2}{78408\beta_2^2}\left\{f^{(6)}(0)\right\}^2 s_n^{17} = 0. \qquad (58)\end{aligned}$$

Unfortunately, a closed-form solution in terms of s_n does not exist for the minimum MSE$(\widehat{\psi})$, but (58) can be solved for s_n via a numerical procedure once estimates of β_2, β_4, and $f^{(6)}(0)$ are determined.

5.4. ψ Estimation for Nucleosome Core Particle

We estimate ψ using an experimental log intensity curve for the molecule nucleosome core particle. Then given $\widehat{\psi}$, we determine from the curve one (or two)

$$\frac{\text{Height}}{\text{Radius}}.$$

In addition, we can estimate R_g from its experimental log intensity curve and using the equation

$$R_g^2 = \frac{\text{Radius}^2}{2} + \frac{\text{Height}^2}{12}. \tag{59}$$

we can determine the height and radius of a cylinder with the same shape of the molecule. Experimental SAXS data for the molecule nucleosome core particle (NCP) is given in Figure 5.1. For this molecule, we determine $\widehat{\psi} = 3.0$ which from Figure 4.2 corresponds to a height/radius of 1.1 or 2.4. Furthermore, from the log intensity curve we can calculate $\widehat{R}_g = 41.4$Å. Then, we use (59) to determine two possible cylinders for this molecule. The first cylinder has dimensions height = 57.1Å and radius = 53.8Å, and the second cylinder has dimensions height = 101.2Å and radius = 41.5Å. Figure 5.2 depicts a digitally created image of the molecule NCP suspended within each of the first cylinder. We see that the first cylinder provides an excellent fit for the molecule. This example illustrates how we can determine a rough estimate for the size and shape of a molecule using only experimental SAXS data.

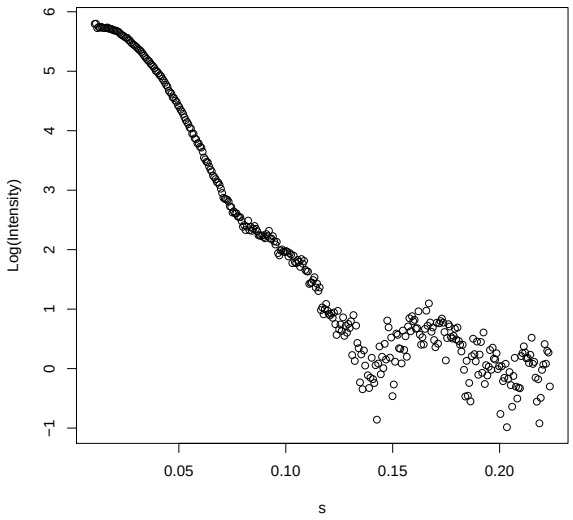


FIGURE 5.1. Plot of experimental SAXS data consisting of log intensity versus scattering angle s for the molecule nucleosome core particle (NCP).

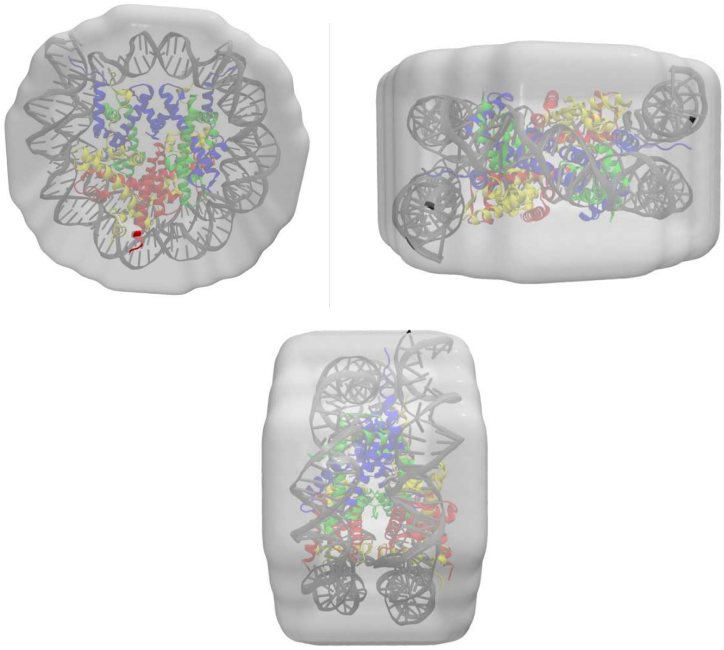


FIGURE 5.2. Digitally created images of the molecule NCP suspended within each of the good-fitting cylinders with height = 101.2 Å, radius = 41.5 Å. (a) Front view of the cylinder. (b) Side view of the cylinder. (c) Top view of the cylinder

5.5. Calculating D_{max} for a Molecule

From the estimated cylinder for a molecule, we can approximate D_{max} for the molecule. Given the height and radius of the cylinder, D_{max} is given by

$$D_{max} = \sqrt{\text{Height}^2 + 4 \times \text{Radius}^2}.$$

We calculate D_{max} for replicates of the molecule nucleosome core particle. Figure 5.3 contains a confidence interval for D_{max} from estimating the cylinder for each log intensity curve. The variance of D_{max} is computed by using a bootstrapping approach on the residuals of the fit of the log intensity curve. Figure 5.3 has results for two wild type and four mutants of the molecule nucleosome core particle. For each molecule, there is a separate 95% confidence interval for the smaller and larger cylinder. All of these confidence intervals were calculated using three replicate intensity curves. The first 12 confidence intervals are for the unsalted samples and the second 12 confidence intervals are the same samples with 50mM of added salt. These D_{max} values can be used as an initial estimate for further analysis.

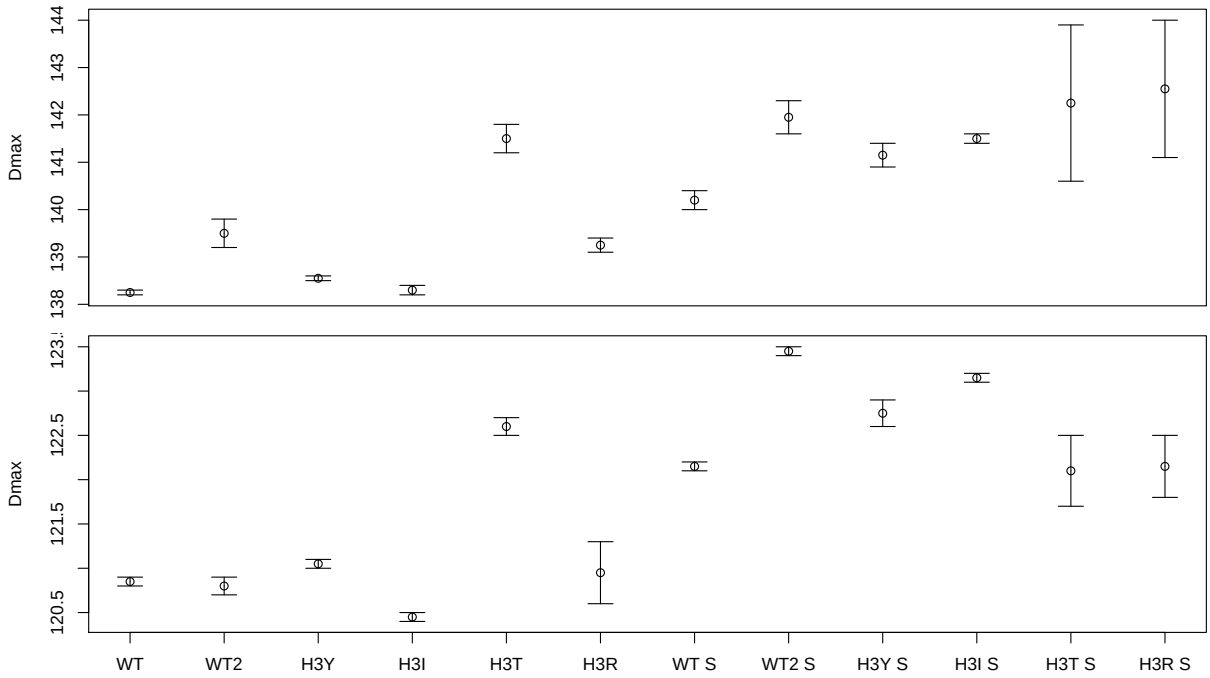


FIGURE 5.3. Results calculating D_{max} for two wild type and four mutants of the molecule nucleosome core particle. For each molecule, there is a separate 95% confidence interval for the smaller and larger cylinder. The first 12 confidence intervals are for the unsalted samples and the second 12 confidence intervals are the same samples with 50mM of added salt.

5.6. Experimental Data Example

Next, we consider an example involving the molecules aldolase and tyrosinase. Figure 5.4 contains an image of both molecules. For each molecule, we have ten replicate intensity curves and we estimate both R_g and ψ using the new procedure. Table 5.1 contains the results of estimating these values and their standard deviation. The estimated R_g value for both molecule is similar and cannot be differentiated given the size of their standard deviations. However, the $\widehat{\psi}$ value for the two molecules is significantly different. Using the new parameter ψ, we are able to distinguish between these two molecules when R_g alone was insufficient.

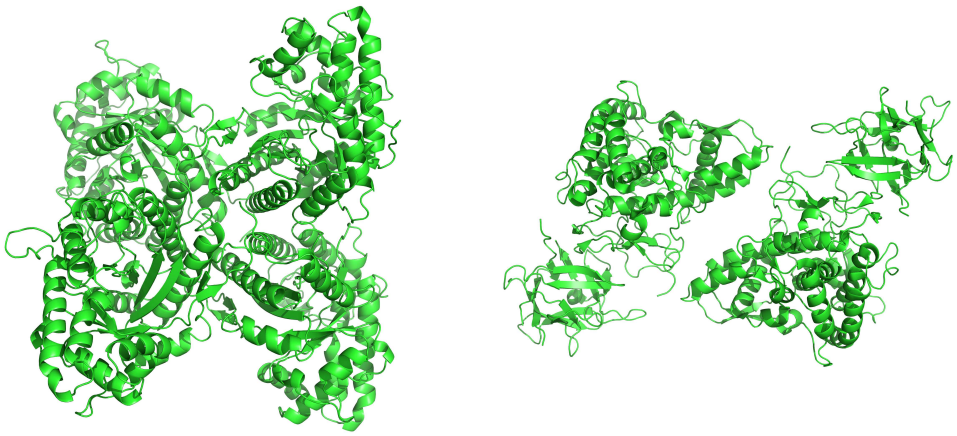


FIGURE 5.4. Digitally created images of the two different molecules. (a) Aldolase (b) Tyrosinase

TABLE 5.1. Results for estimating R_g and ψ using the new procedure for the molecules aldolase and tyrosinase. For each molecule, $\widehat{R}_g$ and its standard deviation are given for both methods.

Molecule	$\widehat{R}_g$	SD($\widehat{R}_g$)	$\widehat{\psi}$	SD($\widehat{\psi}$)
Aldolase	40.1	0.3	4.08	0.02
Tyrosinase	39.9	0.1	3.64	0.01

5.7. LIMITATIONS

We develop a semi-automatic procedure for estimating ψ and R_g from the log intensity curve for a molecule. With these two values, we estimate a good-fitting cylinder for the molecule in solution. The initial points in the estimation procedure can be removed using the modified DFBETAS procedure. However, this procedure is not ready to fully replace good judgment when determining the window of data points to fit to the log intensity curve. Furthermore, for $\widehat{\psi}$ values in the range 2.86 to 3.33 two cylinders are possible in the estimation.

5.8. Appendix

5.8.1. Extended Guinier Analysis Derivation. Guinier analysis involves estimating a molecule's radius of gyration from its experimental SAXS intensity curve. The equation relating R_g and the intensity curve is given by

$$\ln \mathcal{I}(s) = \ln \mathcal{I}(0) - \frac{1}{3} R_g^2 s^2 + \mathrm{O}\left(s^4\right). \tag{60}$$

Using analysis similar to Guinier's, we derive an equation relating R_g and M^4 to the intensity curve for SAXS data. Given the $p(r)$ function for a molecule, its corresponding intensity curve is be defined by

$$\mathcal{I}(s) = 4\pi \int_0^{D_{max}} p(r) \frac{\sin(sr)}{sr}\, dr. \tag{61}$$

Then using Taylor series expansion for $\sin(sr)/sr$ we can write

$$\begin{aligned}
\mathcal{I}(s) &= 4\pi \int_0^{D_{max}} p(r) \left\{ 1 - \frac{(sr)^2}{3!} + \frac{(sr)^4}{5!} - ... \right\} dr \\
&= 4\pi \int_0^{D_{max}} p(r)\, dr - 4\pi \frac{s^2}{3!} \int_0^{D_{max}} r^2 p(r)\, dr + 4\pi \frac{s^4}{5!} \int_0^{D_{max}} r^4 p(r)\, dr + \mathrm{O}\left(s^6\right) \\
&= \mathcal{I}(0) - \frac{1}{3} \mathcal{I}(0) R_g^2 s^2 + \frac{1}{60} \mathcal{I}(0) M^4 s^4 + \mathrm{O}\left(s^6\right) \\
&= \mathcal{I}(0) \left\{ 1 - \frac{1}{3} R_g^2 s^2 + \frac{1}{60} M^4 s^4 + \mathrm{O}\left(s^6\right) \right\}.
\end{aligned}$$

Therefore, for small values of s, we obtain the approximation

$$\mathcal{I}(s)/\mathcal{I}(0) \approx 1 - \frac{1}{3} R_g^2 s^2 + \frac{1}{60} M^4 s^4.$$

Taking the natural log of both sides gives

$$\ln\{\mathcal{I}(s)/\mathcal{I}(0)\} \approx \ln\left(1 - \frac{1}{3}R_g^2 s^2 + \frac{1}{60}M^4 s^4\right). \tag{62}$$

Next, consider the Taylor series expansion of $\ln(1-x)$ centered around $x = 0$:

$$\begin{aligned}\ln(1-x) &= \ln(1) - x - \frac{1}{2}x^2 - ... \\ &\approx -x - \frac{x^2}{2}.\end{aligned}$$

Substituting $x = \frac{1}{3}R_g^2 s^2 - \frac{1}{60}M^4 s^4$ yields

$$\begin{aligned}\ln\left(1 - \frac{1}{3}R_g^2 s^2 + \frac{1}{60}M^4 s^4\right) &\approx -\left(\frac{1}{3}R_g^2 s^2 - \frac{1}{60}M^4 s^4\right) - \frac{1}{2}\left(\frac{1}{3}R_g^2 s^2 - \frac{1}{60}M^4 s^4\right)^2 \\ &= -\frac{1}{3}R_g^2 s^2 + \left(\frac{1}{60}M^4 - \frac{1}{18}R_g^4\right)s^4.\end{aligned}$$

Therefore, (62) can be rewritten as

$$\begin{aligned}\ln\{\mathcal{I}(s)/\mathcal{I}(0)\} &\approx \ln\left(1 - \frac{1}{3}R_g^2 s^2 + \frac{1}{60}M^4 s^4\right) \\ &\approx -\frac{1}{3}R_g^2 s^2 + \left(\frac{1}{60}M^4 - \frac{1}{18}R_g^4\right)s^4.\end{aligned}$$

Thus, the final approximation relating $\ln \mathcal{I}(s)$ to R_g and M^4 is given by

$$\ln \mathcal{I}(s) \approx \ln \mathcal{I}(0) - \frac{1}{3}R_g^2 s^2 + \left(\frac{1}{60}M^4 - \frac{1}{18}R_g^4\right)s^4. \tag{63}$$

This approximation can estimate R_g and M^4 from experimental SAXS data; therefore, it can estimate ψ as well.

5.9. Estimating the Variance of $\widehat{R}_g^2$ and $\widehat{M}^4$

We have an approximation for $\widehat{R}_g^2$ and $\widehat{M}^4$ and we need to determine an estimate for their variance. First, to estimate Var $\left(\widehat{R}_g^2\right)$ using the approximation given in (42), we write

$$\begin{aligned}\text{Var}\left(\widehat{R}_g^2\right) &\approx \text{Var}\left(-3\widehat{\beta}_2\right) \\ &= 9\text{Var}\left(\widehat{\beta}_2\right).\end{aligned}$$

To estimate the variance of $\widehat{M}^4$ using the approximation given in (43), we write

$$\begin{aligned}\text{Var}\left(\widehat{M}^4\right) &\approx \text{Var}\left(60\widehat{\beta}_4 + 30\widehat{\beta}_2^2\right) \\ &= 3600\text{Var}\left(\widehat{\beta}_4\right) + 900\text{Var}\left(\widehat{\beta}_2^2\right) + 3600\text{Cov}\left(\widehat{\beta}_4, \widehat{\beta}_2^2\right).\end{aligned}$$

Thus, we estimate Var $\left(\widehat{\beta}_2^2\right)$, so consider the first two terms of the Taylor series of $\widehat{\beta}_2^2$ centered around β_2 given by

$$\widehat{\beta}_2^2 \approx \beta_2^2 + 2\beta_2\left(\widehat{\beta}_2 - \beta_2\right).$$

Taking the variance of each side yields

$$\begin{aligned}\text{Var}\left(\widehat{\beta}_2^2\right) &\approx \text{Var}\left\{\beta_2^2 + 2\beta_2\left(\widehat{\beta}_2 - \beta_2\right)\right\} \\ &= 4\beta_2^2\text{Var}\left(\widehat{\beta}_2\right) \\ &\approx 4\widehat{\beta}_2^2\text{Var}\left(\widehat{\beta}_2\right).\end{aligned}$$

To complete the estimate for $\text{Var}\left(\widehat{M}^4\right)$, we determine $\text{Cov}\left(\widehat{\beta}_4, \widehat{\beta}_2^2\right)$. Again, we use the first two terms of the Taylor series of $\widehat{\beta}_2^2$ centered around β_2 given by

$$\widehat{\beta}_2^2 \approx \beta_2^2 + 2\beta_2\left(\widehat{\beta}_2 - \beta_2\right).$$

Therefore, we write

$$\begin{aligned}
\text{Cov}\left(\widehat{\beta}_4, \widehat{\beta}_2^2\right) &\approx \text{Cov}\left\{\widehat{\beta}_4, \beta_2^2 + 2\beta_2\left(\widehat{\beta}_2 - \beta_2\right)\right\} \\
&= \text{Cov}\left(\widehat{\beta}_4, \beta_2^2\right) + \text{Cov}\left(\widehat{\beta}_4, 2\beta_2\widehat{\beta}_2\right) + \text{Cov}\left(\widehat{\beta}_4, -2\beta_2^2\right) \\
&= 2\beta_2\text{Cov}\left(\widehat{\beta}_4, \widehat{\beta}_2\right) \\
&\approx 2\widehat{\beta}_2\text{Cov}\left(\widehat{\beta}_4, \widehat{\beta}_2\right).
\end{aligned}$$

Finally, combining each of these results yields the final estimate for $\text{Var}(\widehat{M}^4)$ given by

$$\begin{aligned}
\text{Var}\left(\widehat{M}^4\right) &\approx 3600\text{Var}\left(\widehat{\beta}_4\right) + 900\left\{4\widehat{\beta}_2^2\text{Var}\left(\widehat{\beta}_2\right)\right\} + 3600\left\{2\widehat{\beta}_2\text{Cov}\left(\widehat{\beta}_4, \widehat{\beta}_2\right)\right\} \\
&= 3600\text{Var}\left(\widehat{\beta}_4\right) + 3600\widehat{\beta}_2^2\text{Var}\left(\widehat{\beta}_2\right) + 7200\widehat{\beta}_2\text{Cov}\left(\widehat{\beta}_4, \widehat{\beta}_2\right).
\end{aligned}$$

5.10. Estimating $\text{Var}\left\{\widehat{M}^4/\left(\widehat{R}_g^2\right)^2\right\}$

We determine an estimate of the variance of $\widehat{M}^4/\left(\widehat{R}_g^2\right)^2$. Consider a first-order Taylor series expansion for $\widehat{M}^4/\left(\widehat{R}_g^2\right)^2$ given by

$$\frac{\widehat{M}^4}{\left(\widehat{R}_g^2\right)^2} \approx \frac{M^4}{\left(R_g^2\right)^2} + \frac{1}{\left(R_g^2\right)^2}\left(\widehat{M}^4 - M^4\right) - \frac{M^4}{\left(R_g^2\right)^4}\left\{\left(\widehat{R}_g^2\right)^2 - \left(R_g^2\right)^2\right\}.$$

Taking the variance of each side yields

$$\begin{aligned}\operatorname{Var}\left\{\frac{\widehat{M}^4}{\left(\widehat{R}_g^2\right)^2}\right\} &\approx \operatorname{Var}\left[\frac{M^4}{\left(R_g^2\right)^2}+\frac{1}{\left(R_g^2\right)^2}\left(\widehat{M}^4-M^4\right)-\frac{M^4}{\left(R_g^2\right)^4}\left\{\left(\widehat{R}_g^2\right)^2-\left(R_g^2\right)^2\right\}\right] \\ &= \frac{1}{\left(R_g^2\right)^4}\operatorname{Var}\left(\widehat{M}^4\right)+\frac{\left(M^4\right)^2}{\left(R_g^2\right)^8}\operatorname{Var}\left\{\left(\widehat{R}_g^2\right)^2\right\} \\ &\quad -\frac{\left(M^4\right)^2}{\left(R_g^2\right)^6}\operatorname{Cov}\left\{\widehat{M}^2,\left(\widehat{R}_g^2\right)^2\right\}. \end{aligned} \tag{64}$$

To determine the final result, we estimate $\operatorname{Var}\left\{\left(\widehat{R}_g^2\right)^2\right\}$. Therefore, consider the Taylor series of $\left(\widehat{R}_g^2\right)^2$ centered around R_g^2 given by

$$\left(\widehat{R}_g^2\right)^2 \approx \left(R_g^2\right)^2+2R_g^2\left(\widehat{R}_g^2-R_g^2\right).$$

Taking the variance of both sides yields

$$\begin{aligned}\operatorname{Var}\left\{\left(\widehat{R}_g^2\right)^2\right\} &\approx \operatorname{Var}\left\{\left(R_g^2\right)^2+2R_g^2\left(\widehat{R}_g^2-R_g^2\right)\right\} \\ &= 4\left(R_g^2\right)^2\operatorname{Var}\left(\widehat{R}_g^2\right) \\ &\approx 4\left(\widehat{R}_g^2\right)^2\operatorname{Var}\left(\widehat{R}_g^2\right).\end{aligned}$$

Finally, we estimate $\operatorname{Cov}\left\{\widehat{M}^4,\left(\widehat{R}_g^2\right)^2\right\}$. Again using the Taylor series of $\left(\widehat{R}_g^2\right)^2$ centered around R_g^2 yields

$$\begin{aligned}\operatorname{Cov}\left\{\widehat{M}^4,\left(\widehat{R}_g^2\right)^2\right\} &\approx \operatorname{Cov}\left\{\widehat{M}^4,\left(R_g^2\right)^2+2R_g^2\left(\widehat{R}_g^2-R_g^2\right)\right\} \\ &= 2R_g^2\operatorname{Cov}\left(\widehat{M}^4,\widehat{R}_g^2\right)\end{aligned}$$

$$\approx 2\widehat{R}_g^2 \mathrm{Cov}\left(\widehat{M}^4, \widehat{R}_g^2\right).$$

Combining this result with (64) yields the final estimate for $\mathrm{Var}\left\{\widehat{M}^4 / \left(\widehat{R}_g^2\right)^2\right\}$,

$$\begin{aligned}
\mathrm{Var}\left\{\frac{\widehat{M}^4}{\left(\widehat{R}_g^2\right)^2}\right\} &\approx \frac{1}{\left(\widehat{R}_g^2\right)^4}\mathrm{Var}\left(\widehat{M}^4\right) + \frac{\left(\widehat{M}^4\right)^2}{\left(\widehat{R}_g^2\right)^8} 4\left(\widehat{R}_g^2\right)^2 \mathrm{Var}\left(\widehat{R}_g^2\right) \\
&\quad + \left\{\frac{-\widehat{M}^4}{\left(\widehat{R}_g^2\right)^6}\right\}\left\{2\widehat{R}_g^2\right\}\mathrm{Cov}\left(\widehat{M}^4, \widehat{R}_g^2\right) \\
&= \frac{1}{\left(\widehat{R}_g^2\right)^4}\mathrm{Var}\left(\widehat{M}^4\right) + \frac{4\left(\widehat{M}^4\right)^2}{\left(\widehat{R}_g^2\right)^6}\mathrm{Var}\left(\widehat{R}_g^2\right) - \frac{2\widehat{M}^4}{\left(\widehat{R}_g^2\right)^5}\mathrm{Cov}\left(\widehat{M}^4, \widehat{R}_g^2\right).
\end{aligned}$$

5.11. ψ Software

The ψ software program was written in R. The output contains a plot of log intensity versus s^2 with the range of data points used to estimate ψ. Furthermore, the program estimates R_g, and then uses $\widehat{R}_g$ and $\widehat{\psi}$ to estimate one or two cylinders for the molecule. Additionally, the program outputs D_{max} for each of the two cylinders. Figure 5.5 is the output of the program for one replicate the molecule nucleosome core particle.

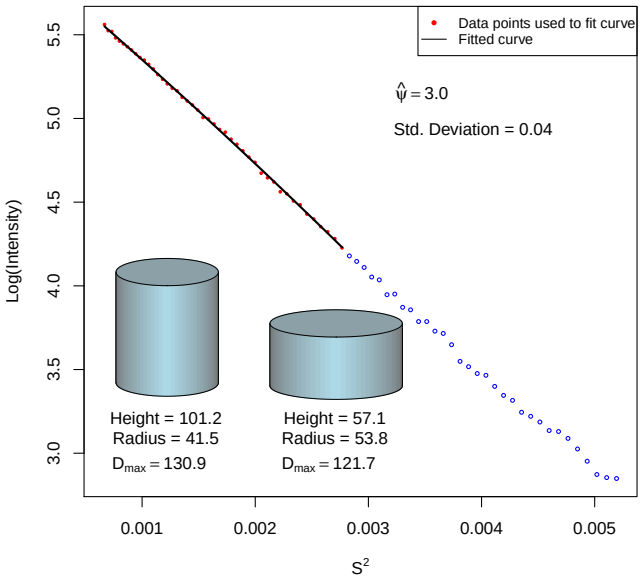


FIGURE 5.5. Output of the ψ program for the molecule nucleosome core particle.

CHAPTER 6

ESTIMATION OF CONCENTRATION RATIOS FROM SAXS EXPERIMENTS WITH APPLICATION TO DETERMINING THE RADIUS OF GYRATION

6.1. Introduction

The concentration of a molecule in solution affects the scattering of the intensity curve for small-angle X-ray scattering (SAXS) data. Figure 6.1 depicts the intensity curves for the molecule nucleosome core particle generated from four different concentrations. Overall, as concentration decreases, there is a negative vertical translation of the intensity curve; however, there may also be a concentration by angle interaction effect. That is, the intensity curves may not be parallel.

Let c_i $(i = 1, \ldots, m)$ denote true concentrations. Assuming the multiplicative model

$$I_{ij} = c_i \exp\left(\beta_0 + \beta s_j^2 + \gamma \frac{c_i}{c_1} s_j^2 + \varepsilon_{ij}\right),$$

we have

$$\ln I_{ij} = (\ln c_i + \beta_0) + \beta s_j^2 + \gamma \frac{c_i}{c_1} s_j^2 + \varepsilon_{ij}. \tag{65}$$

The model assumes that the only difference between expected intensities at two different concentrations i and i' is due to concentration; that is, there are no other differences due to exposure time, X-ray intensity, etc. If the c_i's are known without error, model (65) can be fitted using standard regression techniques. In practice, however, concentrations are not

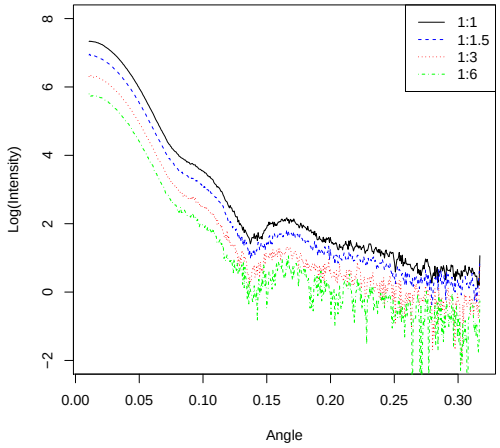


FIGURE 6.1. Log intensity curves for four different concentrations for the molecule nucleosome core particle. The concentration ratios are given in the legend.

achieved exactly due to variation in laboratory procedures (such as successive dilutions of a solution), and only nominal concentrations c_i^N are known.

It is of interest to estimate the true concentration ratios c_i/c_1 and the model (65). To this end, it is convenient to reparameterize (65) as

$$\begin{aligned} \ln I_{ij} &= \alpha_i + \beta s_j^2 + \gamma \exp\left(\ln c_i + \beta_0 - \ln c_1 - \beta_0\right) s_j^2 + \varepsilon_{ij} \\ &= \alpha_i + \beta s_j^2 + \gamma \exp\left(\alpha_i - \alpha_1\right) s_j^2 + \varepsilon_{ij}, \end{aligned} \tag{66}$$

where $\alpha_i = \ln c_i + \beta_0$. In this formulation, the α_i's (though not the c_i's) can be consistently estimated via nonlinear least squares and so the concentration ratios

$$\frac{c_i}{c_1} = \exp\left(\alpha_i - \alpha_1\right)$$

can also be consistently estimated.

6.2. Fitting the Model

Equation (66) is a nonlinear regression model, so we use iterative numerical methods rather than standard linear regression techniques to estimate the unknown parameter vector and its corresponding covariance matrix. To this end, we minimize the sum of squares given by

$$\sum_{ij} \left[\ln I_{ij} - \left\{\alpha_i + \beta s_j^2 + \gamma \exp\left(\alpha_i - \alpha_1\right) s_j^2\right\}\right]^2 \tag{67}$$

with respect to the parameters α_i, β, and γ. This minimization is accomplished using the Gauss-Newton method. We must first obtain reasonable initial values $\widehat{\alpha}_i^{(0)}, \widehat{\beta}^{(0)}$, and $\widehat{\gamma}^{(0)}$ to begin the iterative procedure. We therefore consider the model

$$\ln(I_{ij}) = \delta_i + \theta_i s_j^2 + \varepsilon_{ij} \tag{68}$$

for $i = 1, \ldots, n_c$ where n_c is the number of different concentrations, and $j = 1, \ldots, n_i$. This is a linear regression model containing different parameters for the intercept and slope for each concentration level. Observe that model (66) contains fewer parameters than model (68), making model (66) advantageous over model (68) provided it adequately fits the data. The parameter estimates for model (68) can be found using standard linear regression techniques, yielding $\widehat{\delta_i}$ and $\widehat{\theta}_i, i = 1, \ldots, n_c$.

The initial values for the α_i are given by

$$\widehat{\alpha_i}^{(0)} = \widehat{\delta_i}.$$

To determine initial values for β and γ, we use the first two concentration levels ($i = 1, 2$) and equate the coefficients of the s_j^2 terms for models (66) and (68). This results in the system of equations

$$\begin{cases} \theta_1 &= \beta + \gamma \\ \theta_2 &= \beta + \gamma \exp(\alpha_2 - \alpha_1). \end{cases}$$

Solving this system yields the initial values

$$\widehat{\gamma}^{(0)} = \frac{\widehat{\theta}_2 - \widehat{\theta}_1}{\exp\left(\widehat{\alpha}_2^{(0)} - \widehat{\alpha}_1^{(0)}\right) - 1} \quad \text{and} \quad \widehat{\beta}^{(0)} = \widehat{\theta}_1 - \widehat{\gamma}^{(0)}.$$

We next describe the iterative procedure. Define the vector of parameters

$$\boldsymbol{\beta} := [\alpha_1, \alpha_2, \ldots, \alpha_c, \beta, \gamma,]^\top.$$

The Jacobian matrix is given by

$$J(\boldsymbol{\beta}) = \begin{bmatrix} \frac{\partial f_{11}}{\partial \alpha_1} & \frac{\partial f_{11}}{\partial \alpha_2} & \cdots & \frac{\partial f_{11}}{\partial \gamma} \\ \frac{\partial f_{12}}{\partial \alpha_1} & \frac{\partial f_{12}}{\partial \alpha_2} & \cdots & \frac{\partial f_{12}}{\partial \gamma} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_{cn[c]}}{\partial \alpha_1} & \frac{\partial f_{cn[c]}}{\partial \alpha_2} & \cdots & \frac{\partial f_{cn[c]}}{\partial \gamma} \end{bmatrix},$$

where $f_{ij}(\boldsymbol{\beta}) = \alpha_i + \beta s_j^2 + \gamma \exp(\alpha_i - \alpha_1) s_j^2$. Finally, define

$$Y(\boldsymbol{\beta}) := \begin{bmatrix} \ln I_{11} - f_{11} \\ \ln I_{12} - f_{12} \\ \vdots \\ \ln I_{cn[c]} - f_{cn[c]} \end{bmatrix}.$$

The parameter estimates are then found using the iterative scheme given by

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \left\{ J(\boldsymbol{\beta}^{(m)})' J(\boldsymbol{\beta}^{(m)}) \right\}^{-1} J(\boldsymbol{\beta}^{(m)})' Y(\boldsymbol{\beta}^{(m)}).$$

Convergence can be assessed by iterating until (65) is sufficiently small to achieve a specified level of convergence yields final estimates $\widehat{\alpha_i}$, $\widehat{\beta}$, and $\widehat{\gamma}$.

Using the Jacobian matrix, we compute the variance-covariance matrix Σ for the parameter estimates from model (66) via

$$\Sigma = \sigma^2 \left(J' J \right)^{-1},$$

where $\sigma^2 = \mathrm{Var}(\varepsilon)$. Then asymptotically

$$\begin{bmatrix} \widehat{\alpha}_1 & \widehat{\alpha}_2 & \dots & \widehat{\alpha}_c & \widehat{\beta} & \widehat{\gamma} \end{bmatrix}' \sim N\left(\begin{bmatrix} \alpha_1 & \alpha_2 & \dots & \alpha_c & \beta & \gamma \end{bmatrix}', \Sigma \right).$$

The matrix Σ is estimated by

$$\widehat{\Sigma} = \widehat{\sigma}^2 \left(\widehat{J}' \widehat{J} \right)^{-1},$$

where $\widehat{J}$ is the Jacobian matrix evaluated at the final parameter estimates $\widehat{\alpha_i}$, $\widehat{\beta}$, and $\widehat{\gamma}$. The value $\widehat{\sigma}^2$ is found by evaluating (67) at the parameter estimates and dividng by the proper degrees of freedom.

The confidence interval for $c_i/c_1 = \exp(\alpha_i - \alpha_1)$ is determined by first finding the confidence interval for $\alpha_i - \alpha_1$. Thus, consider

$$\mathrm{Var}(\widehat{\alpha}_i - \widehat{\alpha}_1) = \mathrm{Var}(\widehat{\alpha}_i) + \mathrm{Var}(\widehat{\alpha}_1) - 2\mathrm{Cov}(\widehat{\alpha}_i, \widehat{\alpha}_1).$$

Using this variance formula and standard statistical techniques, we can find a confidence interval for $\alpha_i - \alpha_1$. Exponentiating the lower and upper bound for this confidence interval yields the final confidence interval for c_i/c_1.

6.3. Radius of Gyration Estimation

In addition to providing estimates of concentration ratios, an advantage of model (66) allows estimation of the radius of gyration, R_g, of the molecule while accounting for the concentration by angle interaction effect. Recall, Guinier's formula for obtaining R_g from SAXS data, given by

$$\ln \mathcal{I}(s) = \ln \mathcal{I}(0) - \frac{1}{3} R_g^2 s^2 + \mathrm{O}\left(s^4\right).$$

Thus, the approximations

$$\widehat{R_g^2} \approx -3\widehat{\beta} \quad \text{and} \quad \mathrm{Var}(\widehat{R_g^2}) \approx 9\mathrm{Var}(\widehat{\beta})$$

give the estimates for R_g^2 and its variance, respectively, using model (66).

6.4. Example using SAXS Data

In this example, we compare the estimated concentration ratios, which use only the SAXS data, to ratios obtained through an external measurement. We use the four intensity curves from Figure 6.1, each corresponding to a different concentration. The values in columns two through four of Table 6.1 are obtained external to the SAXS experiment from UV absorption spectra. It is of interest to see if the approach using only SAXS data can reproduce the concentration ratios from the independent UV absorption spectra data. Using this external measurement data, the last column of Table 6.1 is the ratio c_i^{UV}/c_1^{UV}, where c_i^{UV} is the estimate for concentration c_i, obtained independently from SAXS data.

Table 6.2 contains the parameter estimates for model (66) using the SAXS data. This table also contains an estimate for the ratio c_i/c_1 given by $\exp(\widehat{\alpha_i} - \widehat{\alpha_1})$ as well as a confidence interval. Comparing the last column of Table 6.1 with the corresponding confidence interval in Table 6.2 shows that each value fits well within the bounds of the confidence interval. Therefore the SAXS-based estimate of the ratio of concentrations is accurate and independently verified for this data set.

We may also estimate the radius of gyration of the molecule along with its standard deviation using the aforementioned procedure. This method results in the estimate $\widehat{R}_g = 42.90$Å with a standard deviation of 1.92 Å.

Table 6.1. Estimates for UV absorption spectra data for SAXS concentration data shown in Figure 6.1.

Nominal Ratio	Estimate 1	Estimate 2	Estimate 3	c_i^{UV}/c_1^{UV}
1:1	21.40	23.22	–	1
1:1.5	14.97	15.16	13.57	0.65
1:3	7.97	7.16	7.17	0.33
1:6	4.16	4.06	4.45	0.19

TABLE 6.2. Results of fitting model (66) to SAXS concentration data shown in Figure 6.1.

Parameter	Estimate	$\exp(\widehat{\alpha_i} - \widehat{\alpha_1})$	CI of $\exp(\alpha_i - \alpha_1)$
α_1	7.50	1	–
α_2	7.07	0.65	(0.56,0.76)
α_3	6.48	0.36	(0.29,0.44)
α_4	5.93	0.21	(0.17,0.26)
β	-607.13	–	–
γ	-7.05	–	–

6.5. Example using Replicate SAXS Data

In the following example we have data for the molecule nucleosome core particle. This data contains intensity curves for this molecule at five different concentration levels. A plot of the data is shown in Figure 6.2. This picture illustrates the vertical shift of the log intensity curves due to the different concentrations.

Table 6.3 contains the parameter estimates from model (66) along with their corresponding standard deviation estimates. The fourth column in Table 6.3 contains the estimate of the concentration ratios, $\widehat{c}_i/\widehat{c}_1$, based on the model. The fifth column in Table 6.3 contains the corresponding confidence interval for c_i/c_1. The last column of Table 6.3 consists of the UV concentration ratios, c_i^{UV}/c_1^{UV}. All of these UV concentration ratios are within the confidence interval for the estimated concentration ratios in Table 6.3 except for the fifth estimate, which is just outside the confidence interval.

For this set of SAXS data, we also estimate the radius of gyration of the molecule using the aforementioned procedure. The estimate of R_g is 41.50 Å with a standard deviation of 1.54.

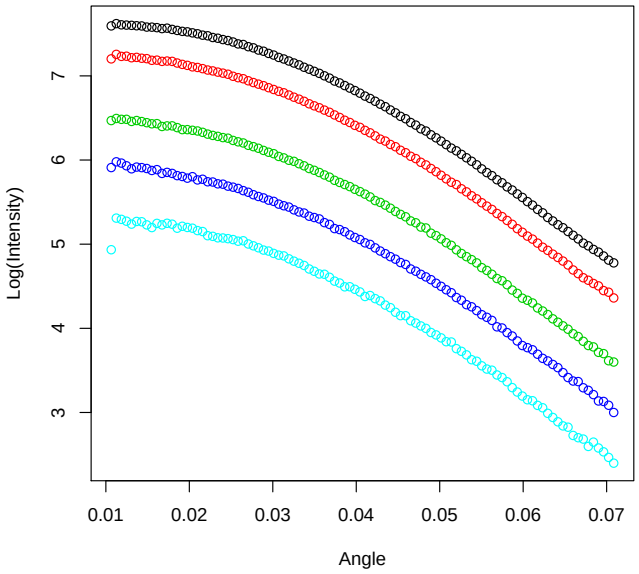


FIGURE 6.2. Log intensity curves for five different concentrations for the molecule nucleosome core particle.

TABLE 6.3. Results of fitting model (66) to SAXS concentration data shown in Figure 6.2.

Parameter	Estimate	Std. Dev.	$\exp(\widehat{\alpha}_i - \widehat{\alpha}_1)$	CI of $\exp(\alpha_i - \alpha_1)$	c_i^{UV}/c_1^{UV}
α_1	7.77	0.10	1	–	1
α_2	7.37	0.08	0.67	(0.63,0.71)	0.66
α_3	6.60	0.08	0.31	(0.29,0.33)	0.33
α_4	6.04	0.09	0.18	(0.17,0.19)	0.17
α_5	5.42	0.09	0.09	(0.08,0.10)	0.06
β	-612.11	48.30	–	–	–
γ	2.87	85.45	–	–	–

6.6. EXAMPLE WITH CONCENTRATION-DEPENDENT DATA

In the following example, we have more data for the molecule NAP. This data contains intensity curves for this molecule at four different concentration levels, and a plot of the data is shown in Figure 6.3. This picture illustrates the vertical shift of the log intensity curves

due to the different concentrations. There is also a subtle concentration by angle interaction effect.

Table 6.4 contains the parameter estimates from model (66) along with their corresponding standard deviation estimates. The fourth column in Table 6.4 contains the estimate of the concentration ratios based on the model. The last column in Table 6.4 contains the corresponding confidence interval for c_i/c_1.

For this set of SAXS data, we also estimate the radius of gyration of the molecule using the aforementioned procedure. The estimate of R_g is 37.20 Å with a standard deviation of 1.81.

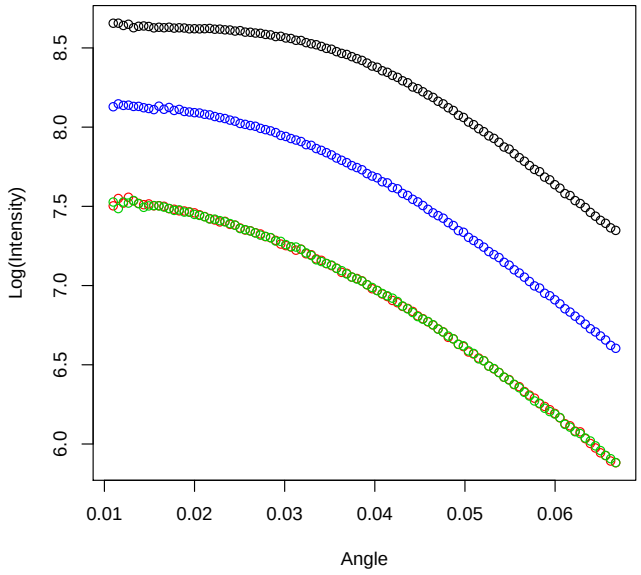


FIGURE 6.3. Log intensity curves for four different concentrations for the molecule NAP.

TABLE 6.4. Results of fitting model (66) to SAXS concentration data shown in Figure 6.3.

Parameter	Estimate	Std. Dev.	$\exp(\widehat{\alpha_i} - \widehat{\alpha_1})$	CI of $\exp(\alpha_i - \alpha_1)$
α_1	8.76	0.05	1	–
α_2	7.61	0.04	0.32	(0.28,0.36)
α_3	7.61	0.04	0.32	(0.28,0.36)
α_4	8.23	0.04	0.59	(0.53,0.65)
β	-461.32	44.80	–	–
γ	188.05	69.67	–	–

6.7. CONCLUSIONS

We devise a model for estimating concentration ratios in the presence of measurement error for concentration, while including a concentration by angle interaction effect in a unified statistical model that requires only standard SAXS data. This model is validated empirically with data for which we have external measurements of concentrations; however, these external measurements are not needed for the model. This model also provides the ability to estimate the molecule's radius of gyration with an estimate of its standard error.

BIBLIOGRAPHY

[1] K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond, "Crystal structure of the nucleosome core particle at 2.8 å resolution," *Nature*, vol. 389, no. 6648, pp. 251–260, 1997.

[2] B. Rupp, *Protein Crystallization Strategies for Structural Genomics*, ch. A Guide to Automation and Data Handling in Protein Crystallization, pp. 9–49. International University Line, 2007.

[3] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne, "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, pp. 235–242, 2000.

[4] R. Maurus, C. M. Overall, R. Bogumil, Y. Luo, A. G. Mauk, M. Smith, and G. D. Brayer, "A myoglobin variant with a polar substitution in a conserved hydrophobic cluster in the heme binding pocket," *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology*, vol. 1341, no. 1, pp. 1–13, 1997.

[5] C. Putnam, M. Hammel, G. Hura, and J. Tainer, "X-ray solution scattering (SAXS) combined with crystallography and computation: Defining accurate macromolecular structures, conformations and assemblies in solution.," *Quarterly Reviews of Biophysics*, vol. 40, no. 3, pp. 191–285, 2007.

[6] Y. Lu, J. C. M., and J. Trewhella, "Invited review: Probing the structure of muscle regulatory proteins using small-angle solution scattering.," *Biopolymers*, vol. 95, no. 8, pp. 505–516, 2011.

[7] R. P. Rambo and J. A. Tainer, "Super-resolution in solution X-ray scattering and its applications to structural systems biology," *Annual Review of Biophysics*, vol. 42, pp. 415–441, 2013.

[8] G. L. Hura, A. L. Menon, M. Hammel, R. P. Rambo, F. L. Poole Ii, S. E. Tsutakawa, F. E. Jenney Jr, S. Classen, K. A. Frankel, R. C. Hopkins, *et al.*, "Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS)," *Nature Methods*, vol. 6, no. 8, pp. 606–612, 2009.

[9] G. L. Hura, H. Budworth, K. N. Dyer, R. P. Rambo, M. Hammel, C. T. McMurray, and J. A. Tainer, "Comprehensive macromolecular conformations mapped by quantitative SAXS analyses," *Nature Methods*, vol. 10, no. 6, pp. 453–454, 2013.

[10] O. Glatter, "A new method for the evaluation of small-angle scattering data," *Journal of Applied Crystallography*, vol. 10, no. 5, pp. 415–421, 1977.

[11] S. Hansen, "Bayesian estimation of hyperparameters for indirect Fourier transformation in small-angle scattering," *Journal of Applied Crystallography*, vol. 33, no. 6, pp. 1415–1421, 2000.

[12] D. I. Svergun, M. V. Petoukhov, and M. H. Koch, "Determination of domain structure of proteins from X-ray solution scattering.," *Biophysical Journal*, vol. 80, pp. 2946–2953, June 2001.

[13] B. D. Bugbee, F. J. Breidt, and M. J. van der Woerd, "Laplace variational approximation for semiparametric regression in the presence of heteroskedastic errors," *Journal of Computational and Graphical Statistics*, no. just-accepted, pp. 00–00, 2014.

[14] A. Guinier, "La diffraction des rayons X aux très petits angles: Application a l'Étude de phénoménes ultramicroscopiques," *Ann. Phys., 11e série*, vol. 12, pp. 161–237, 1939.

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

[16] J. Geweke and S. Porter-Hudak, "The estimation and application of long memory time series models," *Journal of Time Series Analysis*, vol. 4, no. 4, pp. 221–238, 1983.

[17] A. Guinier and G. Fournet, *Small-Angle Scattering of X-rays*. New York: John Wiley & Sons, 1955.

[18] M. V. Petoukhov, P. V. Konarev, A. G. Kikhney, and D. I. Svergun, "*ATSAS* 2.1 – towards automated and web-supported small-angle scattering data analysis," *Journal of Applied Crystallography*, vol. 40, pp. s223–s228, Apr 2007.

[19] F. J. Breidt, A. Erciulescu, and M. van der Woerd, "Autocovariance structures for radial averages in small-angle X-ray scattering experiments," *Journal of Time Series Analysis*, vol. 33, no. 5, pp. 704–717, 2012.

[20] J. Neter, M. H. Kutner, C. J. Nachtsheim, and W. Wasserman, *Applied linear statistical models*, vol. 4. Irwin Chicago, 1996.

[21] D. I. Svergun and M. H. Koch, "Small-angle scattering studies of biological macromolecules in solution," *Reports on Progress in Physics*, vol. 66, no. 10, p. 1735, 2003.

[22] M. D. Morris and S. F. Ebey, "An interesting property of the sample mean under a first-order autoregressive model," *The American Statistician*, vol. 38, no. 2, pp. 127–129, 1984.

[23] M. L. Stein, "Asymptotic properties of centered systematic sampling for predicting integrals of spatial processes," *The Annals of Applied Probability*, pp. 874–880, 1993.

[24] S. N. Lahiri, "On inconsistency of estimators based on spatial data under infill asymptotics," *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 403–417, 1996.

[25] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods*. New York: Springer, 2nd ed., 1991.

[26] J. Chen and A. K. Gupta, *Parametric statistical change point analysis: With applications to genetics, medicine, and finance*. Springer Science & Business Media, 2011.

[27] R. Killick and I. Eckley, "Changepoint: an R package for changepoint analysis," *Journal of Statistical Software*, vol. 58, no. 3, pp. 1–19, 2014.

[28] D. Svergun, C. Barberato, and M. Koch, "CRYSOL - a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates," *J. Appl. Cryst.*, vol. 28, pp. 768–773, 1995.

[29] M. Brehove, T. Wang, J. North, Y. Luo, S. J. Dreher, J. C. Shimko, J. J. Ottesen, K. Luger, and M. G. Poirier, "Histone core phosphorylation regulates dna accessibility," *Journal of Biological Chemistry*, vol. 290, pp. 22612–22621, September 2015.

[30] D. Bingham, E. Schoen, and R. Sitter, "Designing fractional factorial split-plot experiments with few whole-plot factors," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 53, no. 2, pp. 325–339, 2004.

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

[32] D. I. Svergun, "Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing," *Biophysical Journal*, vol. 77, pp. 2879–2886, November 1999.

[33] Y. G. J. Sterckx, A. N. Volkov, W. F. Vranken, J. Kragelj, M. Ringkjøbing Jensen, L. Buts, A. Garcia-Pino, T. Jové, L. van Melderen, M. Blackledge, N. A. J. van Nuland, and R. Loris, "Small-angle X-ray scattering- and nuclear magnetic resonance-derived

conformational ensemble of the highly flexible anti-toxin PaaA2," *Structure*, vol. 22, pp. 854–865, June 2014.

[34] J. Zhang, C. P. Jones, and A. R. Ferré-D'Amaré, "Global analysis of riboswitches by small-angle X-ray scattering and calorimetry," *Biochimica et Biophysica Acta*, vol. 1839, pp. 1020–1029, October 2014.

[35] P. V. Konarev, G. S. Kachalova, A. Y. Ryazanova, E. A. Kubareva, A. S. Karyagina, H. D. Bartunik, and D. I. Svergun, "Flexibility of the linker between the domains of DNA methyltransferase SsoII revealed by small-angle X-ray scattering: implications for transcription," *PloS one*, vol. 9, no. 4, p. e93453, 2014.

[36] R. Killick and I. A. Eckley, "Changepoint: an R package for changepoint analysis," *R package version 0.6, URL http://CRAN. R-project. org/package= changepoint*, 2011.

[37] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[38] R. Shibata, "Selection of the order of an autoregressive model by Akaike's information criterion," *Biometrika*, vol. 63, no. 1, pp. 117–126, 1976.

[39] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods*. Springer, 2009.

[40] G. Casella and R. L. Berger, *Statistical inference*, vol. 2. Duxbury Pacific Grove, CA, 2002.

[41] R. Maurus, C. Overall, R. Bogumil, Y. Luo, A. Mauk, M. Smith, and G. Brayer, "A myoglobin variant with a polar substitution in a conserved hydrophobic cluster in the heme binding pocket.," *Biochimica et Biophysica Acta*, vol. 1341, pp. 1–13, August 1997.

[42] H. Drew, R. Wing, T. Takano, C. Broka, S. Tanaka, I. K., and R. Dickerson, "Structure of a B-DNA dodecamer: conformation and dynamics.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 78, pp. 2179–2183, April 1981.

[43] P. N. Dyer, R. S. Edayathumangalam, C. L. White, Y. Bao, S. Chakravarthy, U. M. Muthurajan, and K. Luger, "Reconstitution of nucleosome core particles from recombinant histones and DNA," in *Chromatin and Chromatin Remodeling Enzymes, Part A* (C. D. Allis and C. Wu, eds.), vol. 375 of *Methods in Enzymology*, pp. 23 – 44, Academic Press, 2003.

[44] C. Yang, M. van der Woerd, U. Muthurajan, J. Hansen, and K. Luger, "Biophysical analysis and small-angle X-ray scattering-derived structures of MeCP2-nucleosome complexes," *Nucleic Acids Research*, vol. 39, pp. 4122–4135, May 2011.

[45] R. A. Johnson, D. W. Wichern, *et al.*, *Applied multivariate statistical analysis*, vol. 4. Prentice hall Englewood Cliffs, NJ, 1992.

APPENDIX A

IMPLEMENT R_G PROGRAM

A.1. R_g Software

The files described in this document are available at

`http://hdl.handle.net/10217/167285`.

A.2. Download R

In order to run this program you need to use the free, publicly available program "R". R is a commonly used programming language for statistical computing and graphics. To download R, visit the site "http://cran.us.r-project.org" and follow the instructions.

A.3. Set Up the Estimation Routines and Examples

(1) Start the application R.

(2) Within R, the working directory must first be changed in order to conveniently select the data. Click "File" > "Change dir..." and then select the folder containing the SAXS intensity curve data. Now it is simple to use any data sets in this folder.

(3) Open the files file1.R, file2.R and file3.R from the R drop-down menu: click "File" > "Open script" and then navigate to the correct file location and click "Open." Do this for each file.

(4) If this is the first time you have run this program in R, you must install the R package "changepoint." From the R drop-down menu:

   (a) Click "Packages" > "Install packages."

(b) Select a geographic location from the "CRAN mirror" menu that pops up; it is best to select a location near you for fast download speed. Click "OK."

(c) Select the "changepoint" package from the "packages" menu that pops up; click "OK." The package will automatically download and install.

Once the package has been installed, you do not need to repeat Step 4 upon subsequent runs.

(5) Highlight everything in file1.R and run the code (Ctrl+A, Ctrl+R for Windows machines).

The program is now ready to analyze the example data sets or user-supplied SAXS data. Data in the input file must be organized into three columns, delimited by spaces or tabs if using a text file (.txt, .dat, etc.), or by commas for a .csv file. The columns must contain the following data in the following order:

angle (s) intensity standard deviation

The second column should NOT contain log intensity.

The R code in file2.R describes the analysis of a single replicate, using a sample SAXS data set for the molecule ovalbumin. The code also describes alternate file formats. See Section A.4 Section A.5 below. The R code in file3.R describes the analysis of multiple replicates, using 10 sample SAXS data sets for the molecule myoglobin. See Section A.6 below.

A.4. Single Replicate Example: User-Specified Initial Angle

Included in this folder is a sample SAXS data set for the molecule ovalbumin. The following R code is included in file2.R. The code runs the analysis using the ovalbumin data

set and should yield Figure A.1, Figure A.2, and Figure A.3. The function `estimate_Rg` has three arguments: the first argument is the name of the data object read from the file, the second argument is the number of replicates, and the third (optional) argument indicates the index i of the initial angle s_i to be used in the analysis (that is, excluding the first $i-1$ data points near zero from the analysis). If the third argument is not included, then the program defaults to automatically determining any initial outlying data points.

```
data = read.table("oval_01C_S008_0_01.dat", header = FALSE)
estimate_Rg(data, 1, 5)
```

The program output is three plots (one may be concealed by the other) containing several pieces of information:

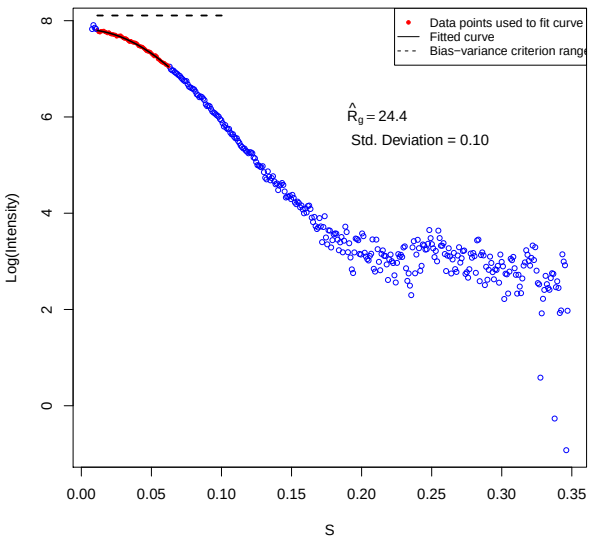


FIGURE A.1. Plot of log intensity vs. s with the estimated R_g value and its standard deviation for a single replicate of ovalbumin.

A.4.1. PLOT OF LOG-INTENSITY VERSUS ANGLE.

- Data points (open blue dots and solid red dots) represent log intensity vs. angle s of the input data; this plot can be used to ensure the input data are correct.

- Specifically, the solid red data points are those that have been chosen for use in curve fitting by minimizing the bias-variance criterion.
- A quadratic fit of the solid red data points is indicated by the solid black curve. This curve is used to estimate R_g and its standard deviation. This curve does not need to fit the data perfectly; some bias is acceptable in return for smaller standard deviation.
- The resulting estimates of R_g and its standard deviation are given.
- A black horizontal dashed line indicates the range of possible values over which the bias-variance criterion is optimized.

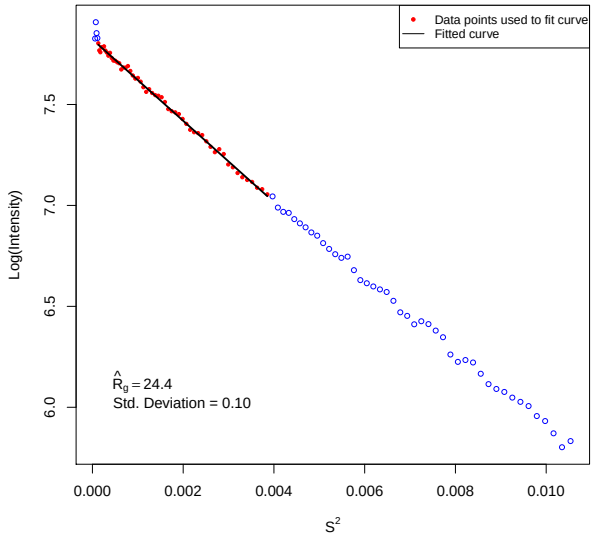


FIGURE A.2. Plot of log intensity vs. s^2 with the estimated R_g value and its standard deviation for a single replicate of ovalbumin.

A.4.2. PLOT OF LOG-INTENSITY VERSUS SQUARED ANGLE.

- Data points (open blue dots and solid red dots) represent log intensity vs. angle squared s^2 of the input data over which the bias-variance criterion is optimized.
- Specifically, the solid red data points are those that have been chosen for use in curve fitting by minimizing the bias-variance criterion.

- A fit of the solid red data points is indicated by the solid black line. This line is used to estimate R_g and its standard deviation. This line does not need to fit the data perfectly; some bias is acceptable in return for smaller standard deviation.
- The resulting estimates of R_g and its standard deviation are given.

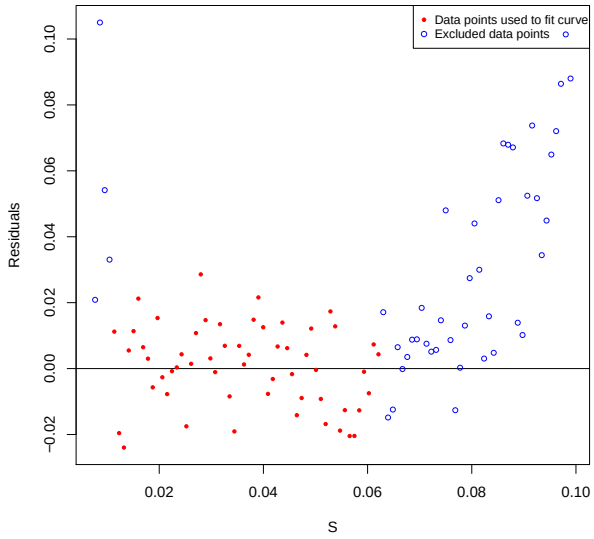


FIGURE A.3. Plot of residuals vs. s for a single replicate of ovalbumin.

A.4.3. PLOT OF RESIDUALS FROM QUADRATIC FIT.

- The points (open blue dots and solid red dots) represent residuals vs. angle s of the input data; this residual plot can be used to ensure that the data window is a reasonable fit.
- Specifically, the solid red data points are those that have been chosen for use in curve fitting by minimizing the bias-variance criterion, and the open blue dots are not used in the fit.

To save a plot as a PDF file, first select the plot by clicking on it. Then, from the R drop-down menu, click "File" > "Save as" > "PDF..." Then select the save location, enter a name for the file, and click "Save."

A.5. Single Replicate Example: Automatic Selection of Initial Angle

By default, the program will automatically determine any initial outlying data points using a modified DFBETAS procedure. If the user does not enter any values for the initial angle, then the program will determine these points automatically and output the number of points removed from the curve. The following R code is included in file2.R. It runs the program using the data set for the molecule ovalbumin and should yield Figure A.4, Figure A.5, and Figure A.6.

```
data = read.table("oval_01C_S008_0_01.dat", header = FALSE)

estimate_Rg(data, 1)
```

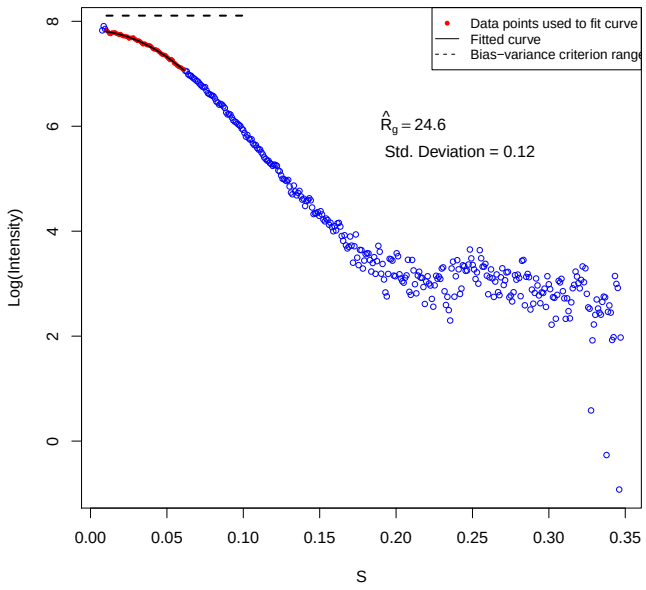


Figure A.4. Plot of log intensity vs. s with the estimated R_g value and its standard deviation for a single replicate of ovalbumin with automatic outlier detection.

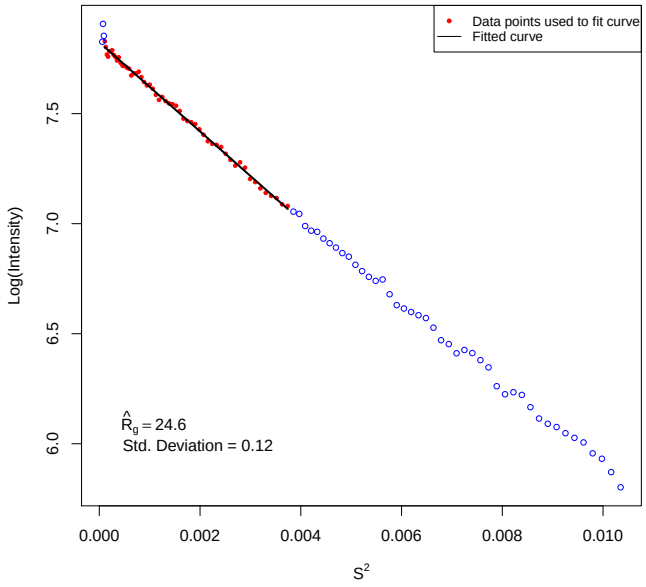


FIGURE A.5. Plot of log intensity vs. s^2 with the estimated R_g value and its standard deviation for a single replicate of ovalbumin with automatic outlier detection.

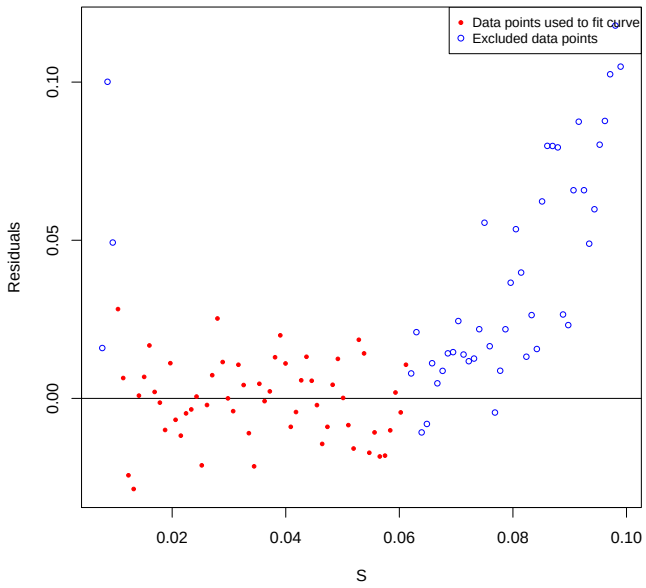


FIGURE A.6. Plot of residuals vs. s for a single replicate of ovalbumin with automatic outlier detection.

A.6. Multiple Replicates Example

An important innovation of this new procedure is the ability to incorporate replicate SAXS intensity curves to determine a more accurate and precise estimate of R_g and its variance. The code below is included in file3.R and demonstrates how to apply this program with replicate data for the molecule myoglobin. First, the replicate data are read in from 10 different files:

```
data1 = read.table("myo2_07D_S215_0_01.dat", header = FALSE)
data2 = read.table("myo2_07D_S215_0_02.dat", header = FALSE)
data3 = read.table("myo2_07D_S215_0_03.dat", header = FALSE)
data4 = read.table("myo2_07D_S215_0_04.dat", header = FALSE)
data5 = read.table("myo2_07D_S215_0_05.dat", header = FALSE)
data6 = read.table("myo2_07D_S215_0_06.dat", header = FALSE)
data7 = read.table("myo2_07D_S215_0_07.dat", header = FALSE)
data8 = read.table("myo2_07D_S215_0_08.dat", header = FALSE)
data9 = read.table("myo2_07D_S215_0_09.dat", header = FALSE)
data10= read.table("myo2_07D_S215_0_10.dat", header = FALSE)
```

See file2.R for alternate file formats.

Next, the data are combined into a matrix with the following columns in the following order:

angle (s), intensity for first replicate, ..., intensity for last replicate

In this example, we first combine all data into a matrix with all ten replicates, then use subsets of the data to illustrate estimation with one, three, and ten replicates:

```
# For illustration, look at one replicate, three replicates, and ten
replicates.
# First combine the data into one big ten-replicate matrix.
# Keep angle and intensity from replicate 1 (columns 1 and 2 but not 3),
# intensity from replicate 2 (column 2 only),
# intensity from replicate 3 (column 2 only),...,
# intensity from replicate 10 (column 2 only).
#
combined_data = cbind(data1[1:400,-3],data2[1:400,2],data3[1:400,2],
data4[1:400,2],data5[1:400,2],data6[1:400,2],
data7[1:400,2],data8[1:400,2],data9[1:400,2],
data10[1:400,2])
```

It remains to specify the initial angle, or let it be selected via automatic outlier detection. In this example, we specify in each case (one, three or ten replicates) that no points are to be deleted:

```
# Run the estimation code with one replicate
# (only the first two columns of the combined data), with no points deleted:
estimate_Rg(combined_data[,1:2], 1, 1)
# Run the estimation code with three replicates
# (only the first four columns of the combined data), with no points
deleted:
estimate_Rg(combined_data[,1:4], 3, rep(1,3))
# Run the estimation code with all ten replicates
```

```
# (all eleven columns of the combined data), with no points deleted:

estimate_Rg(combined_data, 10, rep(1,10))
```

Table A.1 summarizes the results. In each case, the estimate of R_g is similar but using more replicates increases the precision of the estimate.

As with the single replicates case, the program output is three plots (one may be concealed by the other). The only difference is that all the replicate data are plotted. See Figure A.7, Figure A.8, and Figure A.9 for the case of ten replicates.

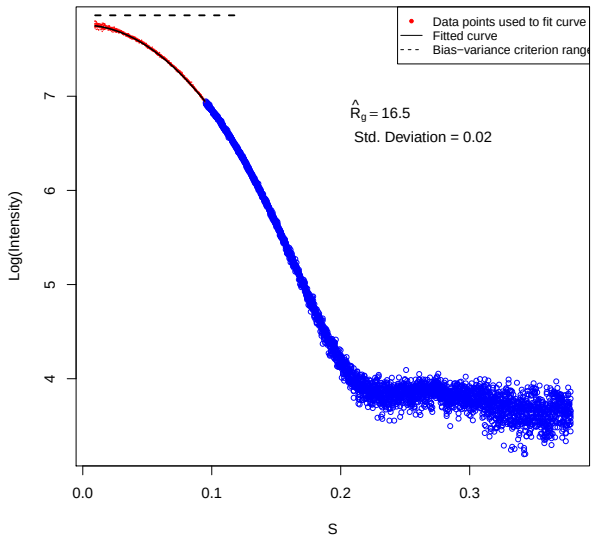


FIGURE A.7. Plot of log intensity vs. s with the estimated R_g value and its standard deviation for ten replicates of myoglobin.

TABLE A.1. Results for estimating R_g using the new procedure for the molecule myoglobin with one, three, and ten replicate SAXS intensity curves. In each case, $\widehat{R}_g$ and its standard deviation are given.

Replicates	$\widehat{R}_g$	SD($\widehat{R}_g$)
1	16.469	0.030
3	16.452	0.029
10	16.452	0.019

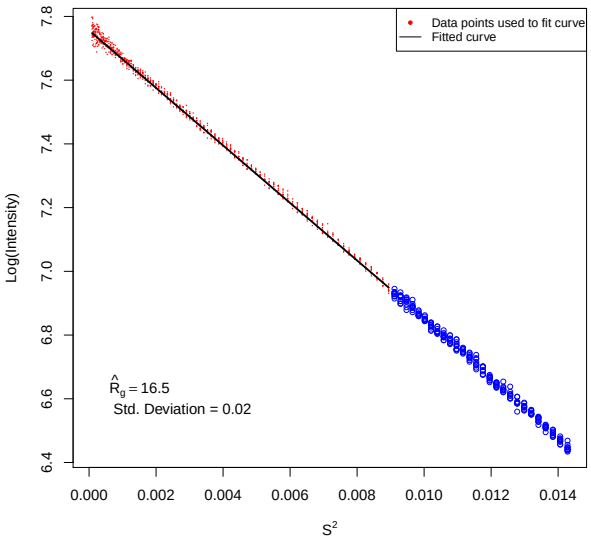


FIGURE A.8. Plot of log intensity vs. s^2 with the estimated R_g value and its standard deviation for ten replicates of myoglobin.

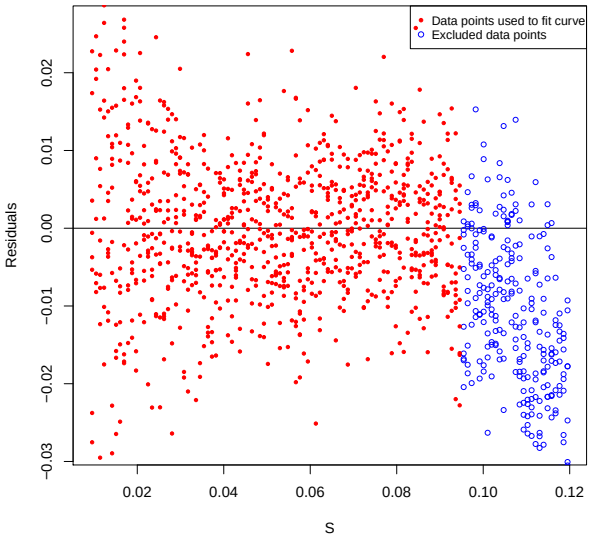


FIGURE A.9. Plot of residuals vs. s for ten replicates of myoglobin.

The third argument can be altered to select different initial points from each replicate. For example, if you want to eliminate the first three points of the fourth replicate while deleting no points from the other nine replicates, you would use the following code:

```
estimate_Rg(combined_data,10,c(1,1,1,4,1,1,1,1,1,1))
```

Alternatively, you can delete the third argument, in which case the program uses the modified DFBETAS criterion as an outlier detection algorithm to determine one, common initial point for all of of the replicate curves. For automatic selection of the common initial point in the example with three replicates, use

```
estimate_Rg(combined_data[,1:4],3)
```

Similarly, for automatic selection of the common initial point in the example with ten replicates, use

```
estimate_Rg(combined_data, 10)
```

A.7. Problems and Possible Solutions

Problem: The following error message appears when executing the program. "Error in [.data.frame(M, , 2) : undefined columns selected"

Solution: Make sure there is only a one-line header in the data file.

Problem: The following warning message appears when reading in a file. "Incomplete final line found by readTableHeader on 'filename'"

Solution: The final line of your text or CSV doesn't have a line feed or carriage return.

APPENDIX B

USING MODIFIED DFBETAS TO DETECT OUTLIERS

B.1. DFBETAS Criterion for $\widehat{R}_g^2$

We have developed an automated statistical procedure to detect outliers, by adapting the standard DFBETAS criterion (e.g., [20], §10.4) to estimation of R_g^2 under model (10). Specifically, we compute

$$\text{DFBETAS}\left(\widehat{R}_{g(-a)}^2\right) = \frac{\widehat{R}_g^2 - \widehat{R}_{g(-a)}^2}{\text{SE}\left(\widehat{R}_g^2\right)}.$$

where $\widehat{R}_{g(-a)}^2$ deletes the first a observations and uses only the angles $s_{a+1}, s_{a+2}, \ldots, s_n$. Therefore, we can remove one outlying point at a time or groups of observations. We remove values if the absolute value of DFBETAS exceeds two, or if it exceeds a size-adjusted cutoff value of $2/\sqrt{\max\{n, n_{(-a)}\}}$, where $n_{(-a)}$ is the number of points used to calculate $\widehat{R}_{g(-a)}^2$.

B.2. DFBETAS Criterion for $\widehat{\psi}$

We are also interested in determining the influence of the ith data point on the estimate of $\widehat{\psi}$. Thus, similarly to our definition of the DFBETAS criterion for $\widehat{R}_g^2$ we define the DFBETAS value for $\widehat{\psi}$ as

$$\text{DFBETAS}\left(\widehat{\psi}_{(-a)}\right) = \frac{\widehat{\psi} - \widehat{\psi}_{(-a)}}{\text{SE}\{\widehat{\psi}\}},$$

where a is the number of consecutive data points deleted in calculating $\widehat{\psi}$. The analogous cutoff value is $2/\sqrt{\max\{n, n_{(-a)}\}}$, where n is the number of data points used to determine

$\widehat{\psi}$ and $n_{(-a)}$ is the number of data points used to determine $\widehat{\psi}_{(-a)}$. In the following section, we will describe simulations to test the new DFBETAS criterion for $\widehat{\psi}$.

B.3. Outlier Simulation Results

To test the performance of the new outlier detection method for $\widehat{R}_g^2$ and $\widehat{\psi}$, we will create a simulation comparing this new outlier detection method to the standard method with no outlier detection. For this simulation, we fit a cubic spline to the model of the molecule, and we will add noise to this fit. Figure B.1 is a plot of s versus log intensity for the molecule myoglobin. The standard method estimates R_g^2 and ψ starting at the first point of the curve, and the new method determines which initial points are outliers before determining the final estimates R_g^2 and ψ. Once the initial outlying points are determined, the new method estimates R_g^2 and ψ the same as the standard algorithm.

There are two common types of outlying behavior that occur with SAXS data. The first type is outlying trend behavior that can be caused by aggregation or interparticle interference. Figure B.2 contains a plot of strong and weak versions of this type of outlying trend. The second type of outlying behavior is single outlying points which is often the result of beam stop scatter. Figure B.3 is an example of this type of single point outlying behavior.

Outlying trend behavior is simulated as follows. A knot is chosen Unif$(5, 20)$, and the coefficient of the first spline is then given random variability that increases or decreases the outlying trend for the initial part of the curve. Finally, appropriate noise is added to the curve and then both methods are used to determine $\widehat{R}_g$ and $\widehat{\psi}$. Using this procedure we are able to test the new outlier detection algorithm on models of molecules. Figure B.4 shows the results of the outlier detection method for $\widehat{\psi}$ for a specific example of outlying trend.

For both types of outlying behavior, we will compare the root MSE values of $\widehat{R}_g$ and $\widehat{\psi}$ using both the regular estimation method without outlier detection and the new outlier detection method. Table B.1 contains the results for trend outlying behavior, and Table B.2 shows the results for single point outlying behavior. In both cases, the root MSE is much smaller for the procedure with the new outlier detection method.

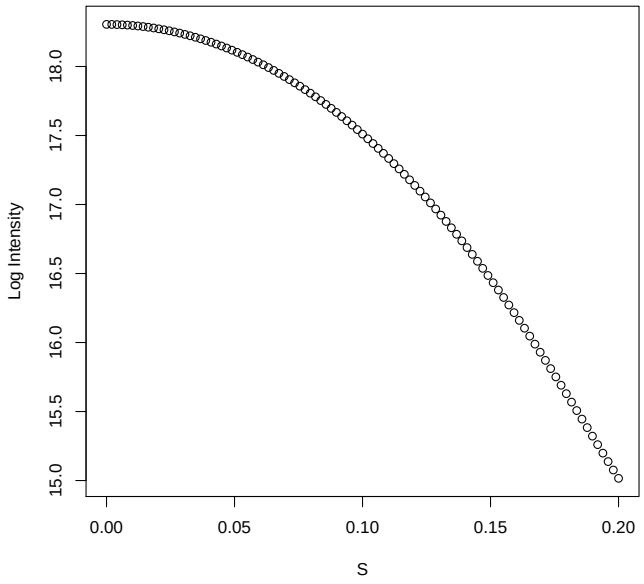


FIGURE B.1. Plot of the theoretical intensity curve for the molecule myoglobin.

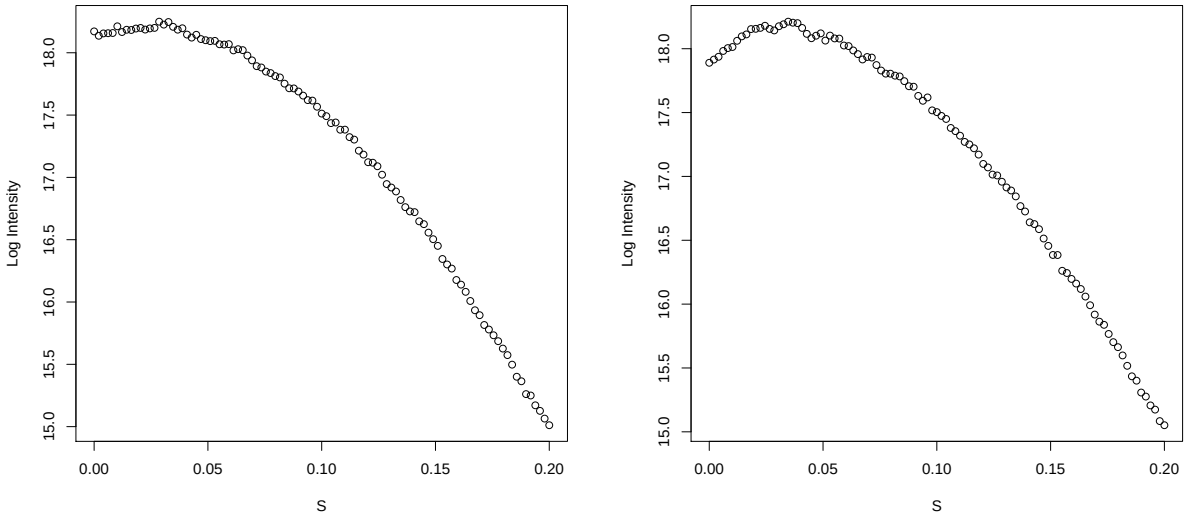


FIGURE B.2. Left: Plot of simulated experimental data with weak outlying trend for the molecule myoglobin. Right: Plot of simulated experimental data with strong outlying trend for the molecule myoglobin.

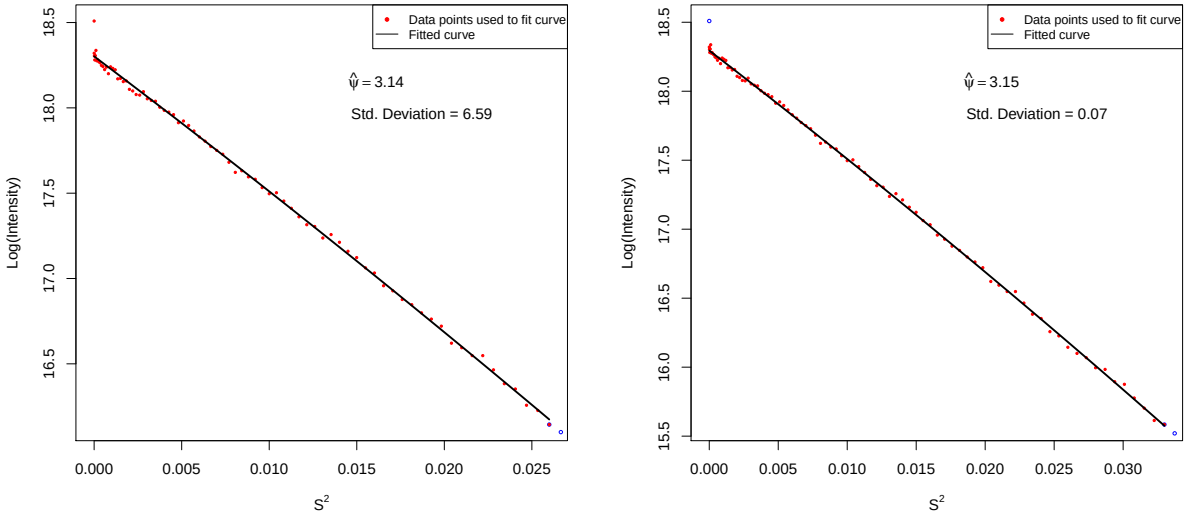


FIGURE B.3. Left: Standard method for a single outlying point. Right: Standard method for a single outlying point.

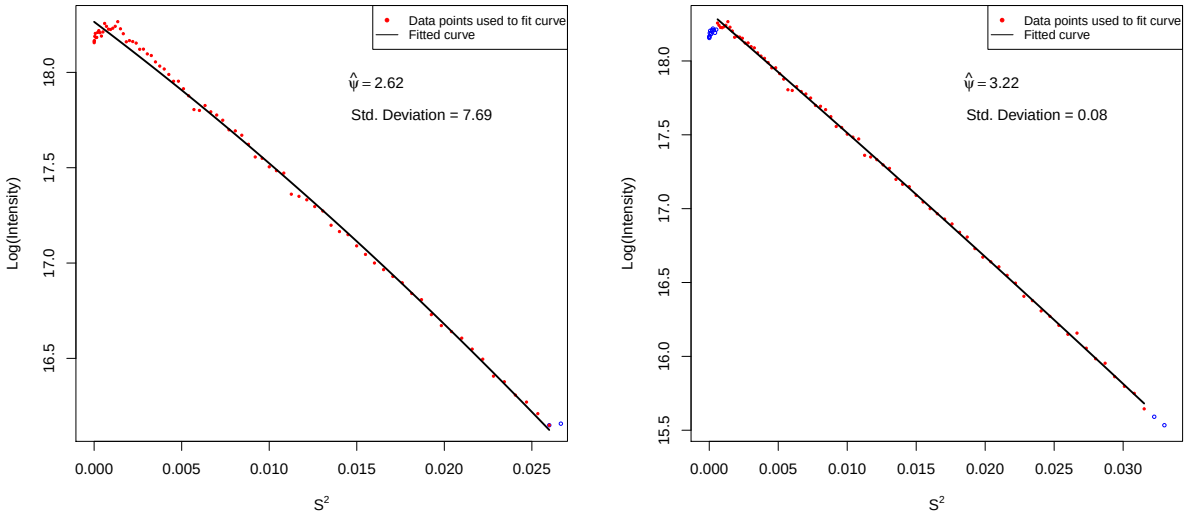


FIGURE B.4. Left: New outlier diagnostics method for a single outlying point. Right: Standard method for outlying trend.

TABLE B.1. Results comparing the root MSE of $\widehat{R}_g$ and $\widehat{\psi}$ using the regular estimation method without outlier detection and the new outlier detection method. Simulation results are based on a sample size of 1000 for the molecule myoglobin with trend outlying behavior.

	Reg. $\widehat{\psi}$	Out. $\widehat{\psi}$	Reg. $\widehat{R}_g$	Out. $\widehat{R}_g$
RMSE	0.4833	0.1275	1.0097	0.2610

TABLE B.2. Results comparing the root MSE of $\widehat{R}_g$ and $\widehat{\psi}$ using the regular estimation method without outlier detection and the new outlier detection method. Simulation results are based on a sample size of 1000 for the molecule myoglobin with single point outlying behavior.

	Reg. $\widehat{\psi}$	Out. $\widehat{\psi}$	Reg. $\widehat{R}_g$	Out. $\widehat{R}_g$
RMSE	0.1950	0.0846	0.4200	0.1814

We also conduced a similar simulation comparing the two procedures on data with no outlying behavior and data with both outlying trend and single point outliers. Figure B.5 contains an example of each of these two simulation setups. Table B.3 contains the results

comparing the two procedures on a simulation with no outlying behavior. The outlier detection method performed slightly better, but both procedures have comparable root MSE values. Finally, Table B.4 compares the two procedures for a simulation with both trend and single point outlying behavior, and the outlier detection method yields smaller root MSE values, as desired.

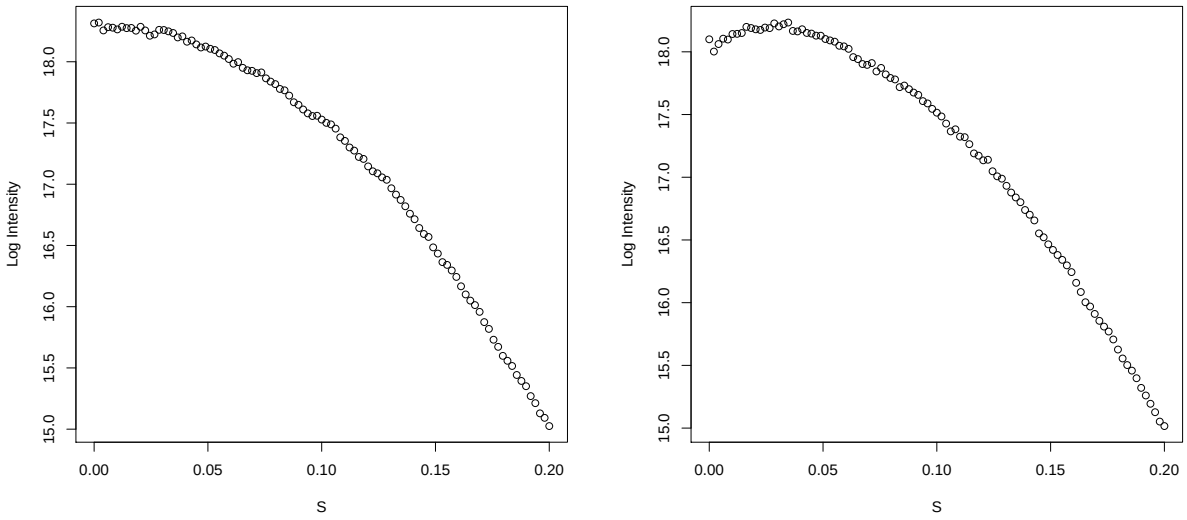


FIGURE B.5. Left: Molecule myoglobin with no outlying behavior. Right: Molecule myoglobin with both outlying trend and a single point outlying behavior.

TABLE B.3. Results comparing the root MSE of $\widehat{R}_g$ and $\widehat{\psi}$ using the regular estimation method without outlier detection and the new outlier detection method. Simulation results are based on a sample size of 1000 for the molecule myoglobin with no outlying behavior.

	Reg. $\widehat{\psi}$	Out. $\widehat{\psi}$	Reg. $\widehat{R}_g$	Out. $\widehat{R}_g$
RMSE_3	0.1341	0.1330	0.1853	0.1851

TABLE B.4. Results comparing the root MSE of $\widehat{R}_g$ and $\widehat{\psi}$ using the regular estimation method without outlier detection and the new outlier detection method. Simulation results are based on a sample size of 1000 for the molecule myoglobin with both trend and single point outlying behavior.

	Reg. $\widehat{\psi}$	Out. $\widehat{\psi}$	Reg. $\widehat{R}_g$	Out. $\widehat{R}_g$
RMSE_3	1.0222	0.1307	1.6468	0.3652