Summer 2008

A Comparison of Bayesian Regression Models Applied in Knot Theory

Sevcan Bilir

A COMPARISON OF BAYESIAN REGRESSION MODELS APPLIED IN KNOT THEORY

A Thesis

Presented to the Faculty

of the Department of Mathematics and Computer Science

McAnulty College and Graduate School of Liberal Arts

Duquesne University

In partial fulfillment of the requirements for

the degree of Master of Science in Computational Mathematics

By

Sevcan Bilir

May 2008

A COMPARISON OF BAYESIAN REGRESSION MODELS APPLIED IN KNOT THEORY

By

Sevcan Bilir

Approved April 17, 2008

APPROVED_____

John Kern, Ph.D., Associate Professor

Department of Mathematics & Computer Science

APPROVED_____

Donald Simon, Ph.D., Associate Professor

Department of Mathematics & Computer Science

APPROVED_____

Mark Mazur, Ph.D., Graduate Director

Department of Mathematics & Computer Science

APPROVED_____

Albert C. Labriola, Ph.D., Dean

McAnulty College and Graduate School of Liberal Arts

ABSTRACT

A COMPARISON OF BAYESIAN REGRESSION MODELS APPLIED IN KNOT THEORY

By

Sevcan Bilir

May 2008

Thesis Supervised by John C. Kern II

This thesis explores variations on a Bayesian regression model used to estimate the mean box length of a random knot as a function of the number of edges of that knot. Specifically, this research recongizes uncertainty in box length variance and compares the resulting inference with that based on an approach that does not recognize such uncertainty. The Bayesian model is then shown to allow straightforward inference on the crossing location of two population regression lines.

ACKNOWLEDGEMENT

I would like to thank my advisor John C. Kern II for his ideas, efforts and support in this work. I would also like to thank Donald Simon and Patrick Juola for their help with computational issues, Mark Mazur and Donald Simon for reviewing the write-up and Eric Rawdon for supplying the study data.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Objective

This thesis is based on a recent study in knot theory [1] which explores the relationship between the number of edges (independent variable) and the box length (dependent variable) of a knot. In this study, a random polygon generation algorithm is used to generate random knots. From these simulated knots, the authors estimate parameters of a function that relates the mean box length to the number of edges of a knot. In this thesis, we use the data from [1] to implement a model that recognizes uncertainty in the variability of the box length values. The resulting inference is then compared with that from [1], where a model that does not recognize uncertainty in the variability of box length values was used.

We propose a normal distribution for box length values at each edge and treat the variability of box length at each edge as random. The primary benefit of treating the box length variability at a given number of edges as random is that the resulting uncertainty estimates for the mean box length function will be more realistic (i.e., greater). However, our analysis may provide results that differ negligibly from those

obtained from a fixed-variance approach, as the data has large enough size to be treated as a population (at least 73,000 knots were simulated at each edge). Improvements of random variance modeling over fixed variance modeling will be assessed after both have been implemented.

A further purpose of this thesis is to explicitly recognize the details behind the fixed-variance model implemented in [1], as such details are not present in their paper. In our study, we will explicitly state the parameter estimation technique for both the fixed and random variance models and then determine how parameter estimations are affected by including a random variance component. Credible intervals will be taken into account when comparing models.

1.2 The Data

In [1], a random knot was generated by using a polygon generation algorithm. Then each knot was measured for the following properties: box length, number of edges and knot-type. A knot is a closed loop with no self-intersection, usually in R^3. The number of edges is the number of segments used in generating a polygon. Box length is the maximum distance between any two vertices. Knot-type can be trefoil and non-trefoil, where trefoil knots are the simplest non-trivial knots which can be obtained by joining the loose ends of a overhand knot [2]. Figure 1.1 shows a trefoil knot with 50 edges and box length as the width of the enclosing prism.

In this research we use two different datasets. The first data set is composed of knots only classified as trefoils; we call this Dataset-I. The second dataset is composed of knots of all types, including trefoils and non-trefoils; we refer to this as Dataset-II. Knots in each dataset assume one of 46 different number of edges, ranging from 50

to 500 by a step size of 10. Table 1.1 shows the number of box length observations at each edge in Dataset-I. Note that Dataset-II has 400,000 observations at each edge.

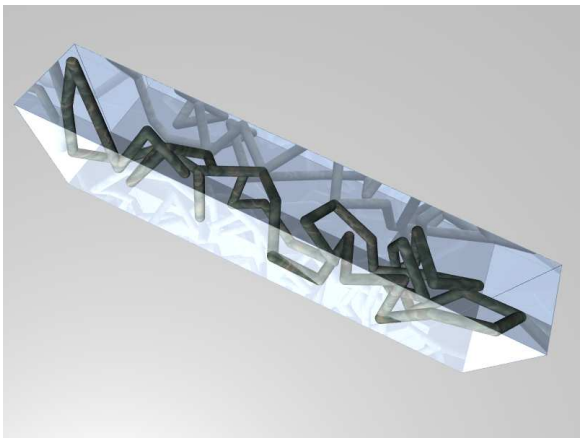

Figure 1.1: A trefoil knot with 50 edges.

1.3 The Model

As mentioned in the data section, the number of edges range from 50 to 500 by a step size of 10. We let the box length measurement of the i^{th} knot at edge j be denoted by x_{ij}. Histograms of box length values for 50, 270 and 500 edges are shown in Figure 1.2.

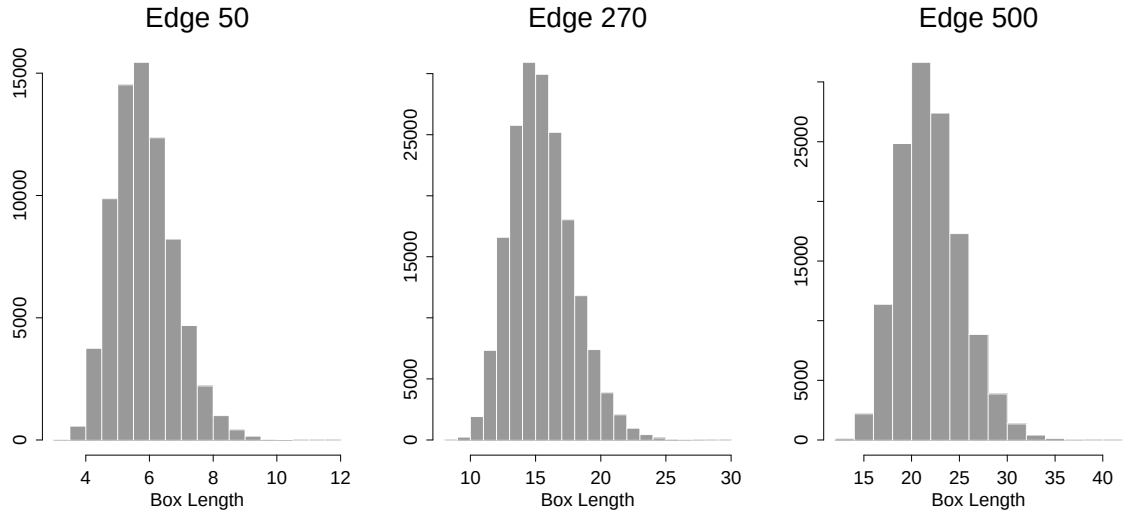

Figure 1.2: Histograms of box lengths at 50, 270 and 500 edges.

Table 1.1: Number of observations at each edge in Dataset-I

# of edges	# of observations	# of edges	# of observations
50	73274	60	89838
70	103034	80	114840
90	125694	100	135162
110	142922	120	150668
130	157022	140	163380
150	167492	160	171612
170	174600	180	177054
190	179752	200	182820
210	183246	220	184914
230	184212	240	184168
250	183436	260	183528
270	182820	280	182020
290	180654	300	178072
310	176682	320	174720
330	172846	340	172210
350	168576	360	167294
370	164906	380	162254
390	159090	400	157198
410	154722	420	152320
430	148726	440	146456
450	142930	460	140502
470	138094	480	135632
490	132844	500	129534

Based on these three histograms, we propose x_{ij} to be normal with mean μ_j and variance σ_j^2,

$$x_{ij} \sim N(\mu_j, \sigma_j^2) \quad \text{for} \quad j = \{50, 60, \ldots, 500\}.$$

As seen in [1], the authors mentioned that the mean value of box length at each edge should scale linearly with respect to number of edges, and the mean function is introduced as $\mu_j = Aj + B\sqrt{j} + C$. In Figure 1.3 the observed mean value of box length at each edge is shown with dots and the box length mean function is fitted to data with $A = 0.00385, B = 0.945$ and $C = -1.02$.

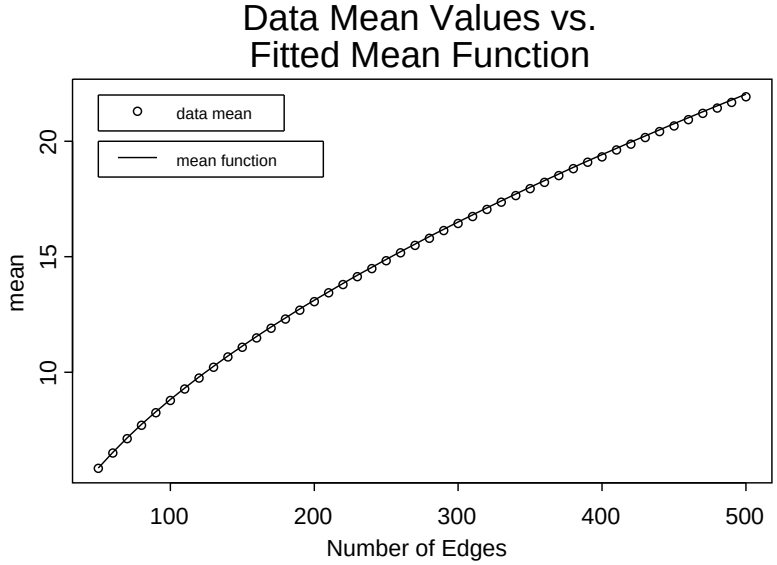

Figure 1.3: Solid dots represent observed mean of box length at each edge. The mean function is introduced as $\mu_j = Aj + B\sqrt{j} + C$ [1] and plotted above with $A = 0.00385, B = 0.945$ and $C = -1.02$.

The likelihood function for $\boldsymbol{\mu}$ and $\boldsymbol{\sigma^2}$ is the product of 46 joint normal distributions for each edge, where n_j denotes the number of observed box length values for edge j and $S = \{50, 60 \ldots 500\}$,

$$L(\boldsymbol{\mu}, \boldsymbol{\sigma^2}|\mathbf{x}) = \prod_{j \in S} \prod_{i=1}^{n_j} \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{1}{2\sigma_j^2}(x_{ij} - \mu_j)^2}. \tag{1.1}$$

We assign noninformative priors (Jeffrey's Prior) for $\boldsymbol{\mu}$ and $\boldsymbol{\sigma^2}$ [3],

$$\pi(\boldsymbol{\mu}) \propto 1 \quad \text{and} \quad \pi(\sigma_j^2) \propto \frac{1}{\sigma_j^2} \ \text{ for } \ j \in S. \tag{1.2}$$

By multiplying these prior distributions for $\boldsymbol{\mu}$ and $\boldsymbol{\sigma^2}$ with the likelihood function $L(\boldsymbol{\mu}, \boldsymbol{\sigma^2}|\mathbf{x})$, we obtain the joint posterior distribution for parameters A, B, C and $\boldsymbol{\sigma^2}$ of the box length mean function denoted $\pi(\boldsymbol{\mu}, \boldsymbol{\sigma^2}|\mathbf{x})$ and given by

$$\pi(\boldsymbol{\mu}, \boldsymbol{\sigma^2}|\mathbf{x}) \propto L(\boldsymbol{\mu}, \boldsymbol{\sigma^2}|\mathbf{x}) \cdot \pi(\boldsymbol{\mu}) \cdot \pi(\boldsymbol{\sigma^2}). \tag{1.3}$$

Given this posterior distribution for the box length function parameters, we use Markov Chain Monte Carlo (MCMC) sampling techniques to draw A, B, C and $\boldsymbol{\sigma^2}$ values from this distribution [4]. Inference is then made on these values.

Chapter 2

Model Implementation

2.1 Markov Chain Monte Carlo Sampling Techniques

Using the model described in Section 1.3, we employ MCMC sampling techniques to estimate the parameters A, B, C and $\boldsymbol{\sigma^2}$. From the joint posterior distribution, we are able to recognize the full conditional distribution for each of the parameters. Hence we use Gibbs sampling[1] (a member of MCMC sampling techniques) to generate a realization from the marginal distribution of each parameter, conditional on the current values of other parameters [6].

2.2 MCMC Calculations

In order to update the parameters using the Gibbs sampling technique, the full conditional distribution for all parameters must be obtained.

The joint posterior distribution is the product of the likelihood function (1.1) and prior distributions (1.2). The explicit form of our joint posterior distribution is

[1] As a general rule, in the case where the full conditional distribution is unobtainable for a parameter, Metropolis or Metropolis-Hastings sampling should be used [5].

therefore

$$\pi(\boldsymbol{\mu}, \boldsymbol{\sigma^2}|x_{ij}) \;\; \propto \;\; \prod_{j\in S}\left(\prod_{i=1}^{n_j}\frac{1}{\sqrt{\sigma_j^2}}e^{-\frac{1}{2\sigma_j^2}\left(x_{ij}-\mu_j\right)^2}\right)\frac{1}{\sigma_j^2}, \tag{2.1}$$

where $\mu_j = Aj + B\sqrt{j} + C$. For generalization of notation, redefine the mean function μ_j as $\mu_j = \sum_{k=1}^{M}\beta_k j^{P_k}$, $j \in S$. Here M represents the number of coefficients in the mean function. Thus for $M = 3, \beta_1 = A$, $P_1 = 1$, $\beta_2 = B$, $P_2 = \frac{1}{2}$, $\beta_3 = C$ and $P_3 = 0$, we have the specific mean function $\mu_j = Aj + B\sqrt{j} + C$.
Substituting $\mu_j = \sum_{k=1}^{M}\beta_k j^{P_k}$ into (2.1) and letting $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_M\}$ yields

$$\pi(\boldsymbol{\beta}, \boldsymbol{\sigma^2}|\mathbf{x}) \propto \prod_{j\in S}\left[\sigma_j^{2^{-}\left(\frac{n_j}{2}+1\right)}\prod_{i=1}^{n_j}e^{-\frac{1}{2\sigma_j^2}\left(x_{ij}-\sum_{k=1}^{M}\beta_k j^{P_k}\right)^2}\right]. \tag{2.2}$$

To more easily recognize the full conditional distributions of our parameters, we compute the log-posterior function. Taking the natural log of (2.2) yields[2]

$$ln(\pi(\boldsymbol{\beta}, \boldsymbol{\sigma^2}|\mathbf{x})) = -\sum_{j\in S}\left[\left(\frac{n_j}{2}+1\right)\ln(\sigma_j^2) + \frac{1}{2\sigma_j^2}\left[\sum_{i=1}^{n_j}\left(x_{ij}-\sum_{k=1}^{M}\beta_k j^{P_k}\right)^2\right]\right]. \tag{2.3}$$

We will present the full conditional distributions for the components of $\boldsymbol{\beta}$ and $\boldsymbol{\sigma^2}$ in the following subsections.

[2]The usage of equality actually incorporates an additive constant on the r.h.s. This slight abuse of equality occurs for all remaining natural logged functions, and does not affect parameter inference.

2.2.1 Updating β_m

In order to get the full conditional distribution of an element of $\boldsymbol{\beta}$, we start by rewriting (2.3) for $j \in S$:

$$\begin{aligned}
\ln(\pi(\boldsymbol{\beta}|\boldsymbol{\sigma^2}, \mathbf{x})) = &-\left(\frac{n_{50}}{2}+1\right)\ln(\sigma^2_{50}) - \frac{1}{2\sigma^2_{50}}\left[\sum_{i=1}^{n_{50}}\left(x_{i50} - \sum_{k=1}^{M}\beta_k 50^{P_k}\right)^2\right] \\
&-\left(\frac{n_{60}}{2}+1\right)\ln(\sigma^2_{60}) - \frac{1}{2\sigma^2_{60}}\left[\sum_{i=1}^{n_{60}}\left(x_{i60} - \sum_{k=1}^{M}\beta_k 60^{P_k}\right)^2\right] \\
&\vdots \\
&-\left(\frac{n_{500}}{2}+1\right)\ln(\sigma^2_{500}) - \frac{1}{2\sigma^2_{500}}\left[\sum_{i=1}^{n_{500}}\left(x_{i500} - \sum_{k=1}^{M}\beta_k 500^{P_k}\right)^2\right]. \quad (2.4)
\end{aligned}$$

In (2.4), we group the terms which do not have β_m's and call them Z_1 and expand the squared terms which include β_m's.

$$\ln(\pi(\boldsymbol{\beta}|\boldsymbol{\sigma^2}, \mathbf{x})) = Z_1 - \sum_{j\in S}\left[\frac{1}{2\sigma^2_j}\sum_{i=1}^{n_j}\left(x^2_{ij} - 2x_{ij}\left(\sum_{k=1}^{M}\beta_k j^{P_k}\right) + \left(\sum_{k=1}^{M}\beta_k j^{P_k}\right)^2\right)\right] \quad (2.5)$$

Now we can rewrite the squared terms in (2.5) as

$$\begin{aligned}
\left(\sum_{k=1}^{M}\beta_k j^{P_k}\right)^2 &= \left(\beta_1 j^{P_1} + \beta_2 j^{P_2} + \ldots + \beta_M j^{P_M}\right)\left(\beta_1 j^{P_1} + \beta_2 j^{P_2} + \ldots + \beta_M j^{P_M}\right) \\
&= K + \beta^2_m j^{2P_m} \\
&\quad + 2\beta_m j^{P_m}\left(\beta_1 j^{P_1} + \ldots + \beta_{m-1} j^{P_{m-1}} + \beta_{m+1} j^{P_{m+1}} + \ldots + \beta_M j^{P_M}\right). \quad (2.6)
\end{aligned}$$

where K represents all the other terms which do not depend on β_m. Hence (2.5) becomes

$$\begin{aligned}\ln(\pi(\boldsymbol{\beta}|\boldsymbol{\sigma^2},\mathbf{x})) &= Z_1 - \sum_{j\in S}\left(\frac{1}{2\sigma_j^2}\sum_{i=1}^{n_j}\left(x_{ij}^2 - 2x_{ij}\sum_{k=1}^{M}\beta_k j^{P_k}\right.\right.\\ &\quad \left.\left. + \beta_m^2 j^{2P_m} + 2\beta_m j^{P_m}\sum_{k\neq m}^{M}\beta_k j^{P_k} + K\right)\right). \qquad (2.7)\end{aligned}$$

In (2.7), we will group all 46 of the $-\frac{1}{2\sigma_j^2}\sum_{i=1}^{n_j}x_{ij}^2$ terms that do not depend on any $\boldsymbol{\beta}$ variables, and add them to Z_1. Call this new sum Z_2. Furthermore, we should name the (irrelevant) constant part of squared term as R_1 in order to complete (2.7) to a perfect square for β_m. Define $\boldsymbol{\beta_{-m}} = \{\beta_1,\ldots,\beta_{m-1},\beta_{m+1},\ldots,\beta_M\}$. Then

$$\ln(\pi(\beta_m|\boldsymbol{\beta_{-m}},\boldsymbol{\sigma^2},\mathbf{x})) \propto Z_2 - \sum_{j\in S}\frac{1}{2\sigma_j^2}\sum_{i=1}^{n_j} j^{2P_m}\left(\beta_m^2 - 2\beta_m j^{-P_m}\left(x_{ij} + \sum_{k\neq m}^{M}\beta_k j^{P_k}\right) + R_1\right). \qquad (2.8)$$

Completing the square in β_m shows $\pi(\beta_m|\boldsymbol{\beta_{-m}},\boldsymbol{\sigma^2},\mathbf{x})$ is a Normal density :

$$\beta_m|\boldsymbol{\beta_{-m}},\boldsymbol{\sigma^2},\mathbf{x} \sim N\left(\frac{\sum_{j\in S}\frac{j^{P_m}}{\sigma_j^2}\left(\sum_{i=1}^{n_j}x_{ij} - \left(\sum_{k\neq m}^{M}\beta_k j^{P_k}\right)\right)}{\sum_{j\in S}\frac{n_j j^{2P_m}}{\sigma_j^2}}, \left(\sum_{j\in S}\frac{n_j j^{2P_m}}{\sigma_j^2}\right)^{-1}\right). \qquad (2.9)$$

2.2.2 Updating A, B and C

We have shown the full conditional distribution for general coefficient β_m. Setting $\beta_1 = A$, $P_1 = 1$, $\beta_2 = B$, $P_2 = \frac{1}{2}$, $\beta_3 = C$ and $P_3 = 0$ gives the full conditional

distributions for parameters A, B, C of the mean function $\mu_j = Aj + B\sqrt{j} + C$.

It is straighforward to see, then, that the full conditional distribution for A, or $\pi(A|B, C, \sigma^{\mathbf{2}}, \mathbf{x})$ is a normal density:

$$A|B, C, \boldsymbol{\sigma^2}, \mathbf{x} \sim N\left(\frac{\sum_{j\in S}\frac{j}{2\sigma_j^2}\sum_{i=1}^{n_j}\left(x_{ij} - Bj^{1/2} - C\right)}{\sum_{j\in S}\frac{j^2 n_j}{2\sigma_j^2}}, \left(\sum_{j\in S}\frac{j^2 n_j}{2\sigma_j^2}\right)^{-1}\right). \quad (2.10)$$

The full conditional distribution $\pi(B|A, C, \sigma^{\mathbf{2}}, \mathbf{x})$ for B also is a normal density:

$$B|A, C, \boldsymbol{\sigma^2}, \mathbf{x} \sim N\left(\frac{\sum_{j\in S}\frac{j^{\frac{1}{2}}}{2\sigma_j^2}\sum_{i=1}^{n_j}\left(x_{ij} - Aj - C\right)}{\sum_{j\in S}\frac{jn_j}{2\sigma_j^2}}, \left(\sum_{j\in S}\frac{jn_j}{2\sigma_j^2}\right)^{-1}\right). \quad (2.11)$$

The full conditional distribution $\pi(C|A, B, \boldsymbol{\sigma^2}, \mathbf{x})$ for C follows in analogous fashion:

$$C|A, B, \boldsymbol{\sigma^2}, \mathbf{x} \sim N\left(\frac{\sum_{j\in S}\frac{1}{2\sigma_j^2}\sum_{i=1}^{n_j}\left(x_{ij} - Bj^{\frac{1}{2}} - Aj\right)}{\sum_{j\in S}\frac{n_j}{2\sigma_j^2}}, \left(\sum_{j\in S}\frac{n_j}{2\sigma_j^2}\right)^{-1}\right). \quad (2.12)$$

2.2.3 Updating σ_j^2

In order to obtain the full conditional for σ_j^2, we rewrite (2.1),

$$\pi(\boldsymbol{\sigma^2}|\boldsymbol{\mu}, \mathbf{x}) \propto \prod_{j \in S} \left(\prod_{i=1}^{n_j} \frac{1}{\sqrt{\sigma_j^2}} e^{-\frac{1}{2\sigma_j^2}(x_{ij} - \mu_j)^2} \right) \frac{1}{\sigma_j^2}$$
$$= \prod_{j \in S} \sigma_j^{2^{-\left(\frac{n_j}{2} + 1\right)}} \prod_{i=1}^{n_j} e^{-\frac{1}{2\sigma_j^2}(x_{ij} - \mu_j)^2}. \tag{2.13}$$

Absorb all multiplicative terms that do not contain a σ_j^2 into the proportionality constant to obtain

$$\pi(\sigma_j^2|\boldsymbol{\mu}, \mathbf{x}) \propto \sigma_j^{2^{-\left(\frac{n_j}{2} + 1\right)}} e^{-\frac{1}{2\sigma_j^2} \sum_{i=1}^{n_j} (x_{ij} - \mu_j)^2}. \tag{2.14}$$

Recall the probability density function for the inverse gamma distribution, defined in [7], as:

$$f(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{\left(\frac{-\beta}{x}\right)} \text{ with } x > 0. \tag{2.15}$$

Based on (2.14) and (2.15), we recognize that the full conditional distribution for σ_j^2 has an inverse gamma distribution with $\alpha = \frac{n_j}{2}$ and $\beta = \sum_{i=1}^{n_j} \frac{(x_{ij} - \mu_j)^2}{2}$,

$$\sigma_j^2|\boldsymbol{\mu}, \mathbf{x} \sim InvGamma\left(\frac{n_j}{2}, \sum_{i=1}^{n_j} \frac{(x_{ij} - \mu_j)^2}{2}\right). \tag{2.16}$$

2.3 Gibbs Sampling Algorithms

Having established the full conditional distribution for all parameters, we present the fixed variance and random variance approaches for making inference about $\boldsymbol{\mu}$. The fixed variance (FV) methodology, which is used in [1], is given explicitly by the following steps:

1. For each j, set σ_j^2 to the empirical variance of the x_{ij}s.

2. Let $\boldsymbol{\sigma^2}$ represent the vector of empirical variances.

3. Select initial values for A_0, B_0 and C_0.

4. Let A_i, B_i, C_i represent the current values of these parameters.

5. Sample A_{i+1} from $\pi(A_{i+1}|B_i, C_i, \boldsymbol{\sigma^2}, \boldsymbol{x})$ in (2.10).

6. Sample B_{i+1} from $\pi(B_{i+1}|A_{i+1}, C_i, \boldsymbol{\sigma^2}, \boldsymbol{x})$ in (2.11).

7. Sample C_{i+1} from $\pi(C_{i+1}|A_{i+1}, B_{i+1}, \boldsymbol{\sigma^2}, \boldsymbol{x})$ in (2.12).

Steps 4 through 7 constitute one iteration, and yield a single realization $\{A_{i+1}$, B_{i+1}, $C_{i+1}\}$. Note that the σ_j^2 values are never updated in the FV approach.

It is in the random variance (RV) methodology where we recognize uncertainty in $\boldsymbol{\sigma^2}$. The RV approach is given explicitly in the following steps:

1. Set σ_j^2 to the empirical variance of the x_{ij}s for $j \in S$.

2. Let $\boldsymbol{\sigma^2}_0$ represent the vector of empirical variances.

3. Select initial values for A_0, B_0 and C_0.

4. Let A_i, B_i, C_i, $\boldsymbol{\sigma^2}_i$ represent the current values of these parameters.

5. Sample A_{i+1} from $\pi(A_{i+1}|B_i, C_i, \boldsymbol{\sigma^2}_i, \boldsymbol{x})$ in (2.10).

6. Sample B_{i+1} from $\pi(B_{i+1}|A_{i+1}, C_i, \boldsymbol{\sigma^2}_i, \boldsymbol{x})$ in (2.11).

7. Sample C_{i+1} from $\pi(C_{i+1}|A_{i+1}, B_{i+1}, \boldsymbol{\sigma^2}_i, \boldsymbol{x})$ in (2.12).

8. For each j, sample σ_j^2 from $\pi(\sigma_j^2|A_{i+1}, B_{i+1}, C_{i+1}, \boldsymbol{x})$ in (2.16) and store the vector of sampled σ_j^2s in $\boldsymbol{\sigma^2}_{i+1}$.

Steps 4 through 8 constitute one iteration, and yield a single realization $\{A_{i+1}$, B_{i+1}, C_{i+1},$\boldsymbol{\sigma^2}_{i+1}\}$.

We implemented a FV java program with 2 million iterations for Dataset-I and 1,000,000 iterations for Dataset-II. A RV java program was also implemented with the same number of iterations. In each approach and dataset, we have lagged the iterations by 1000 to avoid autocorrelations among samples of the same parameter.

Chapter 3

Discussion

3.1 Results

We have applied both the FV and RV models to each dataset. We present first the inference obtained from these two models using Dataset-I and then the inference obtained from these two models using Dataset-II.

3.1.1 Bayesian Analysis using Dataset-I

Fixed Variance Approach

In this research we show the correctness of the Bayesian analysis implementation by presenting convergence of parameters and absence of autocorrelation within iterations. Figure 3.1 shows that generated parameter values of A, B and C converge and Figure 3.2 shows that there is no autocorrelation between randomly generated values of A, B and C.

Since each realization $\{A_i, B_i, C_i\}$ defines a curve μ_j, thousands of realizations define thousands of curves. Figure 3.3 (left) shows the posterior mean of 2000 generated μ_j curves fitted to actual mean values of Dataset-I at each edge. The posterior

mean lies in the range of [data mean-0.0228, data mean+0.0246]. Since the difference between posterior mean and data mean cannot be distinguished easily in Figure 3.3 (left), we present the difference in Figure 3.3 (right) with a bar graph.

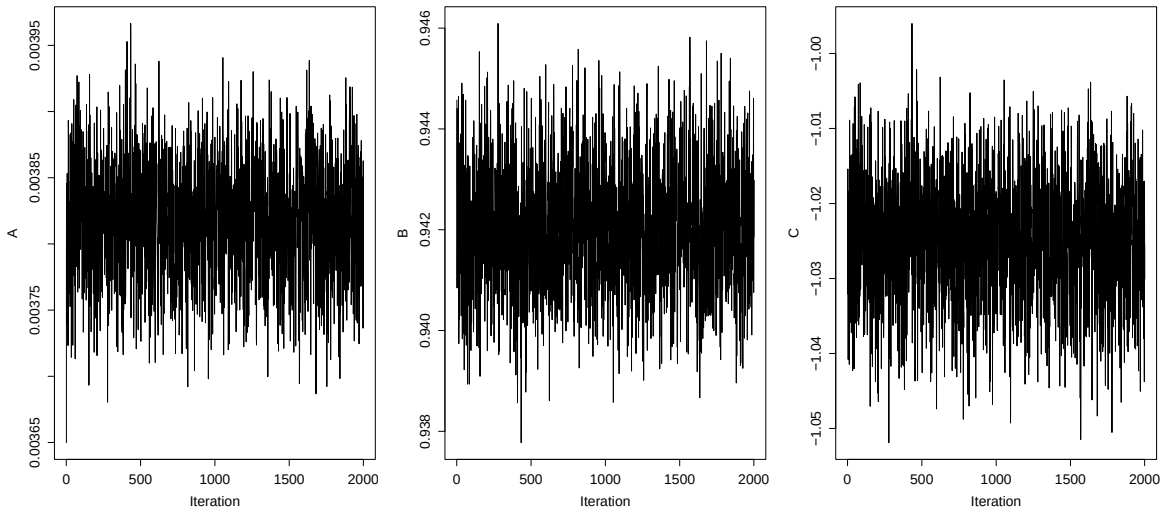

Figure 3.1: Trace plots for A, B and C under the FV model

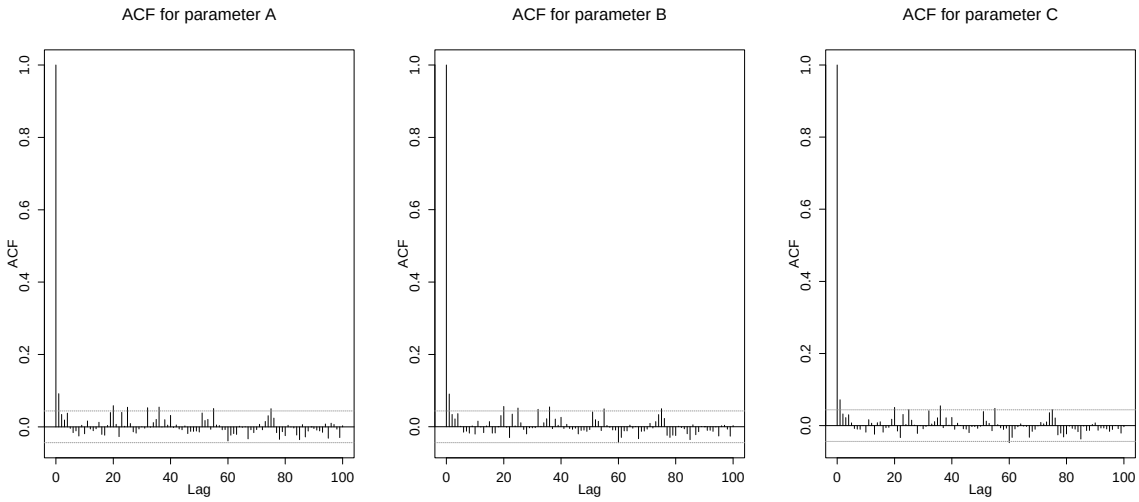

Figure 3.2: Autocorrelation graphs for A, B and C under the FV model

We show a 95% credible interval (CI) for the true function relating mean box length to number of edges by plotting the mean, 97.5% upper bound and 2.5% lower bound for all 2000 mean curve realizations. However, because of the small CI scale,

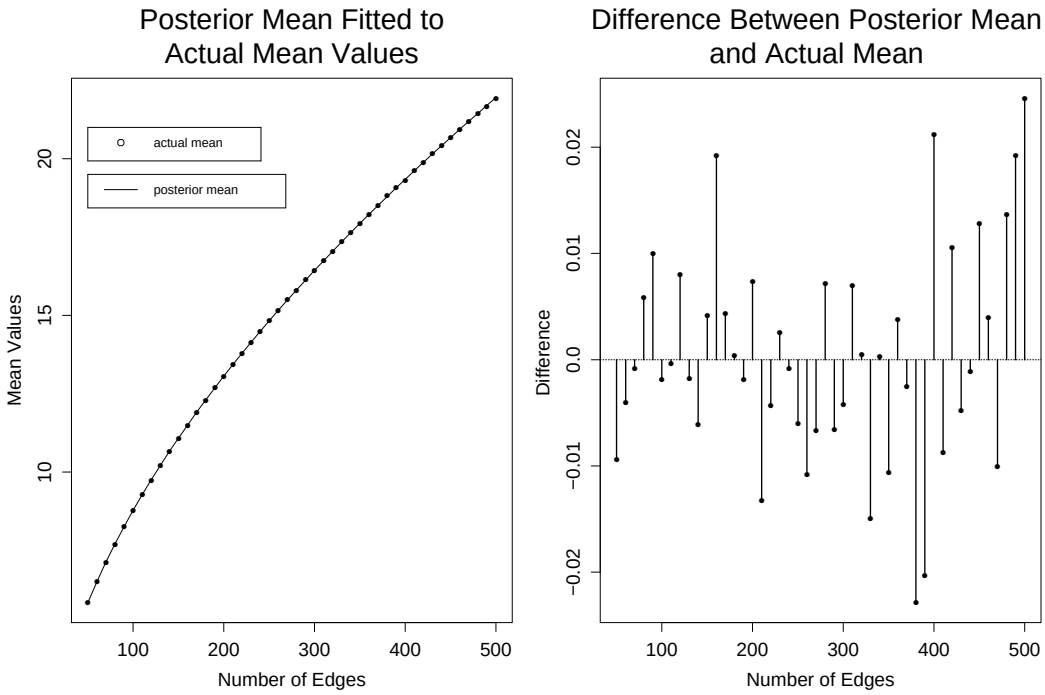

Figure 3.3: Left: Posterior mean is fitted to data mean at each edge. Right: Difference between posterior mean and data mean is presented as bar plot.

the mean and bound curves are difficult to distinguish. We therefore present a 95% CI by fixing the posterior mean at 0 and plotting the 97.5% upper bound and 2.5% lower bound relative to zero (Figure 3.4 (right)). Notice that the CI gets much larger from 300 to 500; this is due to the increased variance associated with more edges (Figure 3.4 (left)).

Random Variance Approach

In this section, we will present our results for the model with random variance of box length at each edge. In this second model, in addition to updating A, B and C values, we use the inverse gamma distribution to update σ_i^2 values (rather than keeping them fixed). Examination of trace plots and autocorrelation plots of variance values at each edge reveal all generated variance values converge with no autocorrelation between generated variance values at each edge. Traceplots of σ_{50}^2 and σ_{500}^2 are given in Figure 3.5 and the corresponding autocorrelation plots are given in Figure 3.6. Figure 3.7 and

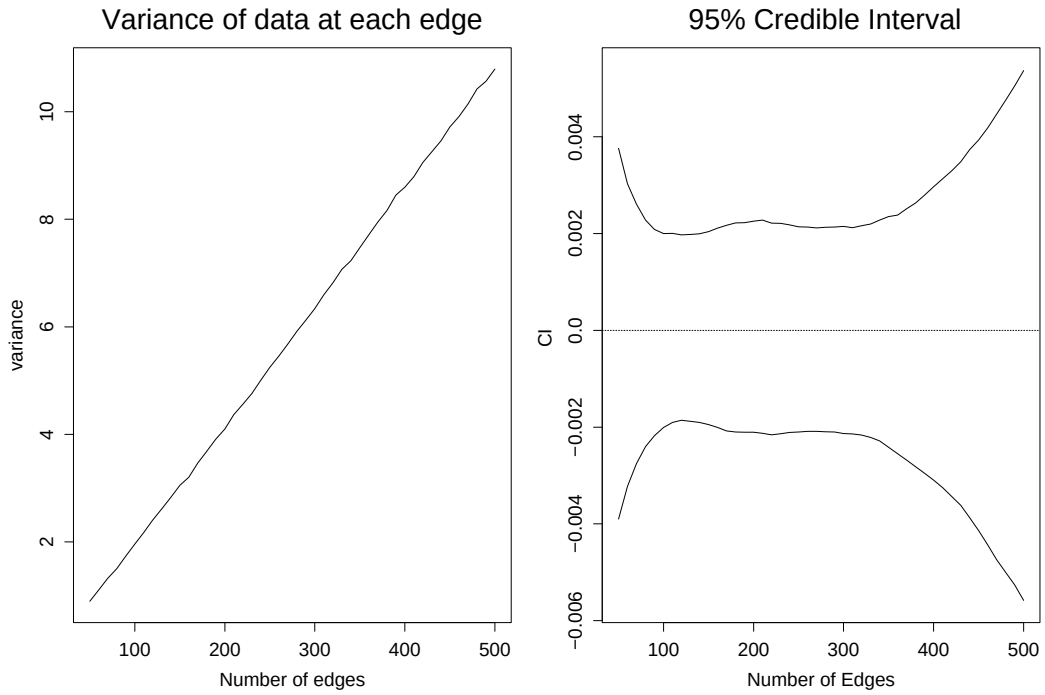

Figure 3.4: Left: Variance of Dataset-I at each edge. Right: 95% Credible interval at each edge.

Figure 3.8 show generated A, B and C values have converged and are independent, respectively.

Figure 3.9 (left) shows the posterior mean curve fitted to data means. In Figure 3.9 (middle), we see that the posterior mean lies in the range of [data mean-0.0228, data mean+0.025], this range is comparable to that with FV approach. Figure 3.9 (right) shows the difference between posterior mean of FV approach and posterior mean of RV approach is negligible ($< 6e^{-05}$) which is an expected result.

Because of the different variance approaches, credible intervals for the FV approach and RV approach should differ. We compare the 95% CI for the two approaches in Figure 3.10 (left). In order to better present the difference between 95% CIs, we plot the difference between the 2.5% lower bound in Figure 3.10 (middle) and the difference between the 97.5% upper bound in Figure 3.10 (right) by subtracting RV bound values from FV bound values. The difference between upper (or lower) bound is of a 10^{-4} order of magnitude.

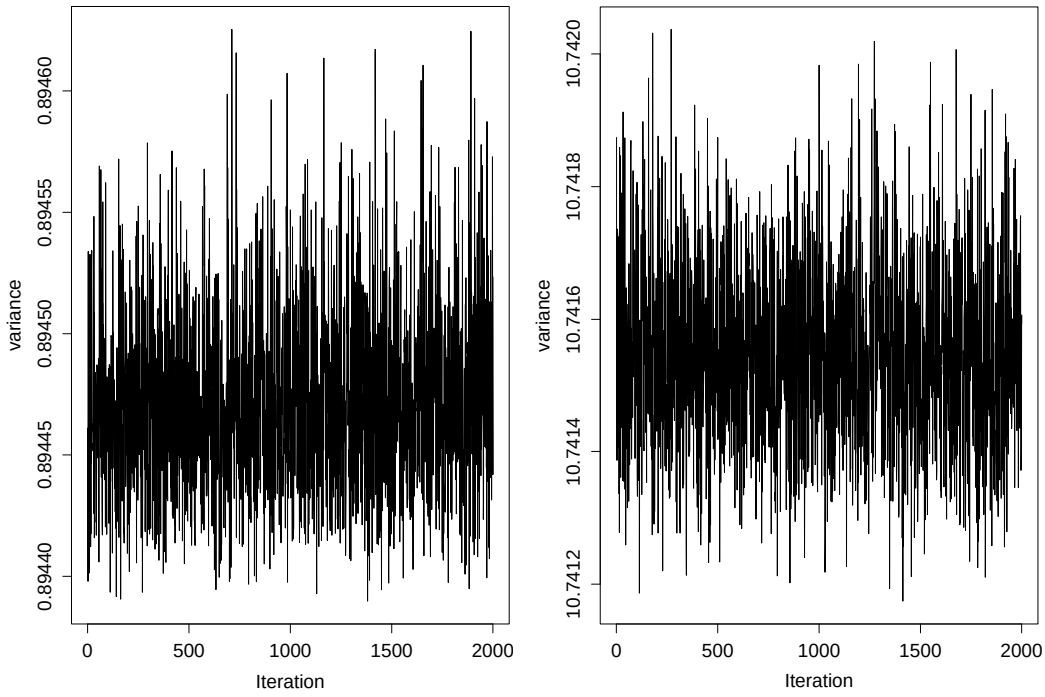

Figure 3.5: Trace plots for σ^2_{50} and σ^2_{500}

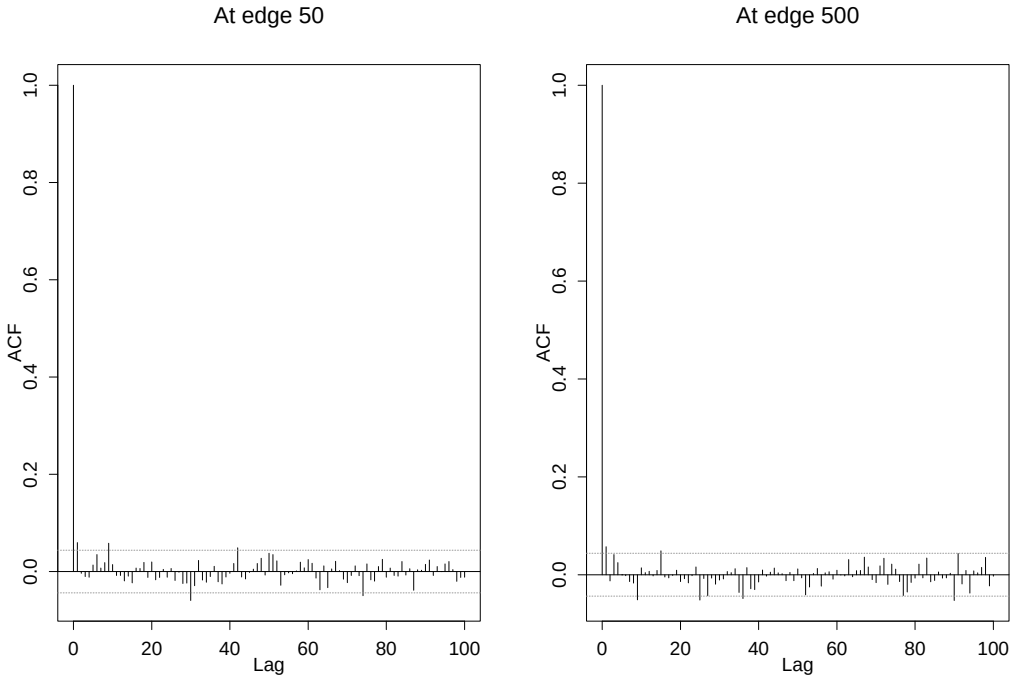

Figure 3.6: Autocorrelation graphs for σ^2_{50} and σ^2_{500}

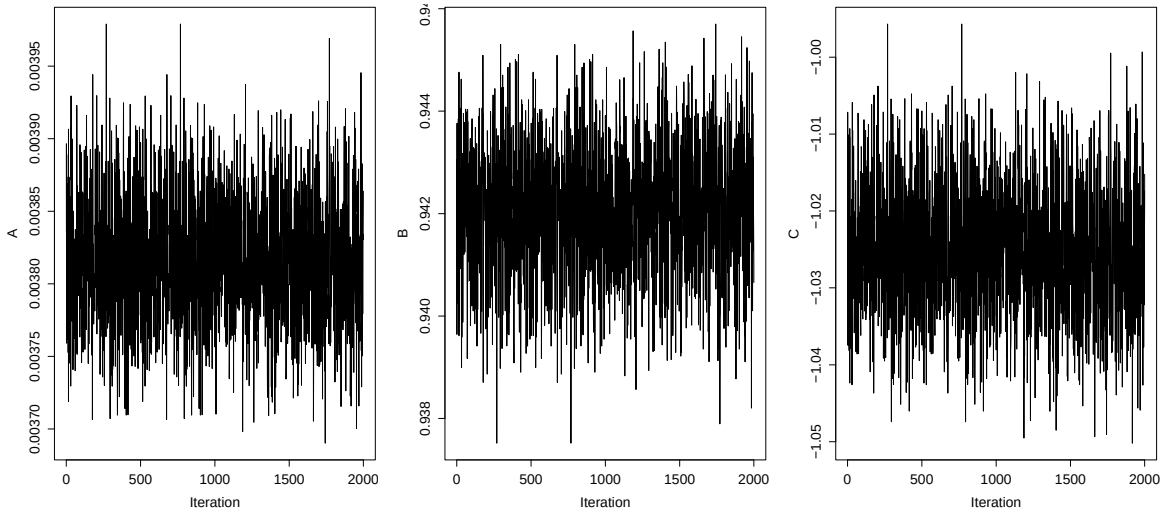

Figure 3.7: Trace plots for A, B and C under the RV model

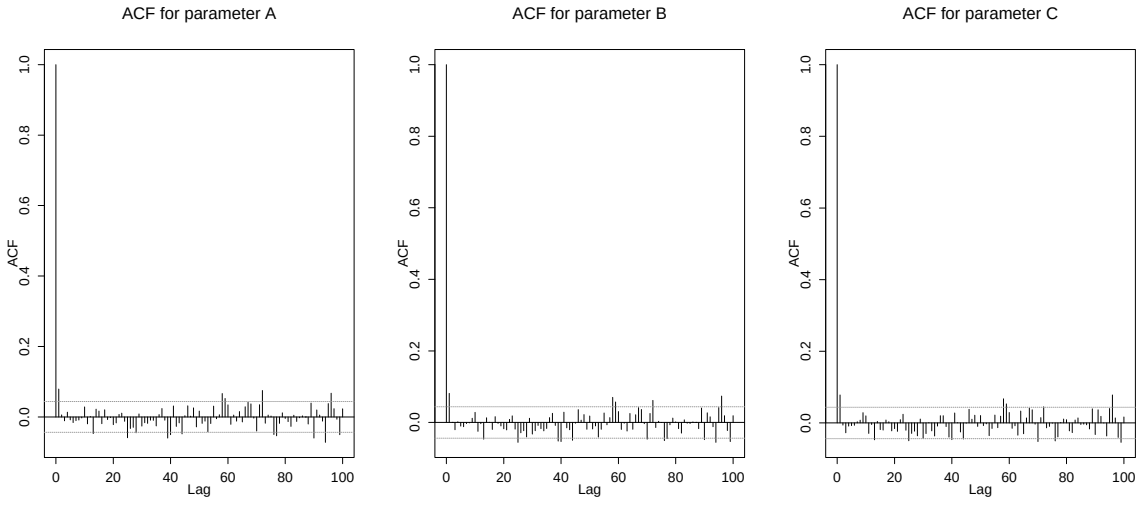

Figure 3.8: Autocorrelation graphs for A, B and C under the RV model

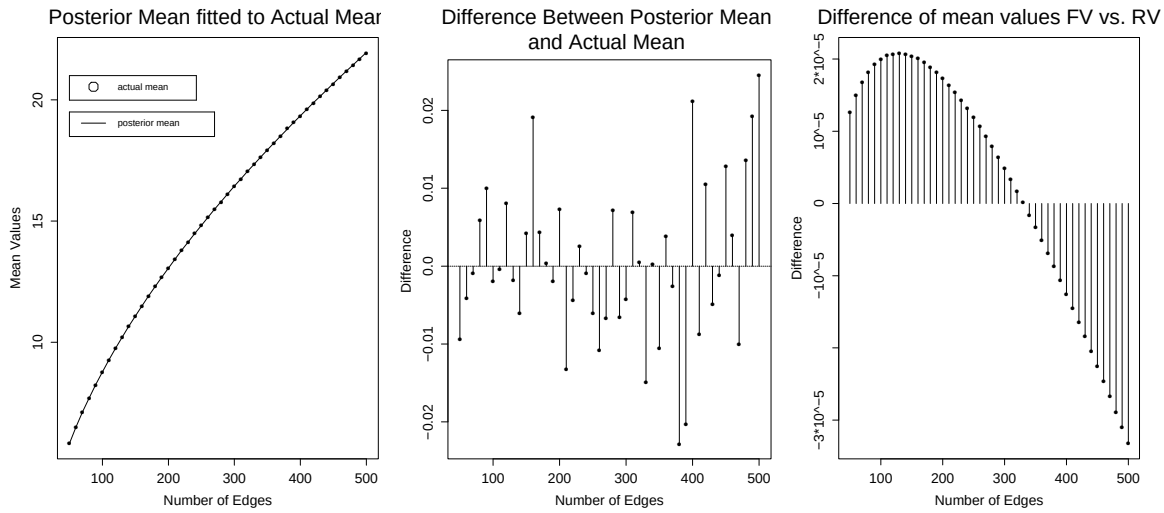

Figure 3.9: Left: Posterior mean is fitted to data mean at each edge. Middle: Difference between posterior mean and data mean is presented as bar plot. Right: Difference between posterior mean of FV approach and RV approach.

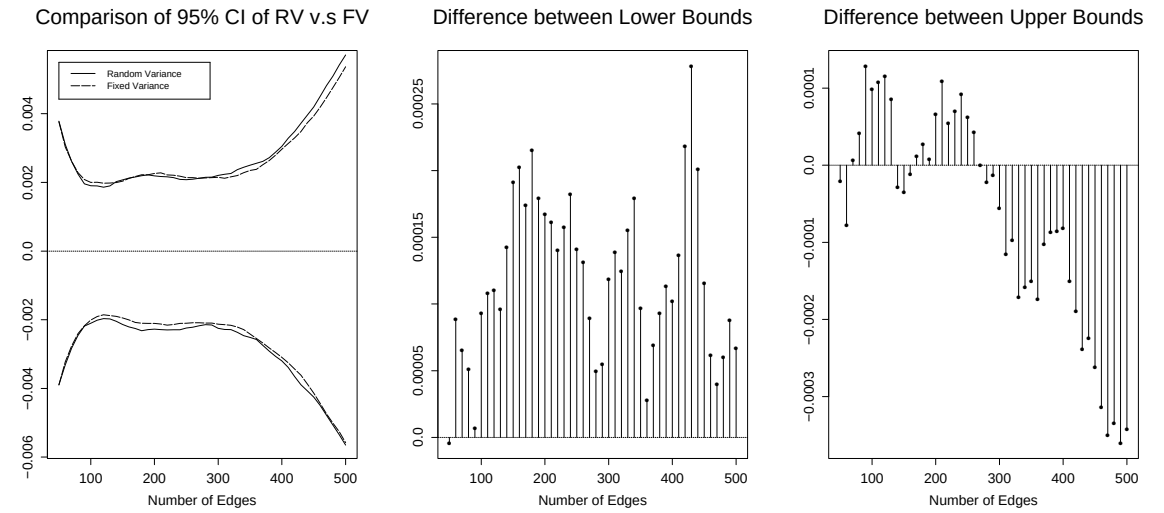

Figure 3.10: Left: 95% CI comparison of FV model and RV model. Middle: Difference between 2.5% lower bound of FV model and RV model. Right: Difference between 97.5% upper bound of FV model and RV model.

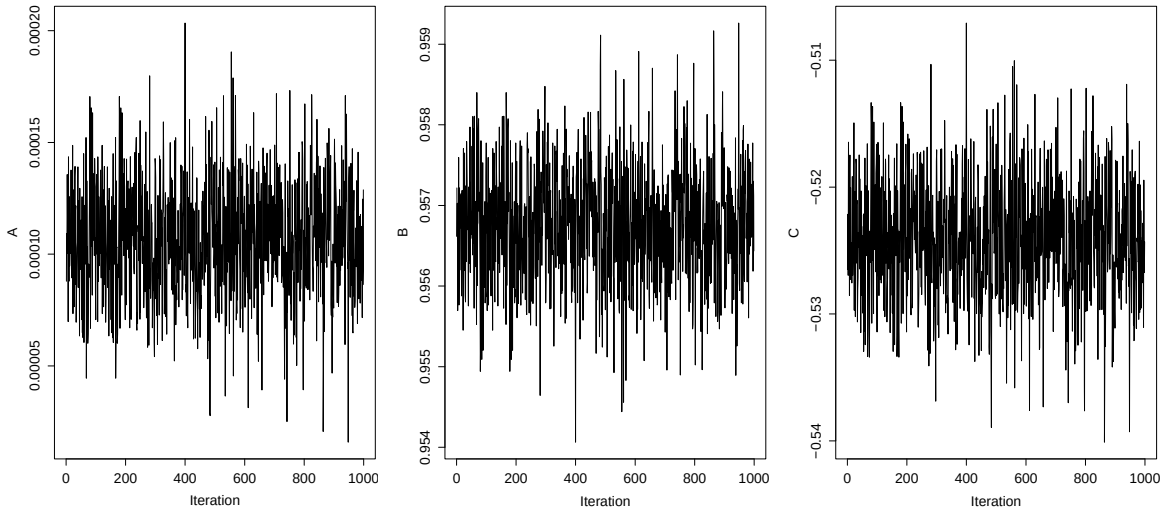

Figure 3.11: Trace plots for A, B and C

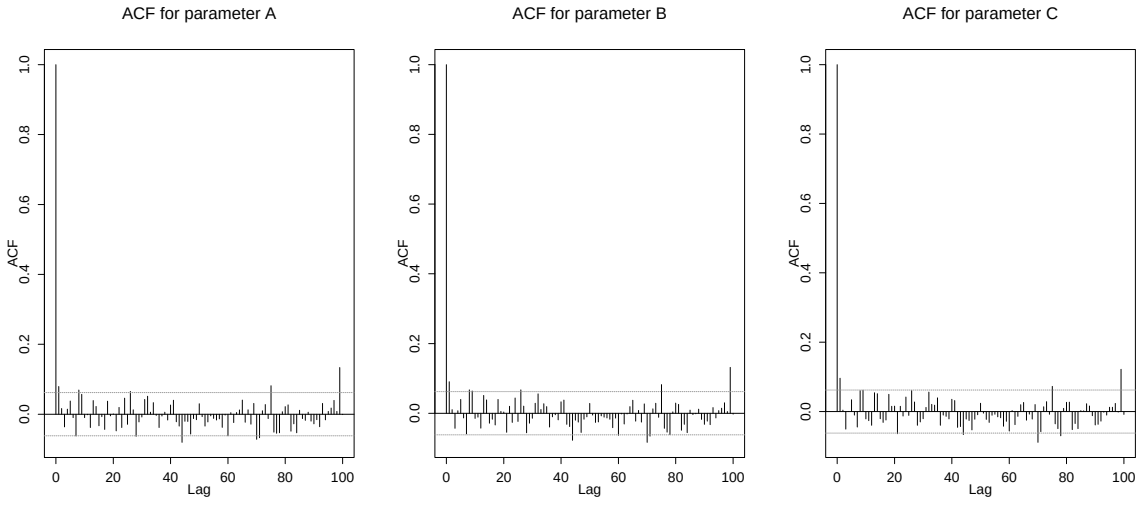

Figure 3.12: Autocorrelation graphs for A, B and C.

3.1.2 Bayesian Analysis using Dataset-II

Fixed Variance Approach

In this section, we complete 1000 iterations with using Dataset-II (400,000 observations at each edge). Figure 3.11 shows that generated A, B and C values converge and Figure 3.12 shows that there is no correlation between generated A, B and C values.

Figure 3.13 (left) shows the pointwise posterior mean of 1000 generated mean curves of box length fitted to actual mean values of Dataset-II at each edge. In Figure 3.13 (right), we see that the posterior mean lies in the range of [data mean-0.009, data mean+0.0071]. Data variance and 95% CI of the posterior mean are shown in Figure 3.14.

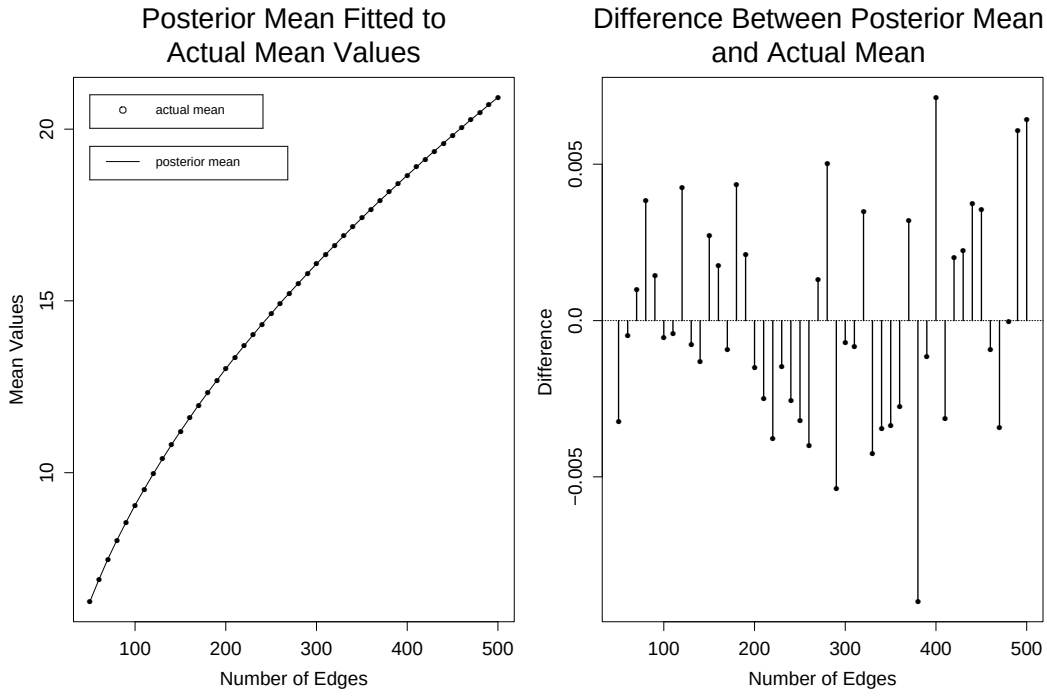

Figure 3.13: Difference between posterior mean and data mean. Left: Posterior mean is fitted to data mean at each edge. Right: Difference between posterior mean and data mean is presented as bar plot.

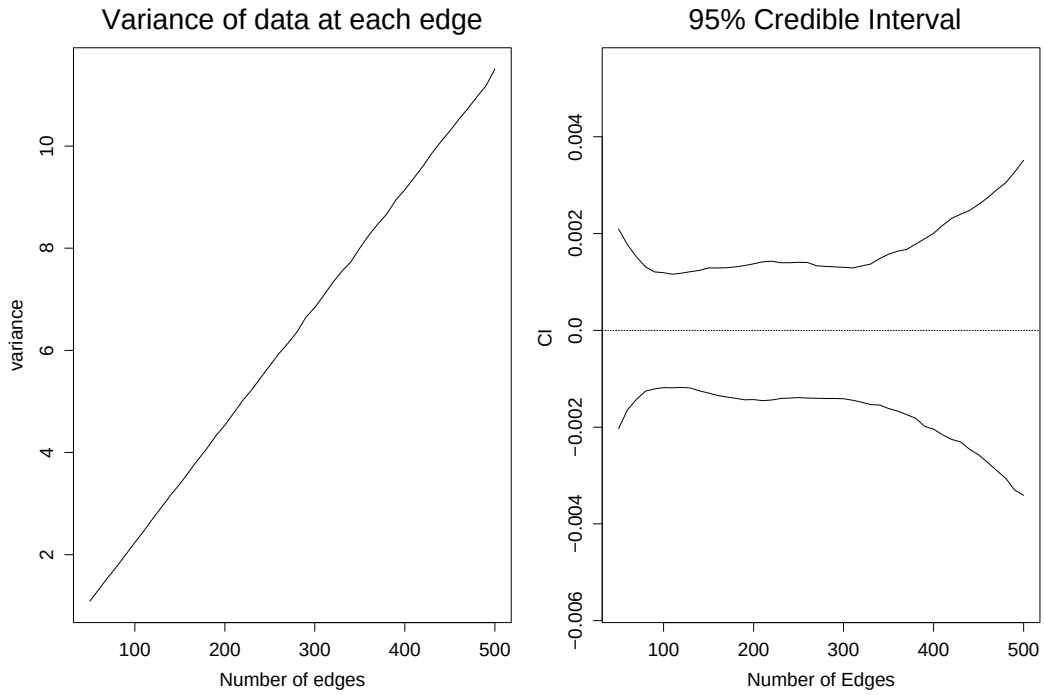

Figure 3.14: Left: Variance of Dataset-II at each edge. Right: 95% Credible interval at each edge.

Random Variance Approach

We now present our results using the RV approach with 1000 iterations of Dataset-II. Figure 3.15 shows that all generated A, B and C values converge and Figure 3.16 shows that there is no autocorrelation between generated A, B and C values.

Figure 3.17 (left) shows the posterior mean is fitted to the mean of Dataset-II. Figure 3.17 (middle) shows that the posterior mean lies in the range of [data mean-0.0066, data mean+0.0116]. Figure 3.17 (right) shows the differences between the posterior mean of the FV approach and the posterior mean of the RV approach.

We compare the 95% CIs of the two different approaches in Figure 3.18 (left). Figure 3.18 (middle) and Figure 3.18 (right) show the difference between the 2.5% lower bounds and the difference between the 97.5% upper bounds, respectively. In Section 3.1.1, for Dataset-I the RV approach mostly provides a larger 95% CI at each edge which is a result of added uncertainty in the variance. For Dataset-II, we could

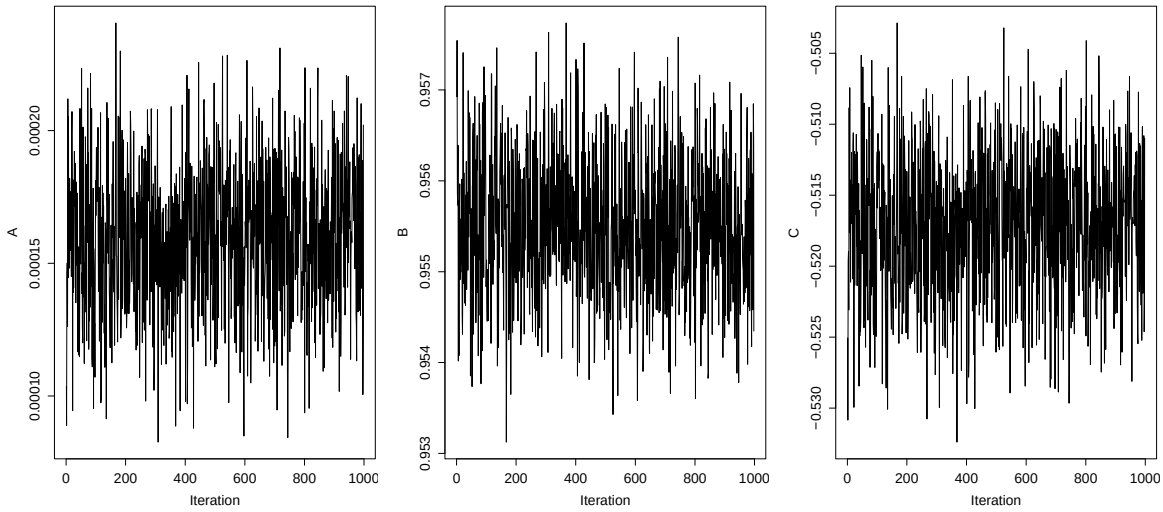

Figure 3.15: Trace plots for A, B and C (RV model).

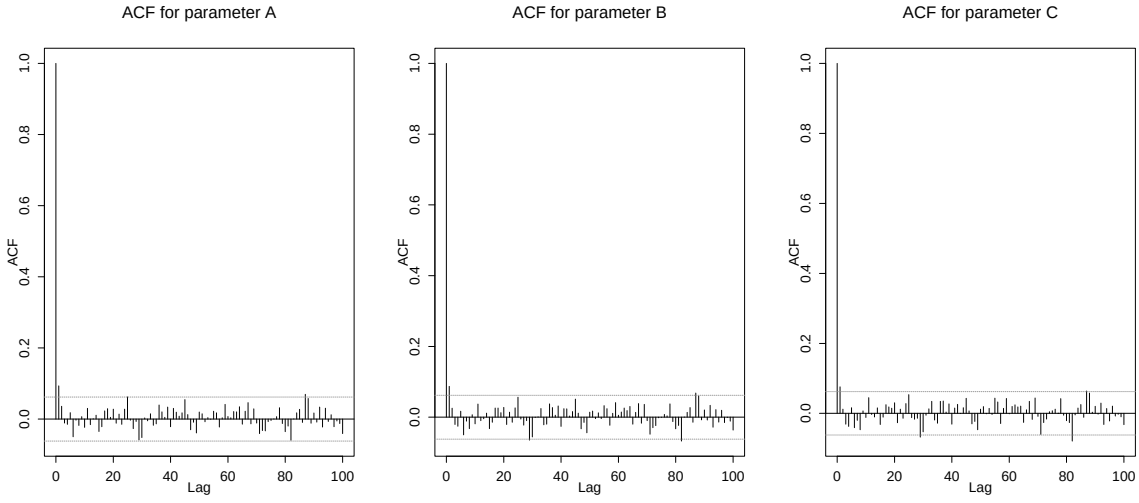

Figure 3.16: Autocorrelation graphs for A, B and C (RV model).

not see the same effect of using RV approach on 95% CI; the RV approach does not necessarily provide a larger CI. The reason is due to the larger size of the dataset which results in less uncertainty in variance.

3.1.3 Intersection of Mean Curves: Dataset-I vs. Dataset-II

A goal of the aforementioned study [1] was explaining the intersection locations of mean function curves for different datasets because of the fact that an intersection

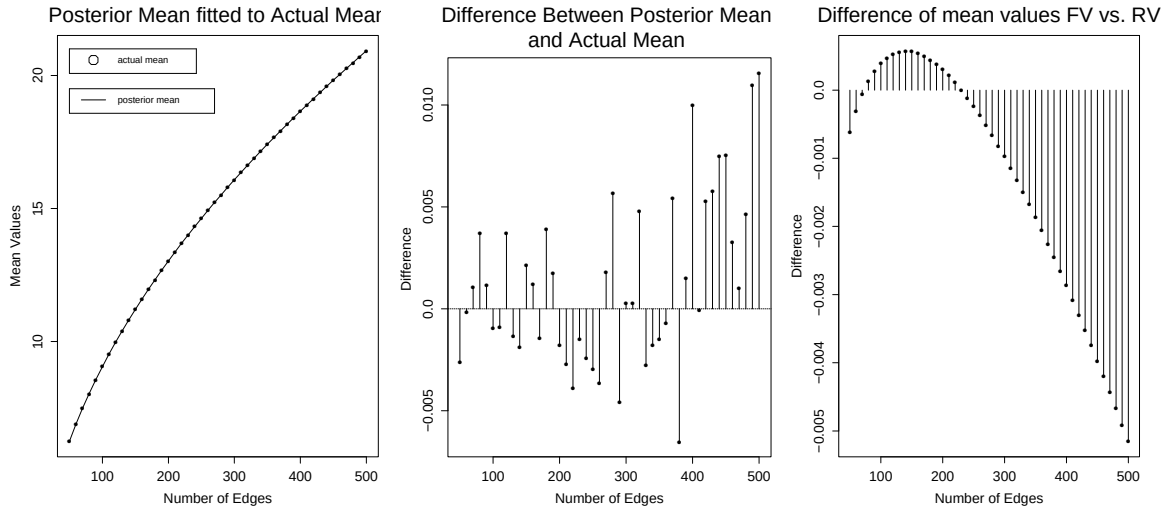

Figure 3.17: Left: Posterior mean is fitted to data mean at each edge. Middle: Difference between posterior mean and data mean is presented as bar plot. Right: Difference between posterior mean of FV approach and RV approach.

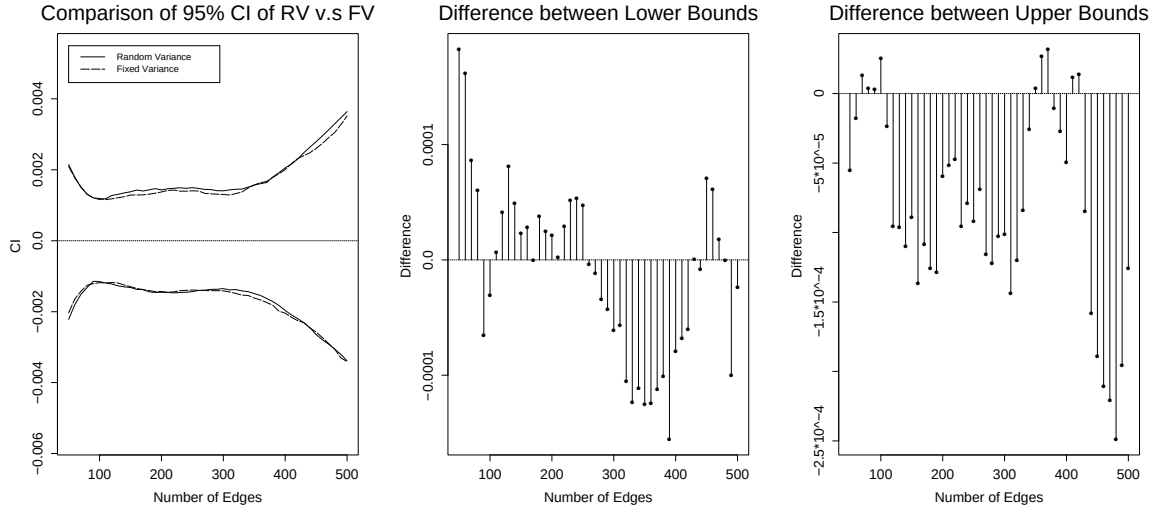

Figure 3.18: Left: 95% CI comparison of FV model and RV model. Middle: Difference between 2.5% lower bound of FV model and RV model. Right: Difference between 97.5% upper bound of FV model and RV model.

point (edge number) defines a transition between rigid and flexible knots. In this research we use Bayesian regression rather than classical regression in order to estimate parameters of mean curve functions. Using classical regression, we can estimate confidence intervals for mean curve functions of Dataset-I and Dataset-II. However, it is difficult to estimate a confidence interval for intersection location of these two

mean function curves. Using Bayesian regression, however, we sample 1000 regressed mean function curves for Dataset-I and Dataset-II, and find the intersection locations of these 1000 mean function curve pairs and get a 95% credible interval for the intersection location of the two mean function curves. Note that credible interval in Bayesian regression is analogous to confidence interval in classical regression [6].

In this section we will present Bayesian inference on the intersection of mean curves for Dataset-I and Dataset-II for the FV approach and the RV approach.

Intersection Inference Under FV Model

We graph the intersection of mean function curves for the first iteration of the FV approach in Figure 3.19 (left). For this first iteration, the curves intersect at edge 190.5429. Figure 3.19 (right) shows the intersection of mean curves at each edge. The 95% CI of the intersection points for the FV model is $[188.7672, 190.3336]$.

Intersection Inference Under RV Model

Figure 3.20 (left) is a graph of the intersection of mean function curves for the first iteration using the RV approach. For this first iteration, they intersect at edge 189.2516. Figure 3.19 (right) shows the intersection of mean curves at each edge. The 95% CI of intersection points for RV is $[188.6193, 190.2381]$.

Comparison of the 95% CI for each dataset shows that considering uncertainty in variance does not yield an appreciable difference in the inference made on the intersection of mean curves of different datasets.

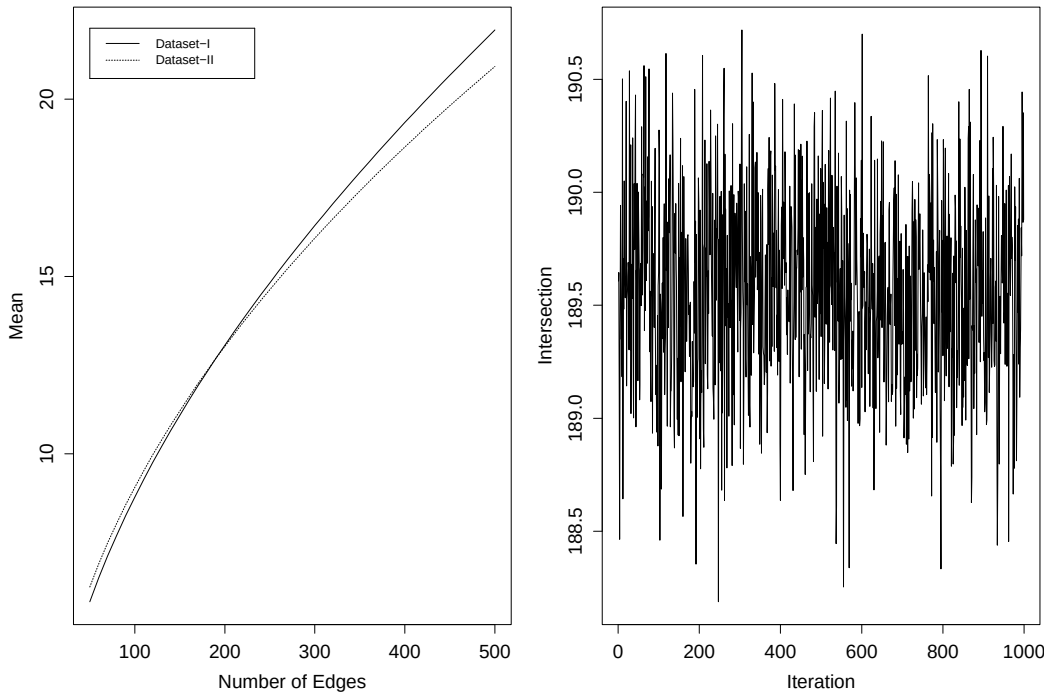

Figure 3.19: Left: Intersection of mean function curves for first iteration. Right: Intersection of mean function curves of Dataset-I and Dataset-II for each iteration.

3.2 Conclusion

Modeling the σ_j^2 as random allows for greater uncertainty in the estimates of A, B, and C. The sample sizes associated with Dataset-I and Dataset-II were so large, however, that there is relatively little uncertainty associated with the empirical estimates of the σ_j^2s. Note that, with the RV approach, the 95% CI for σ_{50}^2 is [0.8944,0.8946], and the empirical σ_{50}^2 equals to 0.8953. Hence, inference on the μ_js for the two datasets shows little or no discrepancy. Inference on the intersection location of the mean functions also shows little discrepancy as a result of the large datasets. However, inference on intersection location does show greater discrepancy when using smaller data size. We repeated the analysis using only 50 observations at each edge. For the FV approach, the 95% CI for the intersection points is $[174.6608, 262.8411]$ and for the RV approach 95% CI for the intersection points is $[143.6612, 270.3738]$. This shows the RV approach has a greater uncertainty on 95% CI estimation.

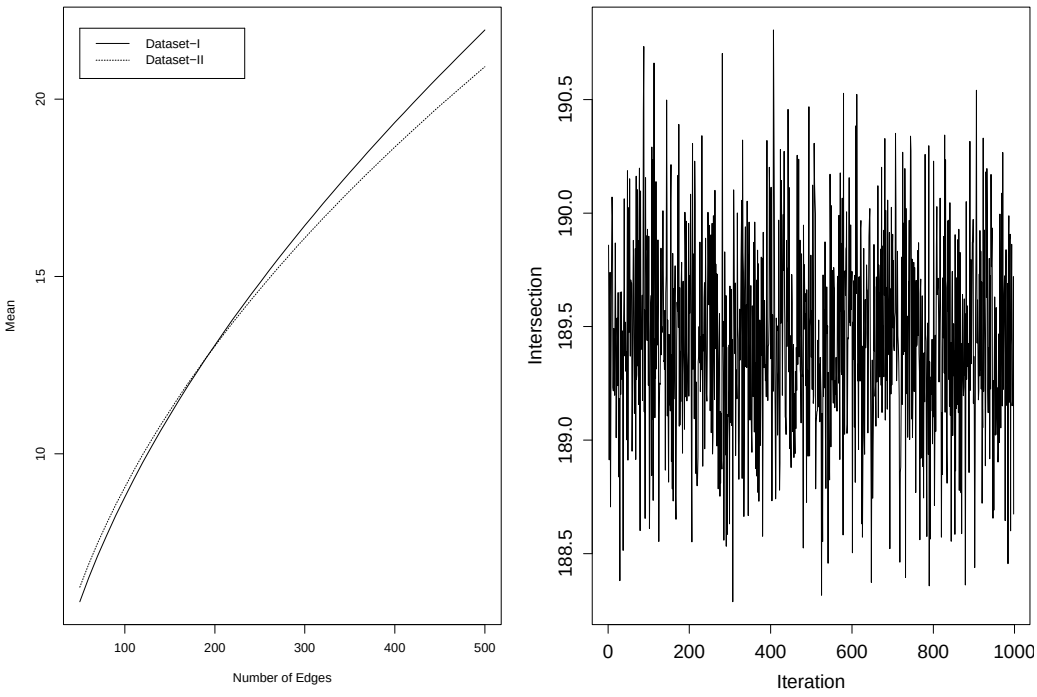

Figure 3.20: Left: Intersection of mean function curves for first iteration. Right: Intersection of mean function curves of Dataset-I and Dataset-II for each iteration.

3.3 Future Work

In our study we use 95% CIs when comparing results. In addition to CIs, considering Monte Carlo standard error in model comparisons may improve the comparison of models. Furthermore, as explained in section 1.3, based on histograms we propose x_{ij} to have normal distribution with mean μ_j and variance σ_j^2 at edge j. On the other hand, histograms are slightly right-skewed which may propose chi-square for other distribution for data points at each edge in further analysis of the same data.

References

[1] Plunkett, P., Piatek, M., Dobay, A., Kern, J.C., Millett, K.C., Stasiak, A., and Rawdon, E.J. "Total curvature and total torsion of knotted polymers". *Macromolecules*, 40(10): 3860-3867, 2007.

[2] Adams C. 2004. "The Knot Book: An Elementary Introduction to the Mathematical Theory of Knots". American Mathematical Society.

[3] Jeffreys, H. "An Invariant Form for the Prior Probability in Estimation Problems". *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* 186 (1007): 453-461, 1946.

[4] Hastings W.K. "Monte Carlo Sampling Methods Using Markov Chains and Their Applications", *Biometrika*, 57(1):97-109, 1970.

[5] Chib S., Greenberg E. "Understanding the MetropolisHastings Algorithm". *The American Statistician*, 49(4), 327335, 1995

[6] Casella G. and George, E.I. "Explaining the Gibbs sampler". *The American Statistician*, 46:167-174, 1992.

[7] Gelman, A., Carlin, J.B.,Stern, H.S., and Rubin, D.B. 1995. "Bayesian Data Analysis". London: Chapman and Hall.