Irregularly Spaced Time Series Data with Time Scale Measurement Error

by

Pulindu Ratnasekera
B.Sc. (Hons.), University of Sri Jayewardenepura, Sri Lanka, 2010

Project Submitted in Partial Fulfillment
of the Requirements for the Degree of

Master of Science

in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Summer 2014

APPROVAL

Name: Pulindu Ratnasekera

Degree: Master of Science

Title of Project: Irregularly Spaced Time Series Data with Time Scale Measurement Error

Examining Committee: Dr. Derek Bingham
Professor, Chair

Dr. Dave Campbell,
Associate Professor, Senior Supervisor

Dr. Tim Swartz,
Professor, Supervisor

Dr. Jiguo Cao,
Associate Professor, Internal Examiner

Date Defended/Approved: May 23rd, 2014

Partial Copyright Licence

Abstract

This project can be mainly divided into two sections. In the first section it attempts to model an irregularly spaced time series data where time scale is being measured with a measurement error. Modelling an irregularly spaced time series data alone is quite challenging as traditional time series techniques only capture equally/regularly spaced time series data. In addition to that, the measurement error in the time scale make it even more challenging to incorporate measurement error models and functional approaches to model the time series. Thus, this project is based on a Bayesian approach to model a flexible regression function when the time scale is being measured with a measurement error. The regression functions are modelled with regression P-splines and the exploration of posterior is carried out using a fully Bayesian method that uses Markov chain monte carlo (MCMC) techniques. In section two, we identify the relationship/dependency between two irregularly spaced time series data sets which were modelled using regression P-splines and a fully Bayesian method, using windowed moving correlations. The validity of the suggested methodology is then explored using two simulations. It is then applied on two irregularly spaced time series data sets each subjected to measurement errors in time scale to identify the dependency between them in terms of statistically significant correlations.

To my loving family in Sri Lanka...

Acknowledgments

First and foremost, I would like to express my gratitude to my supervisor Dr. David Alexander Campbell for the useful comments, remarks and engagement through the learning process of this master's project as well as during my master's program. Secondly, I would like to thank Dr. Tim Swartz and I am most indebted to him for reposing his trust and confidence in me which enabled me to begin this great educational experience at Simon Fraser University.

A special word of thanks should go to my examining committee Dr. David Alexander Campbell, Dr. Tim Swartz and Dr. Jiguo Cao for a patient hearing and for the valuable inputs which I have incorporated into my master's project. I would like to take this opportunity to thank all the faculty members at the Department of Statistics and Actuarial Science who taught me during last two years especially, Dr. David Alexander Campbell, Dr. Tim Swartz, Dr. Jiguo Cao and Mr. Ian Bercovitz. I am grateful for the financial support provided by the Department of Statistics and Actuarial Science. Special thanks to Statistics Workshop Manager Robin Insley for his support and guidance during my time at Simon Fraser University. My sincere gratitude to Sadika, Kelly, and Charlene for their kind assistance.

Furthermore, I would like to thank Chris Carleton for providing me an interesting research idea and Shirin Golchi, Abdollah Safari for their assistance during my master's project. Thanks also to my graduate student colleagues for their friendship and camaraderie and for the fun times we had together. To Wijendra, Kanna and Harsha my Sri Lankan graduate colleagues at SFU, I want to say thanks for helping me with my studies as well for settling in Vancouver.

Finally and most importantly I am fortunate to have a great family in Sri Lanka. I would not have come this far without their encouragement.

Contents

List of Figures

Chapter 1

Introduction

Time series methods are used to identify patterns and for prediction purposes. Time series methods work best for regularly spaced data. For irregularly spaced time series data, methods are still developing. When we encounter measurement errors or errors in covariates in relation to irregularly spaced time series data, even the methods that are currently available tend to fail. Thus, the objective of this project is to address/explore methods for irregularly spaced time series data when the covariate is subject to measurement errors.

1.1 Basics of Time Series

Time series data can be described as sequence of measurements of a variable collected over time. They can be divided mainly into univariate and multivariate time series data and most of the time those measurements are made at regular time intervals. In comparison to standard linear regression, time series data are not necessarily independent and not necessarily identically distributed. One defining characteristic of time series is that this is a list of observations where the ordering matters and changing the order could change the meaning of the data. Time series models describe important patterns, identify of the effect of past data on present data and, forecast future values. We can divide time series models mainly in to two components based on the time domain. The first is, ordinary regression models that use time indices as variables and these can be helpful for an initial description of the data and form the basis of several simple forecasting methods. The second is, models that relate the present value of a series to past values and past prediction errors and these are called ARMA (Auto-Regressive Moving Average) models.

One of the simplest ARMA models is the AR(1) model and it stands for Auto-Regressive order 1 model. The order of the model indicates how many previous measurements we use to predict the

present time. Using parameters δ and θ for predicting state X at time t, the AR(1) model,

$$x_t = \delta + \theta_1 x_{t-1} + \omega_t$$

includes ω_t as independently and identically distributed normal errors with mean zero and a constant variance. In order to assess the order of the AR model we could use the Autocorrelation Function(ACF) and Partial Autocorrelation Function(PACF). ACF gives correlations between the series x_t and lagged values of the series. On the other hand partial correlations are conditional correlations. If we consider a regression context in which y being the response variable and x_1, x_2, and x_3 are predictor variables, the partial correlation between y and x_3 is the correlation between the variables determined taking into account how both y and x_3 are related to x_1 and x_2.

The other part of a ARMA model is the MA component. A moving average (MA) term in a time series model is a past error and again the order of the moving average too can be assessed by looking at the combination of both ACF and PACF. The simplest of the MA models which is MA(1) which models states X at time t with parameters μ and θ as,

$$x_t = \mu + \omega_t + \theta_1 \omega_{t-1}.$$

Again, ω_t can be described as independently and identically distributed normal errors with mean zero and a constant variance.

1.2 Handling Irregularly Spaced Time Series Data

The models discussed in section 1.1 assume that measurements are made at regular intervals. They are therefore not appropriate for irregularly spaced time series (IRTS) data. When it comes to IRTS the key difference from regular series is that we need to keep track of both $\{t_n, X_n\}$, as the observation times are not constant. A very common approach of modelling IRTS is to convert them into regular time series using techniques such as interpolation, and then model the resultant series using regular time series methodologies. Such methods results in higher bias/errors.

As a result it can be said that best way to model IRTS is to model directly from irregularly spaced data rather than converting them into regular spaced data. Some work has been carried out in the statistics literature on this aspect. One such published work [2], models irregularly spaced time series data directly using an extension of AR models considering both stationary and non-stationary circumstances.

Another study [5], presents a computer program (REDFIT), which estimates the AR(1) parameter

directly from unevenly spaced time series data (x_i, t_i) with arbitrary spacing in the following manner and x represent a discrete AR(1) process with times $t_i (i = 1, 2, ..., N)$.

$$x(t_i) = \rho_i x(t_{i-1}) + \epsilon_{t_i}$$
$$\rho_i = \exp(\frac{-(t_1 - t_{i-1})}{\tau})$$

Here the additional unknown quantity τ (a constant representing the characteristic time scale) is estimated from an irregularly spaced time series using a least squares algorithm. Then the estimated AR(1) model is transformed from the time domain into the frequency domain using Lomb-Scargle Fourier Transformation. That paper assumes that paleoclimatic data can be fitted via a AR(1) process, hence begins its model identification using the traditional AR(1) process. Thus the paper defines AR(1) process, autocorrelation function and its spectrum respectively for unevenly spaced time series data.

A study associated with GARCH (Generalized Autoregressive Conditional Hetroskedasticity) modelling [8], too looks into modelling unevenly spaced time series data. This paper looks at the modelling of hetroscedasticity in time series in the context of irregularly sampled time series data. The paper introduces a continuous version of the GARCH model, referred to as COGARCH with a pure jump Levy Process while preserving the main characteristics of the original GARCH process.

1.3 FDA approach of Handling Irregularly Spaced Time Series Data

The papers discussed above, have a common limitation, as they do not account for possible errors in the covariate. In other words there could be measurement errors related to covariates and this is not well captured in the papers discussed above. Thus, this project is an attempt made to model irregularly spaced time series data with errors in covariates from the Functional Data Analysis perspective.

1.3.1 Basics of Functional Data Analysis

Functional data analysis (FDA) is a branch of statistics that analyses data providing information about curves, surfaces or anything else varying over a continuum. The continuum is often time, but may also be spatial location, wavelength, probability, etc. It is worth mentioning that if the continuum is time, it doesn't matter whether time intervals are regularly spaced or not as FDA do account for both scenarios.

When these curves are estimated, it is the assumption that they are intrinsically smooth that defines a FDA. In particular, functional data analyses often make use of the information in the slopes and curvatures of curves, as reflected in their derivatives. Plots of first and second derivatives as functions of t (time), or plots of second derivative values as functions of first derivative values, may reveal important aspects of the processes generating the data. As a consequence, curve estimation methods designed to yield good derivative estimates can play a critical role in FDA.

Models for functional data and methods for their analysis may resemble those for conventional multivariate data, including linear and non-linear regression models, principal components analysis, and many others. More importantly in this project we will be using FDA to model irregularly spaced time series data with measurement errors in covariates.

1.3.2 Measurement Error Models

Measurement error models can be divided into two classes. They are classical measurement error models and Berkson measurement error models. The classical measurement error model assumes a probability distribution for the observed (error contaminated) covariate conditional on unobserved (uncontaminated) covariate. On the other hand Berkson measurement error model assumes a probability distribution for the unobserved (uncontaminated) covariate conditional on the observed (error contaminated) covariate. Out of the two commonly used measurement error models, the classical measurement error model will be used in this project and following set of equations provide a standard format of the model.

$$Y = m(X) + \epsilon$$

$$W = X + U$$

where Y being the response variable and X being the predictor variable with measurement error which is also not observable. Hence we observe W with error, and we use W to approximate X. Here we assume both ϵ and U follow normal distributions. Here our ultimate objective is to approximate our mean function $m(X)$ when covariate X is not observable.

1.4 Motivation / Objectives

The motivating factor which led to carry out this thesis was the limitation encountered in detecting the role of the climate change in the development and demise of Classic Maya civilisation [1]. The original study [1] attempts to make comparisons between two time series data sets which were

collected over a period from 300CE to 1750CE. The two time series data sets were irregularly spaced and had measurement error on the time axis. One of the objectives of [1] was to identify whether there was any relationship between these two time series data sets. However the detection was controversial because of the absence of well dated climate sequences.

Time series data sets recount rainfall amounts and its changes were constructed using Oxygen Isotope ($\delta^{18}O$) and Titanium concentrations. Measurements on Oxygen isotopes were taken from stalagmite samples from Yok Balum cave in terms of Uranium-thorium (U-Th) dates. Those dates had an analytical precision ranging from ± 1 to ± 17. With regard to Titanium concentrations, those measurement were taken from marine sediments from the Cariaco Basin. Analysis in [1] does not examine quantitative relationships between these data sets because they are irregularly spaced and subject to measurement error. Instead they report potential quantitative relationships.

Therefore this thesis attempts to address this limitation which is identifying the statistically significant correlations in the presence of uncertainty in the time scale. The three main objectives of the study are,

- Modelling irregularly spaced time series data
- Incorporating measurement error model into time series model
- Identification of dependency between two irregularly spaced time series functions

Methodology used to overcome these three challenges are discussed in chapter 3 and the analysis carried out to identify correlations between the Oxygen Isotope data and the Titanium data are discussed in chapter 4.

Chapter 2

Literature Review

This chapter will focus on some of the previous work that has been carried out in statistics literature in modelling data when covariates have a measurement error. In statistics literature, existing methods for dealing with measurement error problems can be categorised into functional and structural approaches. The main difference between these two approaches is that, structural approaches assume a parametric distribution for unobserved covariates, where as the functional approaches do not. However in the presence of measurement error the true covariates are not observable. Thus it is very hard to check whether the suggested parametric distribution satisfies the unobserved covariate or not. As a result, most of the time functional approaches are considered above the structural approaches when modelling data with measurement errors.

Even though structural approaches seem to be challenging, considerable work has been carried out modelling data at the presence of covariate measurement errors. Use of flexible parametric approach for avoiding biased inference on response covariate, due to mismeasured continuous covariates has been discussed in [9]. In this study authors consider skew-normal and flexible generalized t-distributions to model the unobserved true covariate. Then for inference and computational purposes they use a Bayesian approach based on conditional independence along with Markov chain monte carlo (MCMC) methods. [6] attempts to model birth weight using the gestational age, where gestational age associated with measurement error. In this study authors use both classical and Berkson measurement error models to handle measurement errors. With regard to parametric assumptions on the true covariate with measurement error, they use Box-Cox transformation rather than assuming that true covariates follow a normal distribution. For estimation and inference purposes, authors suggest a Bayesian approach rather than conventional maximum likelihood / EM algorithm. However they avoid using a fully Bayesian methodology to overcome the computational difficulties. [7] introduce the use of flexible parametric models (mixture of normals) at the presence of departures of standard parametric models when modelling with covariates having measurement

errors. This project considers both classical and Berkson measurement error models discussed above to demonstrate their methodology with a Bayesian (MCMC) approach in the estimation and inferential phase.

Functional approaches used in statistics literature to handle measurement errors are simulation-extrapolation (SIMEX), regression calibration (RC), conditional score approach (CS), corrected score approach (CTS) etc. [4] develops a semi-parametric estimation method for Accelerated Failure Time (AFT) model with covariate subject for measurement errors. They use the traditional measurement error model and estimation and inference is carried out using SIMEX approach.

Additionally [10] discusses the possibility of estimating a regression function non-parametrically at the presence of covariate measurement error. [10] uses Bayesian approaches in modelling a flexible regression function when the predictor variable is measured with the measurement error using a classical measurement error model. Here the regression function is modelled with smoothing splines and the estimation is carried out using partial (Iterative conditional modes) and fully (MCMC) Bayesian methods. Thus, this project uses an extended approach of [10] to model irregularly spaced time series data at the presence of measurement errors in the time domain.

Chapter 3

Methodology

Chapter 3.1 discusses the methodology associated with modelling irregularly spaced time series data in the presence of the measurement error in the time scale as a smooth function approximating the underlying process. Chapter 3.2 discusses the methodology behind identifying the dependency between approximated functions via Windowed moving correlations.

3.1 Approximating Curves with Errors in Covariates

This section introduces the notation for the measurement error model and fits an approximating curve using Bayesian smoothing methodology discussed in [10]. Our objective is to approximate a function $m(X_i)$ for the response R_i at time X_i using following measurement error model

$$R_i = m(X_i) + \epsilon_i, (i = 1, ..., n), \tag{3.1}$$

where ϵ_i is an independent random normal variable with mean 0 and variance σ_ϵ^2. X is not observable, hence surrogate from the observable W as follows,

$$W_{ij} = X_i + U_{ij}, \tag{3.2}$$

where U_{ij}s are independent normal errors with mean 0 and variance σ_u^2. As mentioned above, our primary objective is to approximate function $m(X_i)$ when the covariate X is not observable. For that we introduce another function $g(X_i)$ which is a natural cubic spline estimator of $m(X_i)$. We use the following log likelihood to approximate the mean function $m(X_i)$.

$$LogLikelihood \propto -n \log(\sigma_\epsilon^2) - \frac{1}{2\sigma_\epsilon^2} \sum_{i=1}^{n} \{R_i - g(X_i)\}^2$$

We use a partially improper Gaussian prior for $g(X_i)$ to control the roughness of the approximated mean function $m(X_i)$ $\left(Prior \propto (\frac{\gamma}{2}) exp(-\gamma \int_a^b \{g^{''}(x)\}^2 dx) \right)$. Here the penalty parameter γ controls the roughness of the approximation. If the penalty parameter is close to zero then $m(X)$ will be less smooth. On the other hand, if the penalty parameter tends to infinity, then the approximated curve will be smoother.

Therefore it can be said that this is a Bayesian representation for the penalised least squares estimator when its covariate is not observable by minimizing

$$S(g) = \frac{1}{2\sigma_\epsilon^2} \sum_{i=1}^{n} \{R_i - g(X_i)\}^2 + \left(\gamma \int_a^b \{g^{''}(x)\}^2 dx \right). \tag{3.3}$$

$g(X_i)$ can be represented as a linear combination of basis functions, $g(X) = \phi(X)^T \beta$. Where β is the vector of basis coefficients and $\phi(X) = \{\phi_1(X), ..., \phi_N(X)\}^T, N \lesssim n$ is the corresponding spline basis. The Basis function considered in this project is a truncated polynomial basis and can be written by adding any $(x - t_k)^p$ component to basis if the corresponding $(x - t_k)$ term is positive which is indicated by the plus sign, $\phi(X) = (1, x, x^2, ..., x^p, (x - t_1)_+^p, ..., (x - t_k)_+^p)^T$. We define $t_1, ..., t_k$ to be k equally spaced knots on the range of X_i for convenience purposes, even though they do not necessarily be equally spaced. However the same methodology can be used even with a Fourier basis with some minor changes. Those changes required will be discussed in chapter 3.1.2. We can re-write equation (3.3) with β and $\phi(X)$ in the following form.

$$\frac{1}{2\sigma_\epsilon^2} \sum_{i=1}^{n} \{R_i - \phi(X_i)^T \beta\}^2 + \gamma \beta^T D \beta \tag{3.4}$$

In equation 3.4, D represents a vector with p+1 zeros and k 1's. Now we can think of β as a function of penalty parameter γ and then the optimal vector of coefficients can be obtained from the following expression, with Φ being a nxN matrix with i^{th} row equal to $\phi(X_i)^T$.

$$\hat{\beta} = (\Phi^T \Phi + \gamma D)^{-1} \Phi^T R \tag{3.5}$$

At this point we have five parameters of interest $(\beta, X, \sigma_\epsilon^2, \sigma_u^2, \gamma)$ and their joint posterior distribution can be written using a latent variable model in the following form,

$$[\beta, X, \sigma_\epsilon^2, \sigma_u^2, \gamma | R, W] \propto [R|\beta, X, \sigma_\epsilon^2][X|W, \sigma_u^2][\beta|\gamma][\sigma_u^2][\sigma_\epsilon^2][W][\gamma] \tag{3.6}$$

having prior distributions on all parameters including the hyper parameters μ_x, σ_x (parameters of the prior distribution of X) and the variance components $\sigma_\epsilon, \sigma_U$.

List of prior distributions,

- $\sigma_\epsilon^2 \sim IG(A_\epsilon, C_\epsilon)$ - Variance of the error in R
- $\sigma_u^2 \sim IG(A_u, C_u)$ - Variance of the error in X
- $\gamma \sim G(A_\gamma, C_\gamma)$ - Penalty parameter
- $\mu_x \sim N(d_x, \tau_x^2)$ - Mean of the prior distribution of X
- $\sigma_x^2 \sim IG(A_x, C_x)$ - Variance of the prior distribution of X

Hence we can write joint posterior density (as per [10]) as follows.

$$exp\{-\frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{n}\{R_i - \phi(X_i)^T\beta\}^2 - \frac{1}{2\sigma_u^2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\{W_{ij} - X_i\}^2 - \frac{1}{2\sigma_x^2}\sum_{i=1}^{n}\{X_i - \mu_x\}^2 - \frac{1}{2\tau_x^2}\{\mu_x - d_x\}^2\}*$$

$$exp\{-(\gamma/2)\{\phi(X)^T\beta\}^T D\{\phi(X)^T\beta\} - \frac{1}{C_\epsilon\sigma_\epsilon^2} - \frac{1}{C_U\sigma_U^2} - \frac{\gamma}{C_\gamma} - \frac{1}{C_x\sigma_x^2}\}*$$

$$\sigma_\epsilon^{-2(n/2+A_\epsilon+1)} * \sigma_U^{-2(1/2\sum_{i=1}^{n} m_i + A_U + 1)} * \sigma_x^{-2(n/2+A_x+1)} * \gamma^{A_\gamma + M/2 - 1}$$

3.1.1 Bayesian Implementation for Regression P-Splines for measurement Error Model

This section discusses the estimation of function $g(X)$ which approximates the true function $m(X)$. The estimation is based on the methodology suggested in [10] which gives emphasis to $\phi(X)$, the basis function (in this case a truncated polynomial basis) and β, the basis coefficients.

For fixed-knot P-splines, g(x) can be written as, $g(X) = \phi^T(X)\beta$. We can write $g = (g(X_1); ...; g(X_n))^T$ as $g = \Phi\beta$ where Φ being, a vector $\phi(X)$ evaluated based on a vector X. Here N (=1+p+k) represents number of basis functions where, p is the order of the polynomial and k is the number of knots. With regard to $\phi(X)$, we apportion $\phi(X) = (\phi_1^T(X), \phi_2^T(X)^T)$ where $\phi_1^T(X)$ is the first p+1 elements and similarly apportion β as $\beta = (\beta_1^T, \beta_2^T)^T$. The improper prior (a Gaussian with infinite variance) of β_1 can be written as N(0, δI) and the Gaussian prior of β_2 can be written as N(0, $\gamma^{(-1)}I$). The diagonal matrix D_* can be written as, $D_* = \sigma_\epsilon^2 diag(I/\delta, \gamma I)$.

β, can be sampled using the following full conditional distribution (obtained from the joint posterior

distribution discussed in chapter 3.1) and its respective parameters in the following manner.

$$\beta|R, X, W \sim N(QH, Q)$$

$$H = \sigma_\epsilon^{-2} \sum_{i=1}^{n} \phi(X_i) R_i = \sigma_\epsilon^{-2} \Phi^T R$$

$$Q = \sigma_\epsilon^2 \left(\sum_{i=1}^{n} \phi(X_i)\phi^T(X_i) + D_* \right)^{-1} = \sigma_\epsilon^{-2} (\Phi^T \Phi + D_*)^{-1}$$

Once we have the basis coefficients β together with the basis function evaluated at initial covariate values, then we could find an initial estimate for $g(x)$ as follows.

$$g(x) = \phi^T(x)\beta$$

This initial sample of $g(x)$ can be used to sample values from rest of the full condition distributions which were extracted from the joint posterior distribution discussed in chapter 3.1. Full conditional distributions obtained were as follows (as stated in [10]).

Full conditional distribution of "X",

$$[X_i|W_i, g, \sigma_\epsilon^2, \sigma_u^2, R, W] \propto exp\left(-\frac{1}{2\sigma_u^2} \sum_{j=1}^{m_i} (W_{ij} - X_i)^2 - \frac{1}{2\sigma_\epsilon^2}(R_i - g(X_i))^2 - \frac{1}{2\sigma_x^2}(X_i - \mu_x) \right) \tag{3.7}$$

Full conditional distribution of "σ_ϵ^2",

$$\sigma_\epsilon^2|g, X, R, W \sim IG\left(A_\epsilon + n/2, [1/C_\epsilon, (1/2)\sum_{i=1}^{n}\{R_i - g(X_i)\}^2]^{-1} \right) \tag{3.8}$$

Full conditional distribution of "σ_u^2",

$$\sigma_u^2|X \sim IG\left(A_u + (1/2)\sum_{i=1}^{n} m_i, [1/C_u, (1/2)\sum_{i=1}^{n}\sum_{j=1}^{m_i}\{W_{ij} - X_i\}^2]^{-1} \right) \tag{3.9}$$

Full conditional distribution of "γ",

$$\gamma|\beta \sim G\left(A_\gamma + \frac{k}{2}, [1/C_\gamma + (1/2)\beta_2^T \beta_2]^{-1} \right) \tag{3.10}$$

Full conditional distribution of "μ_x",

$$\mu_x|X \sim normal\left(\frac{n\bar{x}\tau_x + d_x\sigma_x^2}{n\tau_x^2 + \sigma_x^2}, \frac{\sigma_x^2\tau_x^2}{(nt_x^2 + \sigma_x^2)}\right) \tag{3.11}$$

Full conditional distribution of "σ_x",

$$\sigma_x|X \sim IG\left(A_x + n/2, [C_x^{-1} + (1/2)\sum_{i=1}^{m}(X_i - \mu_x)^2]^{-1}\right) \tag{3.12}$$

Except for the full conditional distribution of X, all other full conditional distributions have known forms. Therefore the project uses Metropolis Hastings within Gibbs to sample from full conditional distributions. In Metropolis Hastings step, candidate values of X_{prop} are generated from a normal proposal distribution with a mean of current value of $X_{current}$. This Project uses an adaptive Metropolis Hastings algorithm. Thus the algorithm begins with an initial guess for standard deviation, which is adjusted during the sampling process according to the acceptance rate. The adjustment stops when number of iterations reach half of the sample size and this first half of the samples will be discarded as burn-in.

3.1.2 Changes required when using a Fourier Basis Function

Use of a Fourier basis in place of a truncated polynomial basis require few key changes to the methodology discussed in chapter 3.1. Changes required have an effect on estimation of β, basis coefficients and full conditional distribution $\gamma|\beta$.

Estimation of basis coefficients β requires the calculation of matrix D_*. In comparison to degree of polynomial and number of knots in truncated polynomial basis, Fourier basis only has the number of basis functions as a parameter. Thus D_* becomes a diagonal matrix of the following form.

$$D_* = diag(1/\delta, nbasis)$$

Fourier basis also affects the form of full conditional distribution $\gamma|\beta$. Fourier basis no longer needs to be apportioned into basis coefficients, $\beta = (\beta_1^T, \beta_2^T)^T$. Thus the full conditional distribution will take the following form.

$$\gamma|\beta \sim G(A_\gamma + nbasis/2, \{C_\gamma^{-1} + \beta^T\beta/2\}^{-1})$$

Apart from these changes, methodology stated in [10] can be followed when approximating curves

with Fourier basis.

3.2 Windowed Moving Correlations on Approximated curves with Errors in Covariates

The primary objective of this project is to identify the dependency between the two time series data sets (Oxygen Isotope data and Titanium Data) in terms of significant correlations. The concept of windowed moving correlation can be used to identify this dependency between the two time series data sets. The moving correlations especially becomes useful in identifying dependency when we cannot differentiate our time series data sets between response and predictor functions.

The methodology for calculation of windowed moving correlations is somewhat similar to the calculation of moving averages, that we use in traditional time series analysis. When calculating windowed moving correlations we define a ”Window Size”, which is similar to ”Moving Average Cycle” in moving averages. In moving correlations we calculate the correlations between our two time series data sets for a predetermined window size. Window size is a subset of the sample size and in order to calculate moving correlations both data sets should have a equal length.

In this project, the moving correlations were calculated using the following methodology. The study obtains estimated basis coefficients (β) from each of the two posterior samples relating to two data sets of interest. Those coefficients were evaluated on a fine grid which gives a sample size of equal length to each data set. Those two vectors of equal length were then used in the windowed moving correlation calculations. The final result of the code will be a vector of windowed moving correlations. From which we could determine whether our two time series data sets (Oxygen Isotope data and Titanium Data) relate to each other with significant correlations or not.

Chapter 4

Simulation Study

This chapter demonstrates the methodology suggested in chapter 3 by using two simulations followed by a analysis on a real data set. In chapter 4.1 we explore the methodology suggested in chapter 3.1. It approximates a curve which is subject to a covariate measurement error using a Truncated polynomial Basis. This is a replication of one of the simulations in [10].

Simulation 2 will be based on a real life data set (Gait Data), where it uses a Fourier basis to approximate both response and predictor functions which are subjected to measurement error. The dependency between response and predictor functions will be evaluated using the functional regression mechanism and the windowed moving correlations. Time series data will be used in chapter 4.3, where it explores the method suggested in chapter 3.1 with a Truncated polynomial basis in approximating both Oxygen Isotope data and Titanium data. The dependency between the two time series data sets will be evaluated using windowed moving correlations. Here we use windowed moving correlations in place of functional regression mechanism as we cannot differentiate Oxygen and Titanium data in terms of predictor and response functions.

4.1 Simulation 1 - Truncated Polynomial Basis

Simulation considered in this section is a recreation of [10] and was carried out using the following model.

$$m(x) = \frac{sin(\pi x/2)}{1 + 2x^2\{sign(x) + 1\}}$$

The objective is to find an approximated curve for m(x), where X is subjected to measurement error. The simulation was carried out with a sample size of 100. A Truncated polynomial basis was used with 15 basis functions (polynomial degree(p)=4, number of knots(k)=10). In this simulation

we generated two values for W (m=2) to approximate the unobservable covariate X. Noise in W was created using a normal distribution, N(mean=0, SD=0.05). This noise was added on top of the desired range of X which is a sequence from -3 to +3 having a sample size of 100. The measurement error model used in the simulation takes the following form.

$$R_i = m(X_i) + \epsilon_i, (i = 1, ..., 100)$$

$$W_{ij} = X_i + U_{ij}$$

$$\epsilon_i \sim N(0, \sigma_\epsilon^2)$$

$$U_{ij} \sim N(0, \sigma_u^2)$$

Following prior distributions were used in the simulation.

- $\sigma_\epsilon^2 \sim IG(1, 1)$ - Variance of the error of R
- $\sigma_u^2 \sim IG(1, 1)$ - Variance of the error of X
- $\gamma \sim G(3, 1000)$ - Penalty parameter
- $\mu_x \sim N(0, 10^2)$ - Mean of the prior distribution of X
- $\sigma_x^2 \sim IG(1, 1)$ - Variance of the prior distribution of X

With these prior parameters we update full conditional distributions (eq. 3.7 to 3.12) and use these updated distributions for sampling purposes. The sampling was carried out iteratively for 10,000 iterations and the first 5000 samples were removed as a burn-in. The rest of the samples were used in the approximation. The figure 4.1 gives the approximation for the function m(X). The trace plots of each of the full conditional distributions from Gibbs sampling methodology are provided in figure 4.2 and samples obtained for unobservable X from Metropolis Hastings algorithm are shown in figure 4.3.

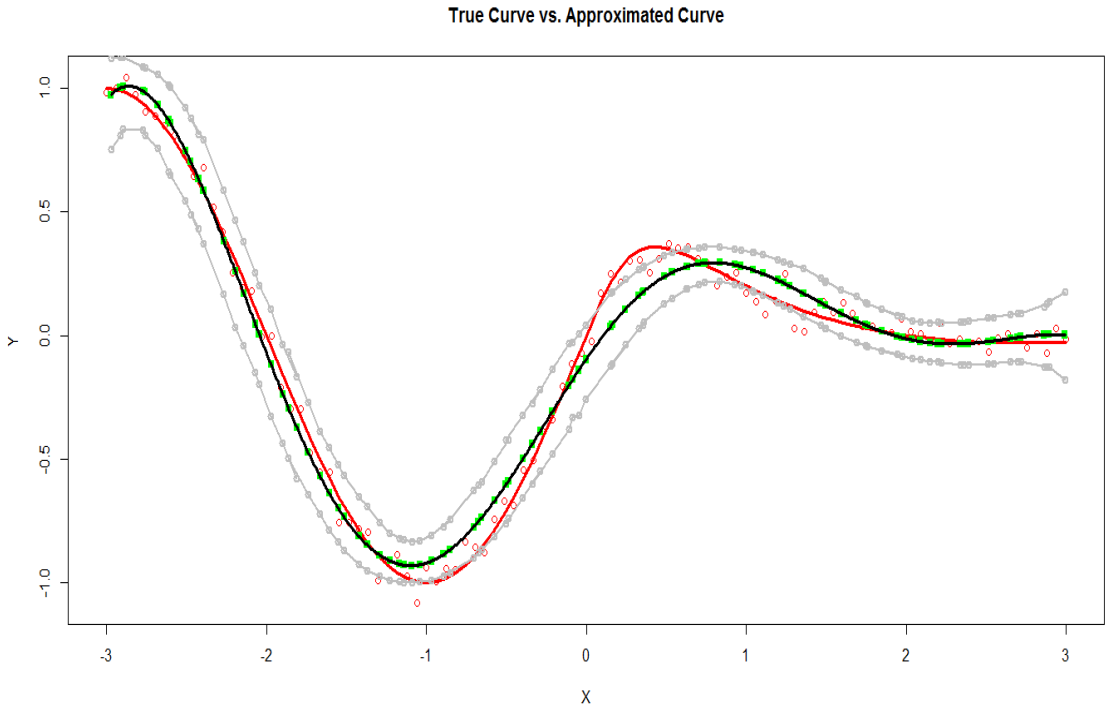

Figure 4.1: Approximated Curve for True function m(X):

Figure compares the approximated curve with the true curve. The true curve is given in "Red". "Green" points are the point wise estimates for the true curve. The "Black" curve is the curve which approximates the true curve which is obtained by evaluating the estimated coefficients on a fine grid. The approximation was carried out using posterior means on the second half of the iterations as first half was discarded as burn-in. Estimation uncertainty is indicated by the two Gray lines which are 90% point wise confidence intervals.

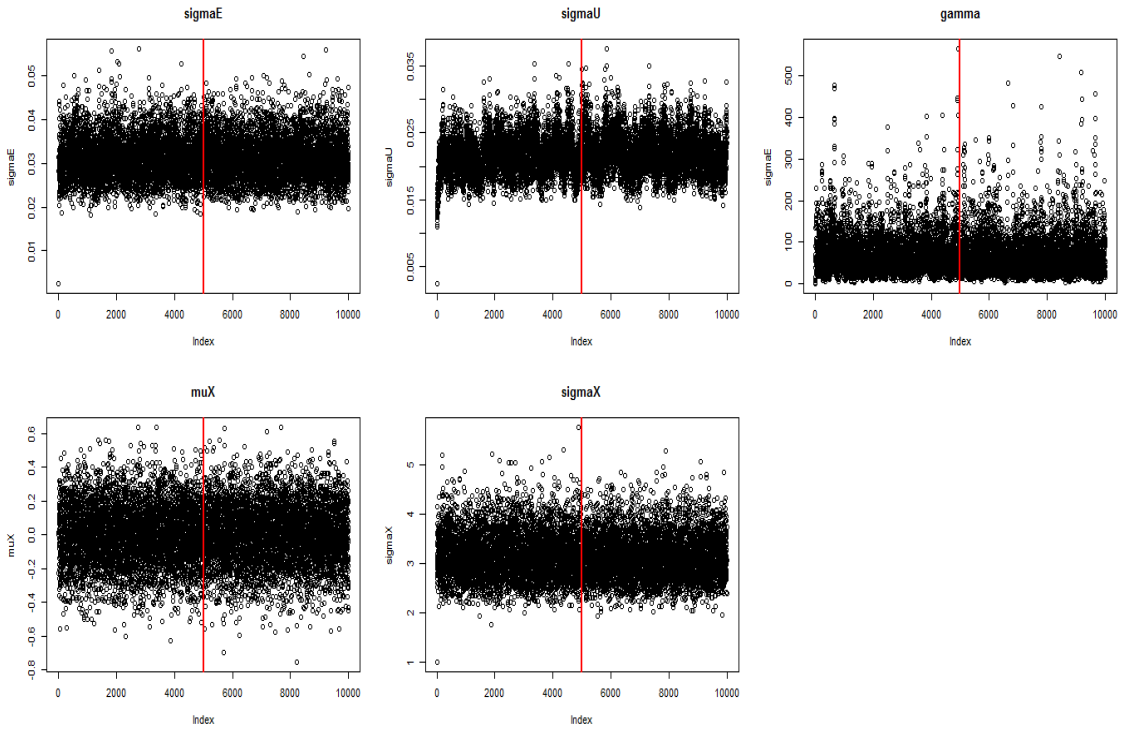

Figure 4.2: Trace Plots - Full Conditional Distributions from Gibbs Sampling:

Five figures in the graph represent each of the full conditional distributions sampled during the MCMC iterations. The vertical lines on each of the five figures give an idea on burn-in. Estimation we carried out using the samples obtained beyond the vertical line, as previous samples were discarded as burn-in.

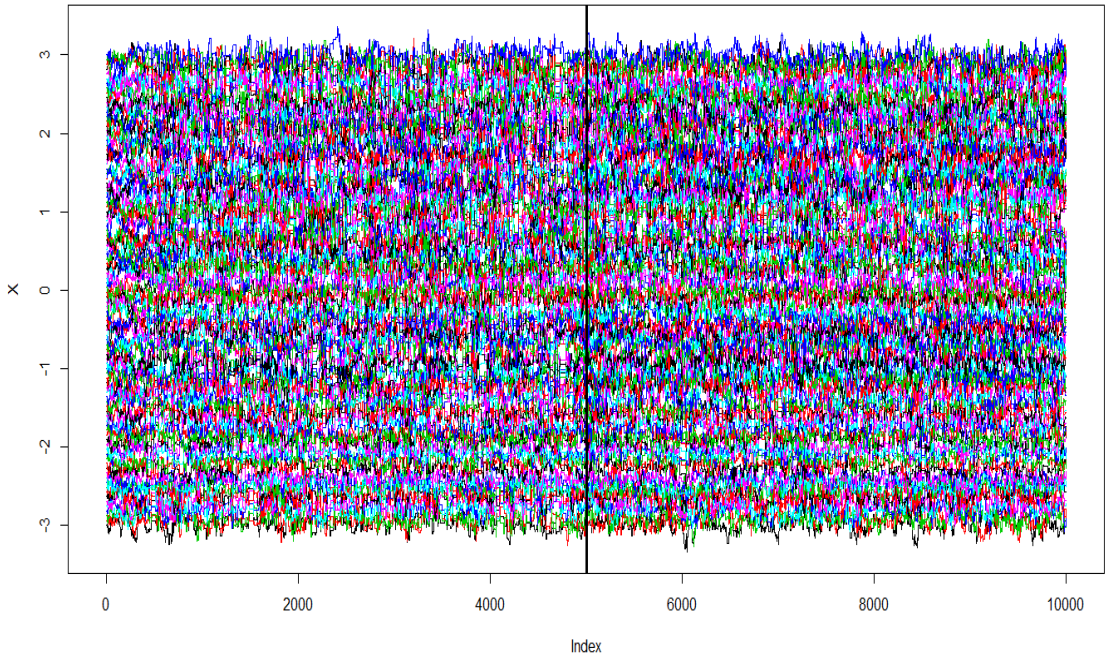

Figure 4.3: Samples of Unobservable X from Metropolis Hastings algorithm:

Figure represent the sampling of unobservable covariate(X) using the M-H algorithm. The vertical line at iteration 5000 indicate the stoppage for adapting the M-H algorithm for variance adjustment and also the point for burn-in. The samples prior to vertical axis were discarded as burn-in and the samples after vertical axis were only considered for estimation purposes

Figure 4.1 indicates that the methodology in chapter 3.1 succeeds in providing a reasonable approximation for the true curve. The figures 4.2 and 4.3 ensure that sampling from full conditionals are stable. In order to identify the quality of the approximation, we calculated Mean Square Errors (MSE) for different values of σ_U.

- MSE = Bias^2 + Variance
- Bias = mean(true value - fitted)
- Variance = Variance of point wise estimate for a time point over MCMC iterations averaged over time points
- Average MSE over time points = Average Bias(over time points and MCMC iterations) + Average Variance(of fitted time point values over MCMC runs)

Project uses two MSE calculations. The First, in the direction of Y axis for different values of measurement error and the other in the direction of X axis for different values of measurement error. Therefore an ideal plot would be a plot with increasing MSE values for increasing measurement errors. To generate noisy data, we considered following values for the standard deviation (σ_U).

$$\sigma_U = [0.05, 0.1, 0.15, 0.25, 0.4, 0.5, 0.75, 1, 2, 3]$$

$$U_{ij} \propto N(0, \sigma_u^2)$$

$$W_{ij} = X_i + U_{ij}$$

$$R_i = m(rowMeans(W_i)) + \epsilon_i$$

For different values of σ_U, we have different W_i and R_i. The MSE calculations were carried out in the directions of both Y axis and X axis using those values respectively.

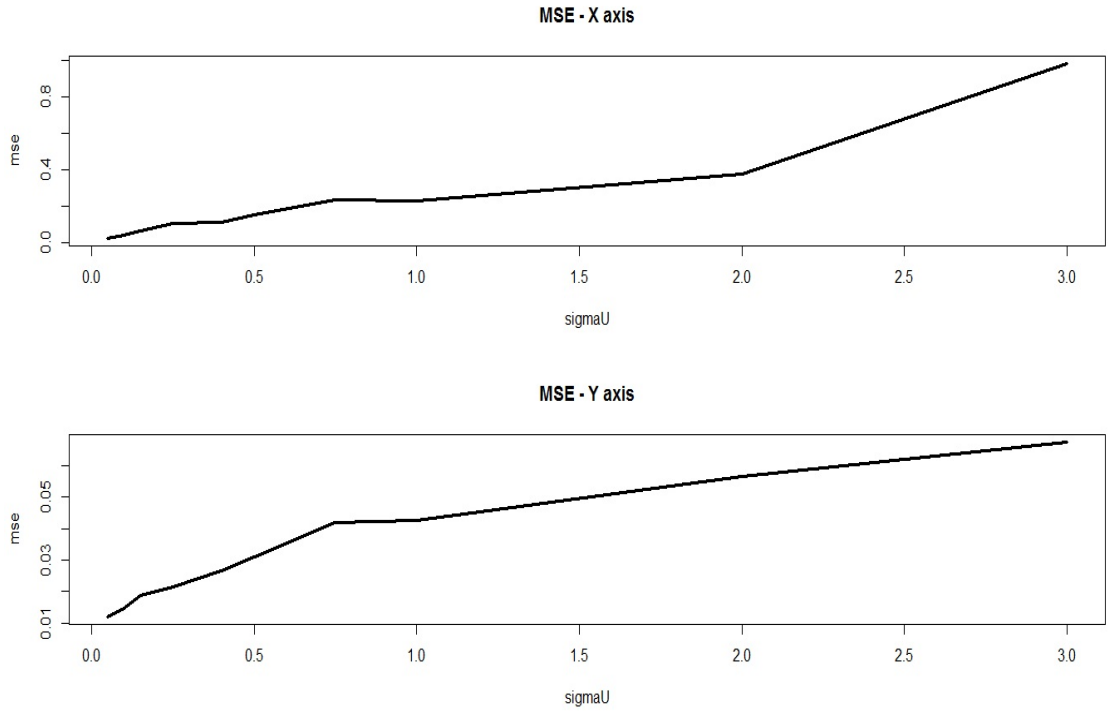

Figure 4.4: MSE for different values of noise on Unobservable Covariate:

Top panel of the figure provides the MSE calculation for approximated Xs for different values of σ_U. Similarly, bottom panel provides the MSE calculation for approximated Rs for different values of σ_U. Both MSE curves suggest that method suggest in chapter 3.1 succeeds in providing a reasonable fit as MSE tend to increase with more noise on data.

4.2 Simulation 2 - Gait Data / Fourier Basis

Simulation 2 will be based on the real data set, namely Gait data [3] and explores the methodology suggested in chapters 3.1 and 3.2. The Gait data is from Motion Analysis Laboratory at Children's Hospital, San Diego, CA, and consisted of the angles formed by the hip and the knee of each of 39 children over their gait cycles. Objective is to measure "The Control of the Hip Angle that has over Knee Angle". However it should be noted that this data set does not have a measurement error in its time axis. Therefore we introduce a measurement error to the data set for our simulation purposes.

During this simulation [10] was used to smooth both knee angle and hip angles for each of the 39 students. The smoothing process was carried out using 10,000 iterations, where 10,000 samples were obtained for each of the full conditional distributions using both Metropolis Hastings and Gibbs sampling methods. The first $1/2$ of the iterations were removed as burn-in and rest of the 5,000 samples were used in the estimation. Prior distributions that were considered in chapter 4.1 were used as it is.

In Gait data, time points had a range from 0.5 to 19.5. Thus basis range was set up from 0 to 20. The measurement error on covariate was generated randomly from a normal distribution with 0 mean and a variance of σ_U for different values of σ_U (0.05, 1, 0.15, 0.25, 0.4, 0.5, 0.75, 1, 2, 3). This simulation has a sample of size 20. It should be noted that the figure 4.5 was obtained with a measurement of 0.05 (σ_U). During the simulation, for each true observation two values (W) were generated having this measurement error. Figures 4.5 and 4.6 provide the approximated mean curves for both hip and knee angles of child 2.

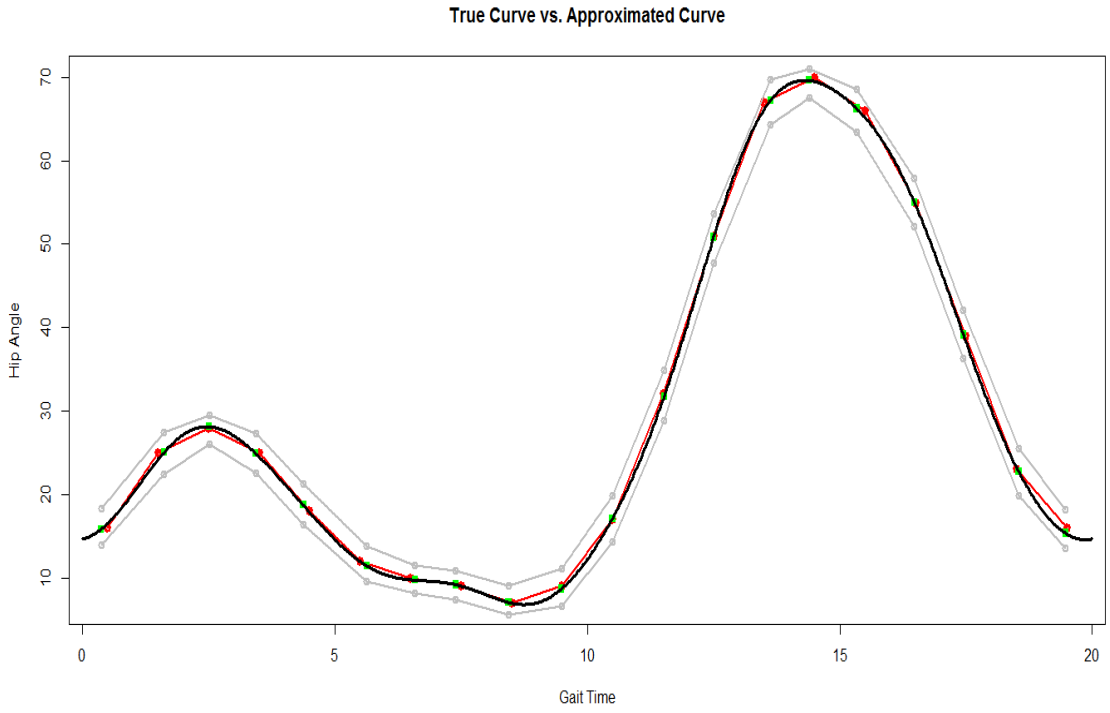

Figure 4.5: Approximated Curve for Knee Angle of Child 2:

Figure compares the approximated curve to the true curve of the knee angle of child 2 obtain over a time period from 0.5 to 19.5. True curve is given in "Red". "Green" points are the point wise estimates for the true curve. The "Black" curve is the curve which approximates the true curve which is obtained by evaluating the estimated coefficients on a fine grid. The approximation was carried out using posterior means on the second half of the iterations as first half was discarded as burn-in. Estimation uncertainty is indicated by the two Gray lines which are 90% point wise confidence intervals.

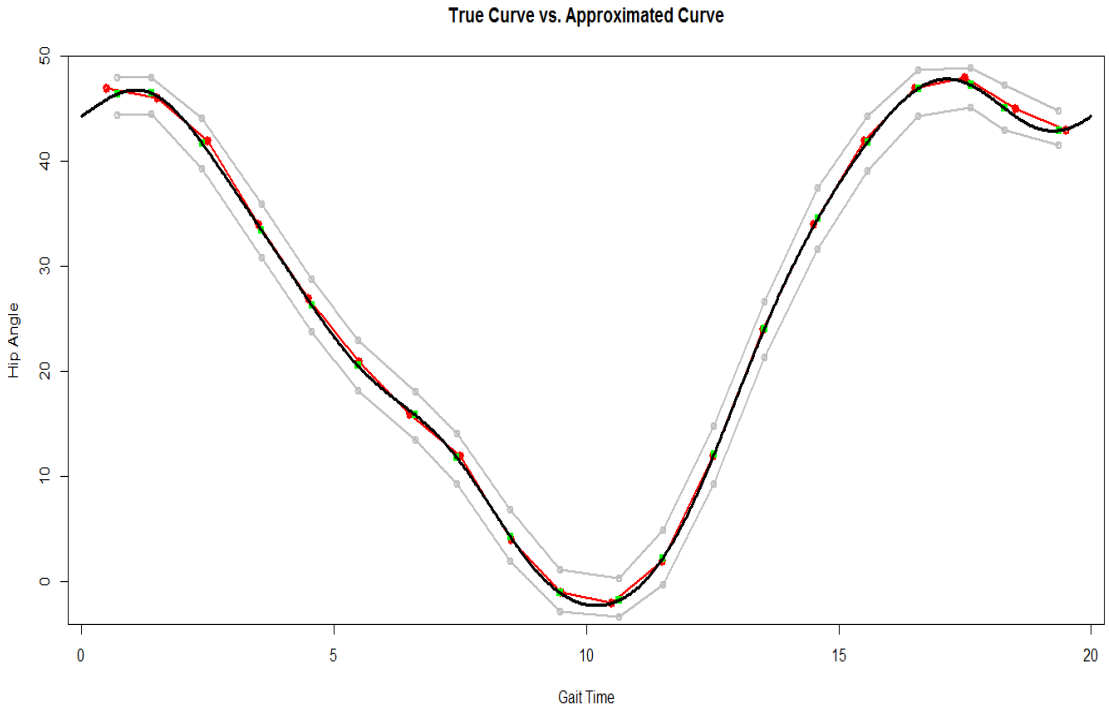

Figure 4.6: Approximated Curve for Hip Angle of Child 2:

Figure compares the approximated curve to the true curve of the hip angle of child 2 obtain over a time period from 0.5 to 19.5. True curve is given in "Red". "Green" points are the point wise estimates for the true curve. The "Black" curve is the curve which approximates the true curve which is obtained by evaluating the estimated coefficients on a fine grid. The approximation was carried out using posterior means on the second half of the iterations as first half was discarded as burn-in. Estimation uncertainty is indicated by the two Gray lines which are 90% point wise confidence intervals.

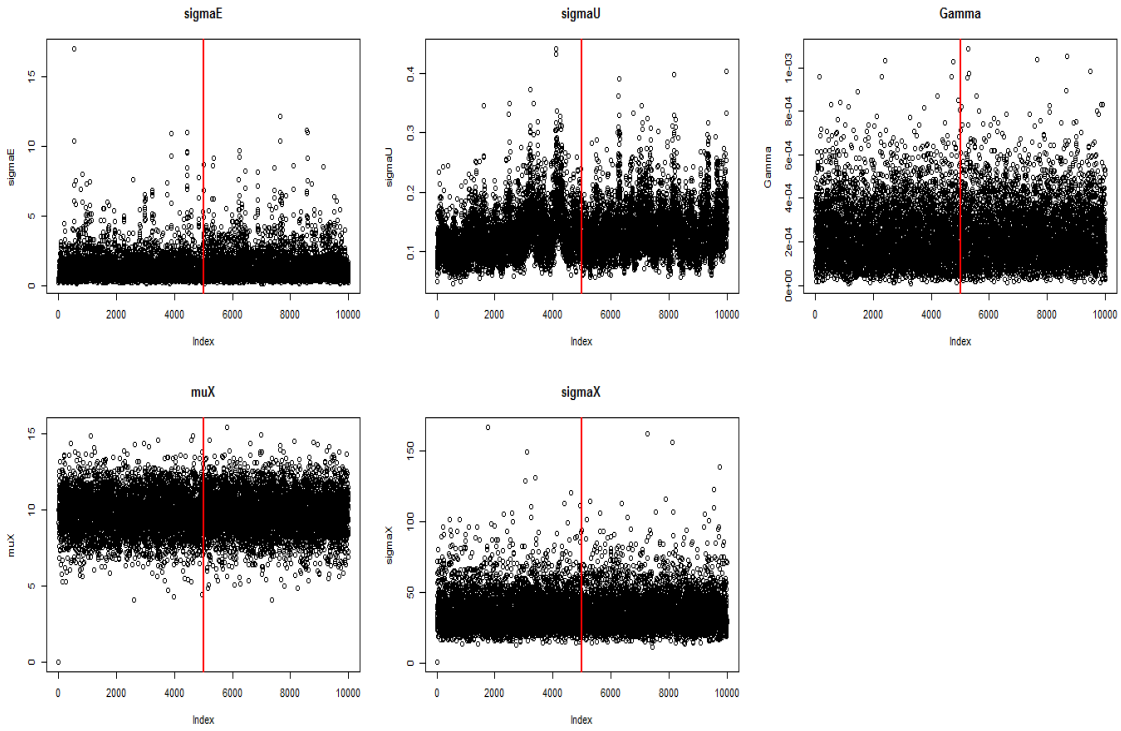

Figure 4.7: Convergence of the Full Conditional Distributions - Knee Angle of Child 2:

Five figures on the graph represent each of the full conditional distributions sampled during the MCMC iterations. The vertical lines on each of the five figures give an idea on burn-in. For estimation we used the samples obtained beyond the vertical line as previous samples were discarded as burn-in.

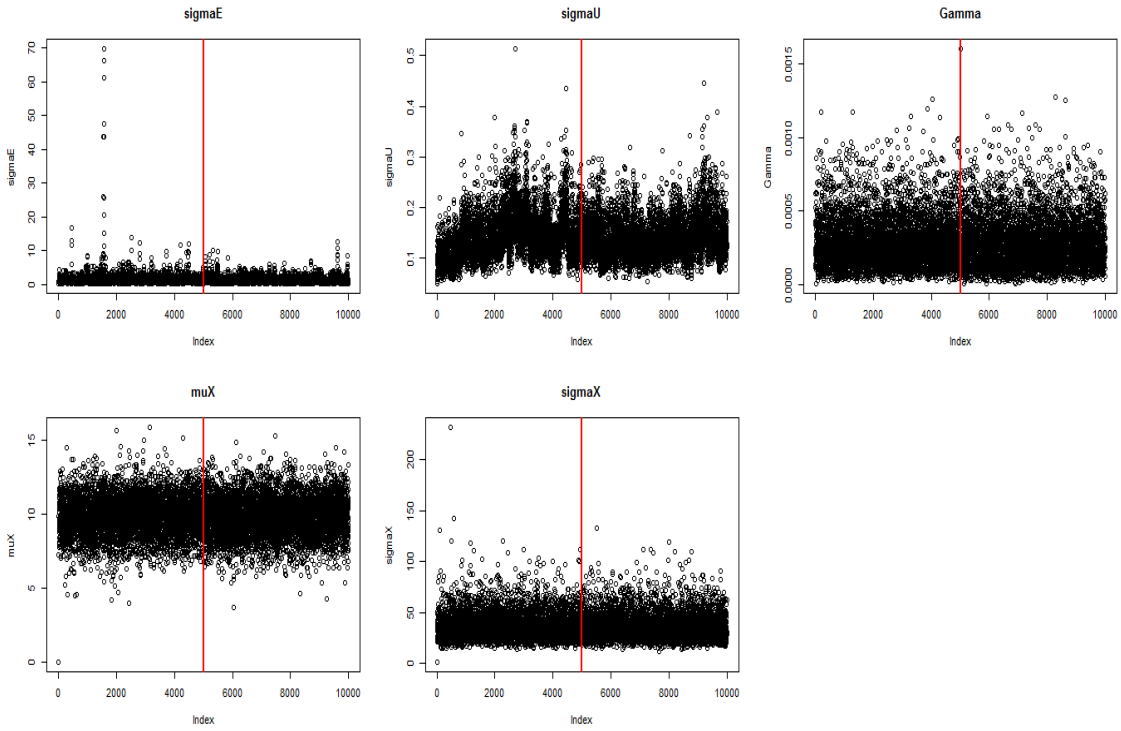

Figure 4.8: Convergence of the Full Conditional Distributions - Hip Angle of Child 2:

Five figures on the graph represent each of the full conditional distributions sampled during the MCMC iterations. The vertical lines on each of the five figures give an idea on burn-in. For estimation we used the samples obtained beyond the vertical line as previous samples were discarded as burn-in.

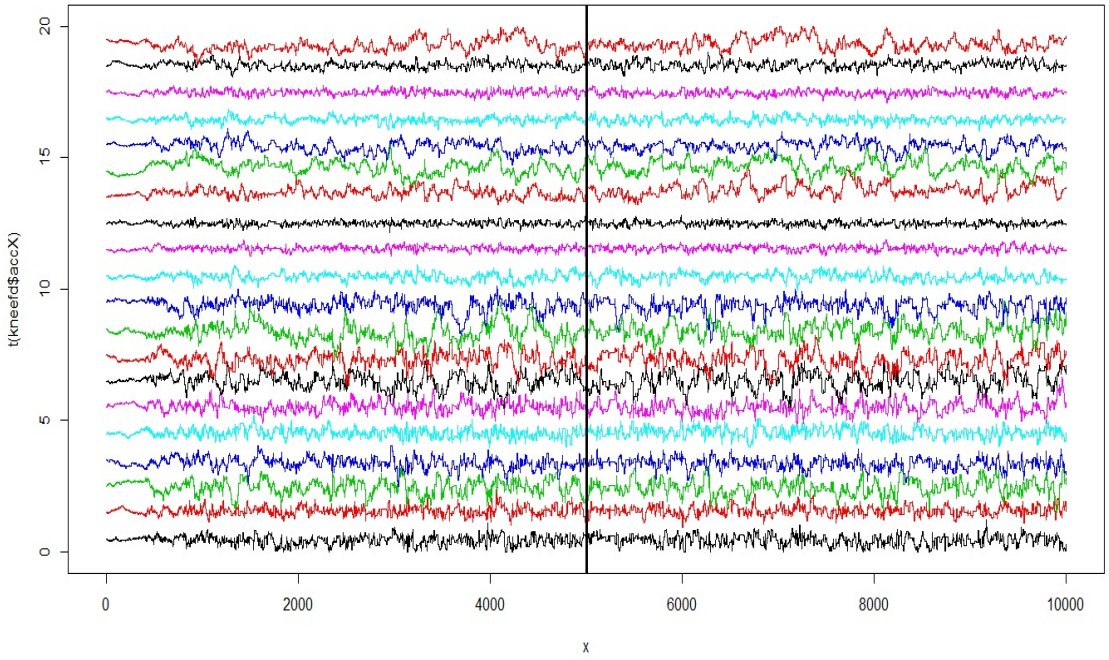

Figure 4.9: Samples of Unobservable Gait times - Knee Angle of Child 2:

Figure represent the sampling of unobservable gait times of the knee angle of child 2 using M-H algorithm. The vertical line at iteration 5000 indicate the stoppage for adapting the M-H algorithm for variance adjustment and also the point for burn-in. The samples prior to vertical axis were discarded as burn-in and the samples after vertical axis were only considered in the estimation

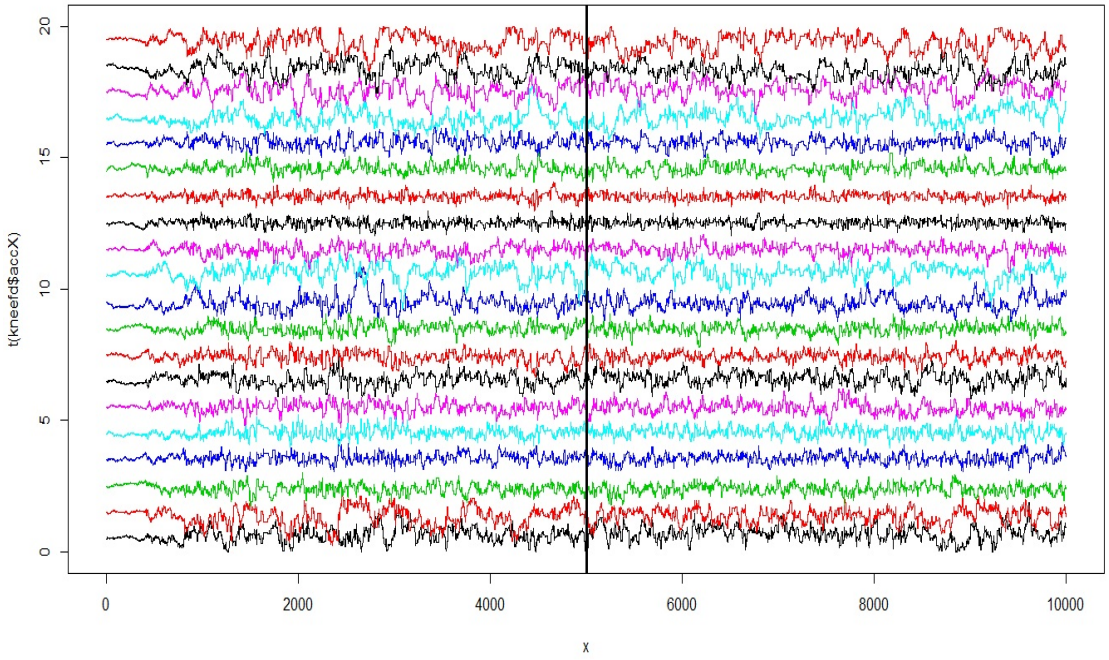

Figure 4.10: Samples of Unobservable Gait times - Hip Angle of Child 2:

Figure represent the sampling of unobservable gait times of hip angle of child 2 using M-H algorithm. The vertical line at iteration 5000 indicate the stoppage for adapting the M-H algorithm for variance adjustment and also the point for burn-in. The samples prior to vertical axis were discarded as burn-in and the samples after vertical axis were only considered in the estimation

Figure 4.5 and 4.6 indicate that the methodology in chapter 3.1 succeeds in providing a reasonable approximation for the true curves, the knee and hip angles of child 2. Figures 4.7 and 4.8 ensure that sampling from full conditionals are stable. we assessed the fit of the approximated curves via MSE plots introducing different measurement error on true data.

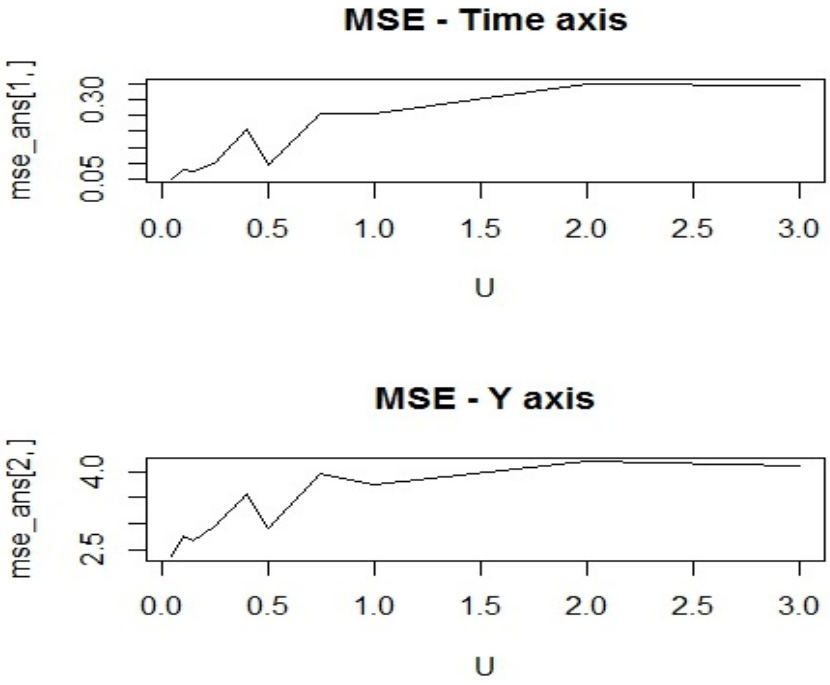

Figure 4.11: MSE calculation for Knee Angle of Child 2:

Top panel of the figure provides the MSE calculation for approximated Xs for different values of σ_U. Similarly, bottom panel provides the MSE calculation for approximated Rs for different values of σ_U. both MSE curves suggest that method suggest in chapter 3.1 succeeds in providing an approximation as MSE tend to increase with more noise on data.

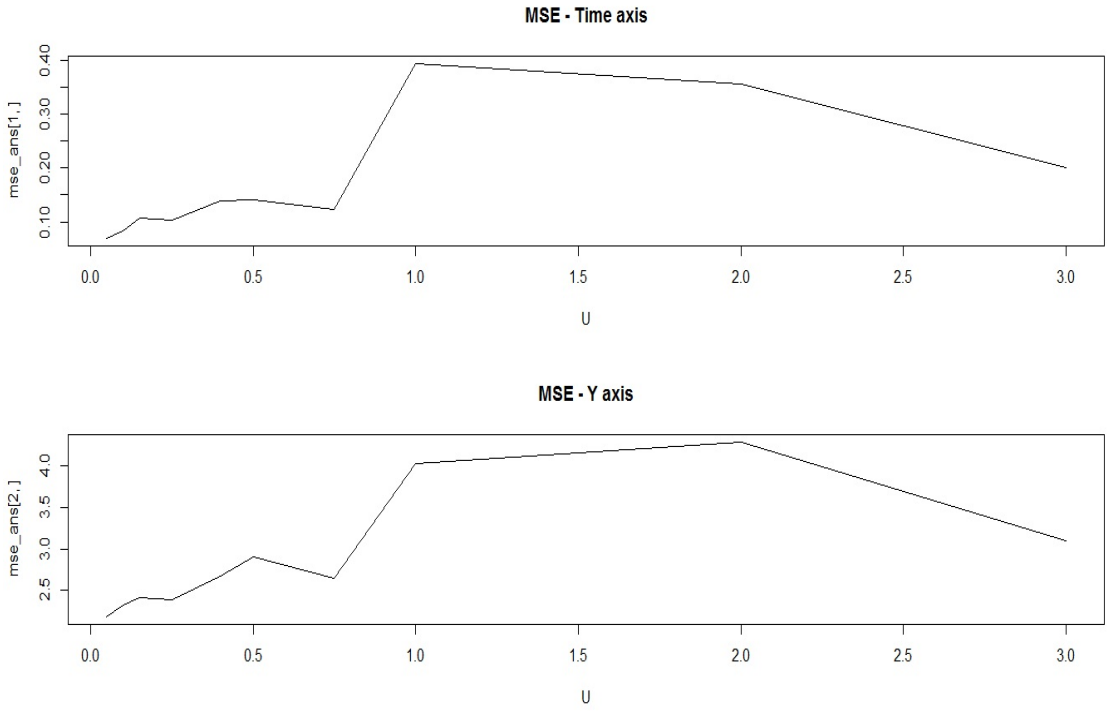

Figure 4.12: MSE calculation for Hip Angle of Child 2:

Top panel of the figure provides the MSE calculation for the approximated Xs for different values of σ_U. Similarly, bottom panel provides the MSE calculation for the approximated Rs for different values of σ_U. By looking at both panels it can be seen that MSE tend to decrease with the increase in measurement error after an increase. However ideally this should increase if the approximation works well. The reason for this decrease in MSE could be a fact due to less variability in true hip angle

The same procedure was used in approximating the Knee and the Hip angles of rest of the 38 individuals and these results were used in regressing Hip angle on Knee angle using the functional regression methodology. The model of interest in functional regression analysis can be given as follows,

$$y_i(t) = \omega_0(t) + \sum_{j=1}^{q-1} x_{ij}(t)\omega_j(t) + \epsilon_i(t)$$

We define $y_i(t)$, a functional vector of length N and it represents our response function knee angle. $x_{ij}(t)$ is the functional predictor and it represents hip angle. The parameters that need to be estimated are functions. We define $\omega_0(t)$ and $\omega_j(t)$ as the functional parameters to be estimated. $\omega_0(t)$

represents intercept function and $\omega_j(t)$ represents hip angle coefficient function. The estimation of these functional parameters were carried out using an iterative approach. At each iteration 39 curves of hip angles and knee angles were approximated using methodology suggested in chapter 3.1. These approximated curves from the knee and the hip angles were regressed at each iteration. The results of the functional regression analysis are provided in figure 4.13.

Figure 4.13 provides estimated intercept function and the estimated hip regression coefficient function with their uncertainty in terms of 90% confidence intervals. From which we could observe that more hip bend results in more knee bend. Here both red curves on figure 4.13 were obtained by regressing 39 mean curves of the hip angles and 39 mean curves of the knee angles. The point wise confidence intervals were obtained by taking the quantiles, regressing 39 curves approximated at each iteration during the MCMC sampling process.

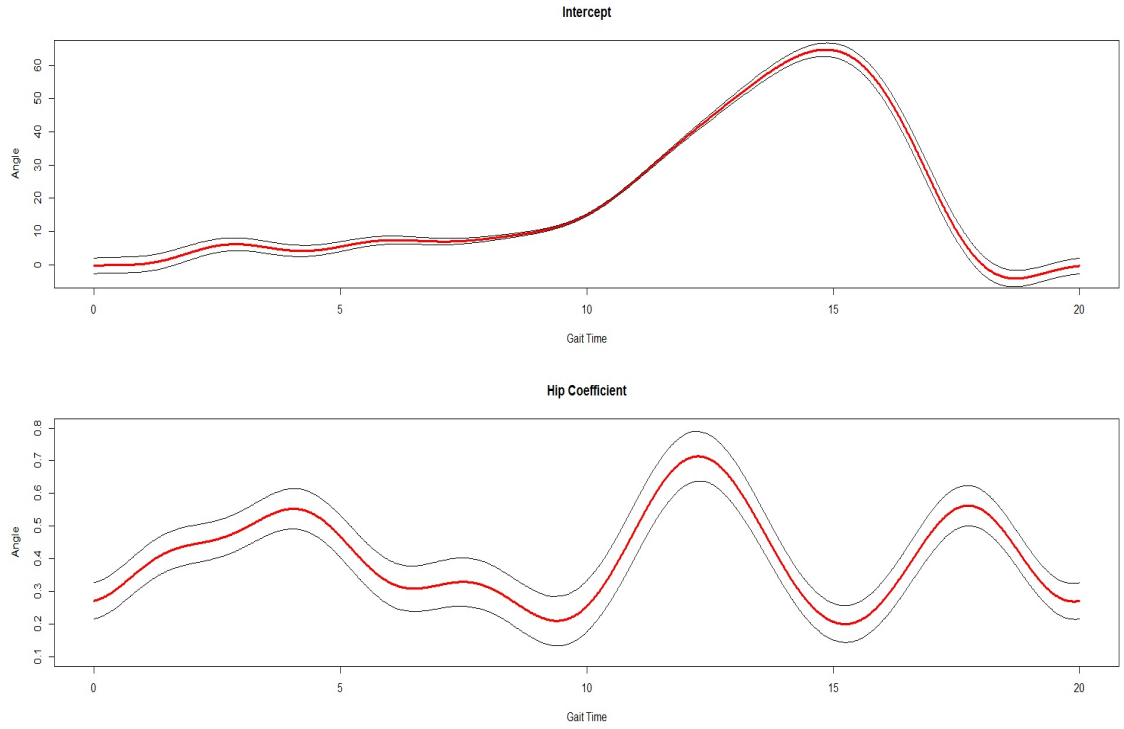

Figure 4.13: Intercept and Hip Regression Coefficient with their 90% confidence intervals:

Figure provides the estimated intercept function (top panel) and the estimated hip regression coefficient function (bottom panel) for the Gait cycle with 90% point wise confidence intervals. Here the red line represent the mean curve from latter 5000 samples ignoring first 5000 samples as burn-in. The 90% point wise confidence intervals were obtained by taking respective quantiles for each point considering their respective sample paths

The study moves into windowed moving correlations between the Knee angle data and the hip angle data of child 2 to explore the methodology suggested in chapter 3.2. Figure 4.14 provides moving correlations between Knee angle data and Hip angle data of child 2 for a window size of 25%. The Knee angle and the Hip angle data were evaluated on a fine grid of 1000 points with its respective estimated basis coefficients to obtain two vectors of equal length. The point wise confidence intervals were calculated by taking quantiles, of windowed moving correlations, which were calculated iteratively for each point considering their sample paths.

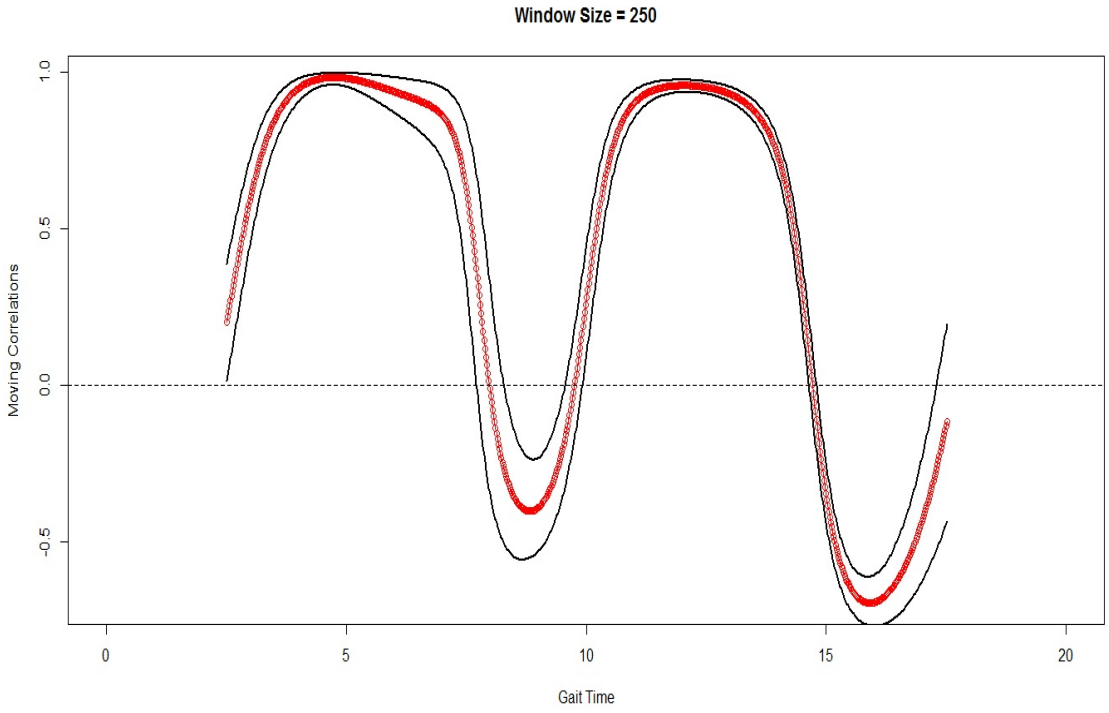

Figure 4.14: Windowed Moving Correlations between Knee angle and Hip angle data:

Figure provides windowed moving correlations calculated at a window size of 25% when two data sets have the same length of 1000 data points. The two Black lines indicate the 90% point wise confidence intervals.

If compare the two figures 4.13 and 4.14, whenever the hip coefficient has a upward trend in figure 4.13, the corresponding moving correlations have a positive correlation in figure 4.14 and vice versa. Not only that, windowed moving correlations plot confirm that there are significant correlations between knee and hip angle through out its range. It confirms the fact that both functional regression or moving correlations could be used to identify the dependency between two functions. Furthermore, moving correlations will be extremely useful above functional regression, when we cannot distinguish the two functions in terms of predictor and response functions.

4.3 Analysis of Climate Change Data

This section discusses the analysis on time series data which created the motivation to carry out this project. The project takes into account two irregularly spaced time series data sets which were subject to time scale measurement error. The two time series data sets are, Oxygen isotope ($\delta^{18}O$)

measurements that were taken from stalagmite samples from the Yok Balum cave and Titanium concentrations which were taken from marine sediments in the Cariaco Basin. The primary objective with these two data sets is to identify the dependency between them from a statistical view point as [1] only rely on graphical interpretation.

To overcome this challenge we first model irregularly spaced time series data sets using the methodology suggested in chapter 3.1 and then we find the dependency between the two data sets using windowed moving correlations.

4.3.1 Approximating a Curve for Oxygen Isotope Data

The approximation was carried out based on 1440 data points collected over a period of 2000 years. Oxygen Isotope measurements had a range from 0 to 1. The point wise errors on the Oxygen Isotope measurement (σ_ϵ) was given. From a programming perspective we were aware of σ_ϵ, thus we no longer need to sample σ_ϵ during the approximation process. The time scale had a range from 300CE to 1750CE and it has a measurement error from ± 1 to ± 17. Again from a programming perspective we had to divide all time values by 2000 including measurement errors to overcome numerical complexities in the program. Instead of sampling σ_U, we used $\frac{8.5}{2000}$ as our σ_U in the approximation. Sampling was carried using 25,000 iterations, removing first 10,000 samples as burn-in. This project uses a truncated polynomial basis with 20 basis functions (p=4(degree of polynomial), k=15(number of knots)) in this approximation. The prior distributions used were as follows,

- $\gamma \sim G(1, 10^{-3})$
- $\mu_x \sim N(0.5, 1)$
- $\sigma_x^2 \sim IG(0.5, 1)$

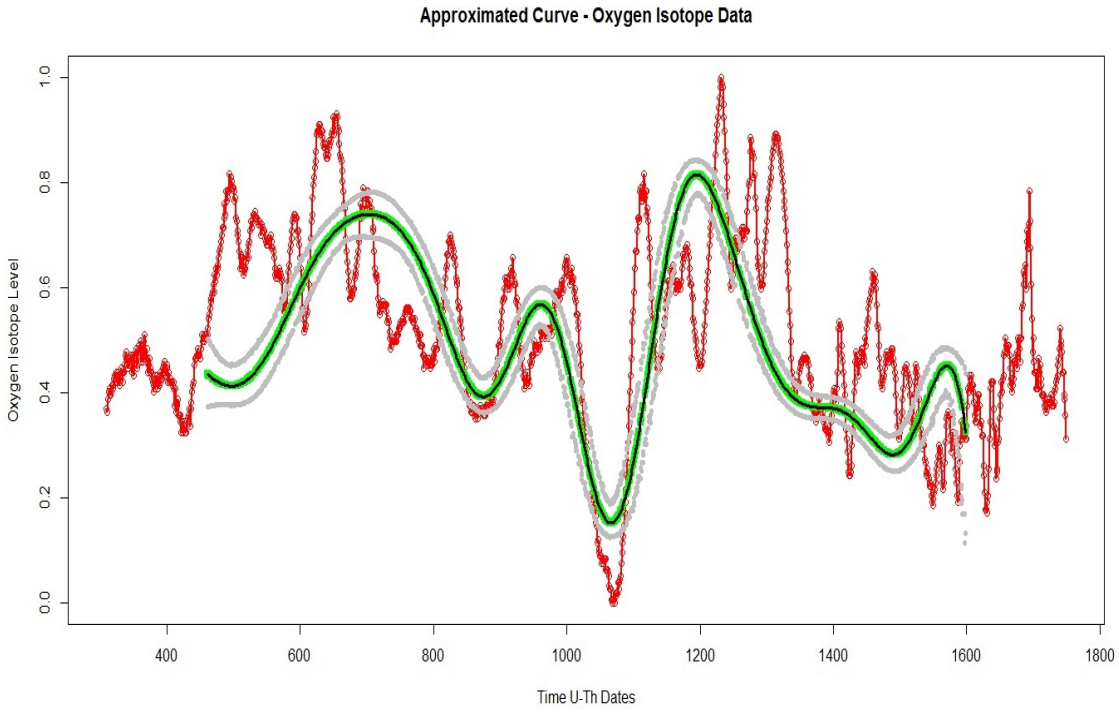

Figure 4.15: Approximated Curve for Noisy Oxygen Isotope Data:

Figure provides a approximated curve for Oxygen Isotope data. The noisy data is given in "Red". "Green" points are the point wise estimates for Oxygen Isotope data eliminating measurement error. The "Black" curve is the curve obtained by evaluating the estimated coefficients on a fine grid. The approximation was carried out using posterior means. Estimation uncertainty is indicated by the two Gray lines which are 90% point wise confidence intervals.

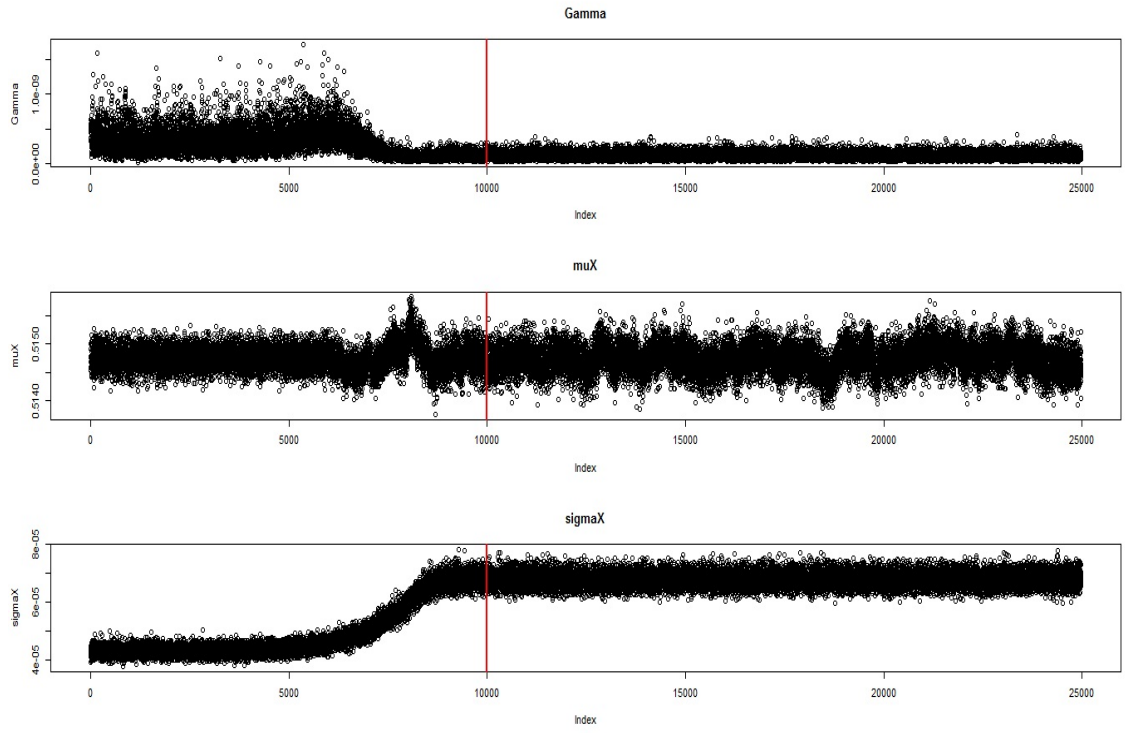

Figure 4.16: Trace Plots - Full Conditional Distributions from Gibbs Sampling:

Three plots on the figure represent each of the full conditional distributions sampled during the MCMC iterations. The vertical lines on each of the three figures give an idea on burn-in. For estimation we used the samples obtained beyond the vertical line as previous samples were discarded as burn-in.

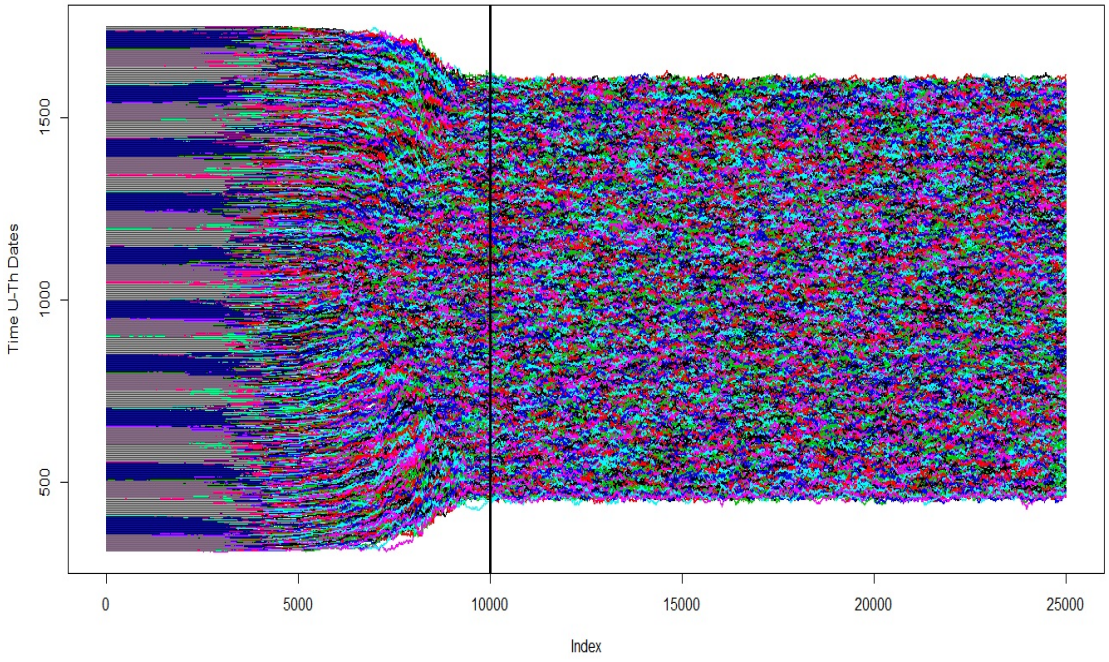

Figure 4.17: Samples of Unobservable time ”t” from Metropolis Hastings algorithm:

Figure represents the sampling of noisy covariate(time points) using the M-H algorithm. The vertical line at iteration 10000 indicate the stoppage for adapting the M-H algorithm for variance adjustment and also the point for burn-in. The samples prior to vertical axis were discarded as burn-in and the samples after vertical axis were only considered for estimation purposes

Figure 4.15 indicates that, approximated curve for Oxygen Isotope data provides a reasonable fit. However observing same plot we can identify that the mean curve fails in approximating two extremes of the noisy data. The reason for this drawback can be seen in figure 4.17. The sampled Xs from Metropolis Hastings algorithms tend to converge to the middle in figure 4.17. In other words Xs attempt to converge to a single μ_x value. This is a drawback in [10] as the model in [10] attempts to sample Xs against a single μ_x at each iteration for a sample size of 1440 data points. This makes sampled values to pull toward to the middle of the plot as seen in figure 4.17. This adversely results in final approximation as seen in figure 4.15.

To overcome this issue, this project slightly modifies the model in [10] by introducing a fixed value for σ_x rather than sampling σ_x from its full conditional distribution. The objective is to minimize the effect of μ_x in the model so that Xs will not pull towards to the middle of the data range. In other

words it is recommended to use a very informative prior for σ_x. Therefore, this project uses the following prior distributions.

- $\gamma \sim G(1, 10^{-3})$
- $\mu_x \sim N(0.5, 1000)$
- $\sigma_x^2 = 0.0005$

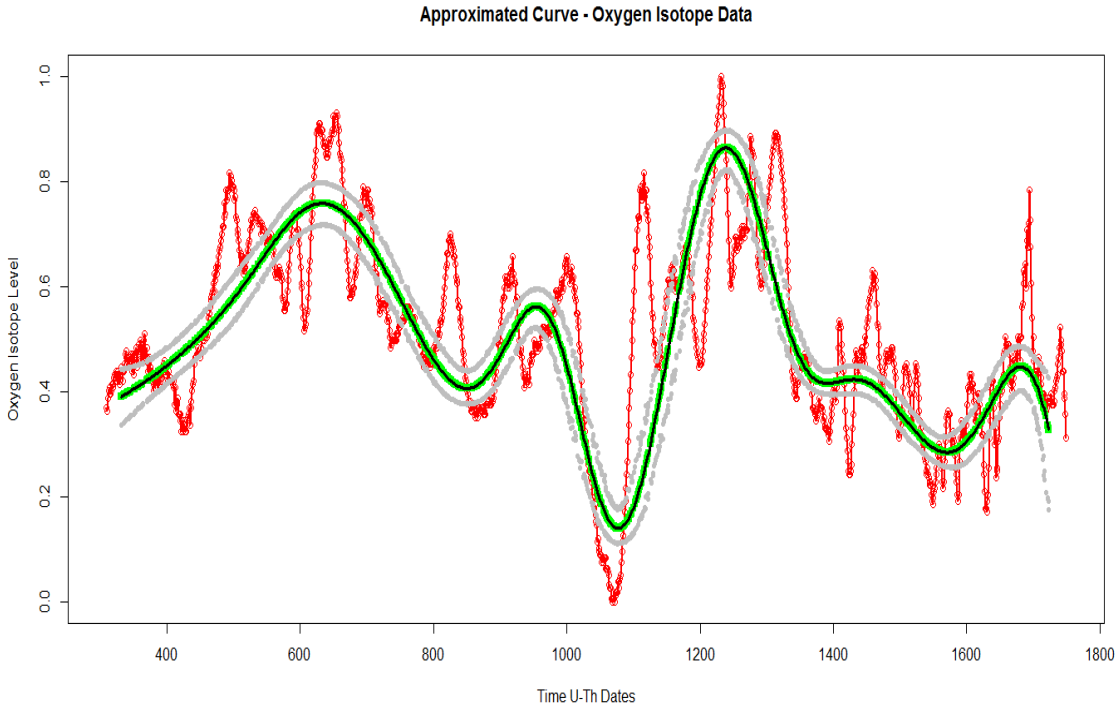

Figure 4.18: Approximated Curve for Noisy Oxygen Isotope Data:

Figure provides a approximated curve for Oxygen Isotope data. The noisy data is given in "Red". "Green" points are the point wise estimates for Oxygen Isotope data eliminating measurement error. The "Black" curve is the curve obtained by evaluating the estimated coefficients on a fine grid. The approximation was carried out using posterior means. Estimation uncertainty is indicated by the two Gray lines which are 90% point wise confidence intervals.

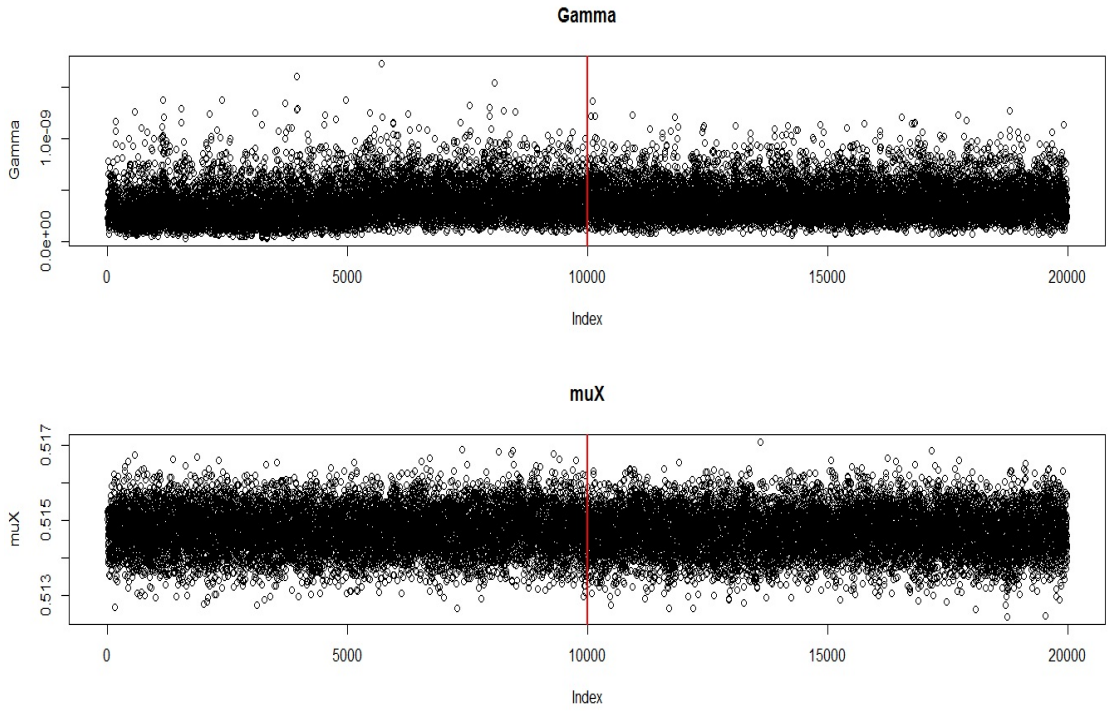

Figure 4.19: Trace Plots - Full Conditional Distributions from Gibbs Sampling:

Three plots on the figure represent each of the full conditional distributions sampled during the MCMC iterations. Here the vertical lines on each of the three figures give an idea on burn-in. For estimation we use the samples obtained beyond the vertical line as previous samples were discarded as burn-in.

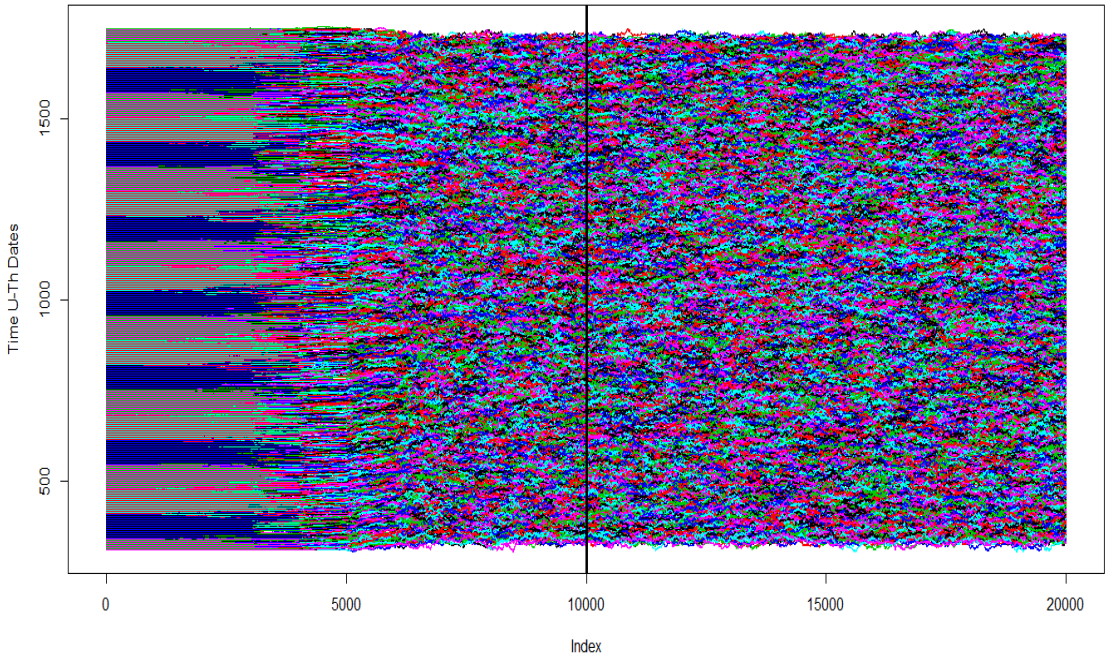

Figure 4.20: Samples of Unobservable time ”t” from Metropolis Hastings algorithm:

Figure represent the sampling of noisy covariate(time points) using the M-H algorithm. The vertical line at iteration 10000 indicate the stoppage for adapting the M-H algorithm for variance adjustment and also the point for burn-in. The samples prior to vertical axis were discarded as burn-in and the samples after vertical axis were only considered for estimation purposes

Figure 4.18, indicates that the slight alteration to the model in [10] succeeds in providing an approximation for the entire range of the data. Even the sampled values from Metropolis Hastings algorithm as shown in figure 4.20 indicate that they are no longer pulled to the middle of the plot, which was the case in figure 4.17. The figure 4.19 and figure 4.20 confirm that all full conditional distributions are stable, hence our posterior is obtained. 90% posterior confidence intervals given on figure 4.18 in Gray does not capture most of the variations in noisy Oxygen data. However we do not wish to capture all of the variation within our confidence intervals as our primary objective is to find a smooth underlying process for noisy data.

4.3.2 Approximating a Curve for Titanium Concentration Data

The approximation of Titanium data will be carried out based on two time series data sets each having 264 data points collected over a period of 2000 years. The time scale had a range from 300CE to 1750CE. Both error on time measurements and error in Titanium levels were not given. Hence σ_ϵ and σ_U were sampled during the MCMC iterations. Similar to Oxygen Isotope approximation, all time values were divided by 2000 for numerical reasons and sampling was carried out for 25,000 iterations, removing first 10,000 as burn-in. For Titanium data a truncated polynomial basis with 45 basis functions (p=4(degree of polynomial), k=40(number of knots)) was used in the approximation. The prior distributions used were as follows,

- $\sigma_\epsilon^2 \sim IG(1, 0.5)$
- $\sigma_U^2 \sim IG(1, 10^{-3})$
- $\gamma \sim G(1, 10^{-10})$
- $\mu \sim N(0.5, 1)$
- $\sigma^2 \sim IG(0.5, 1)$

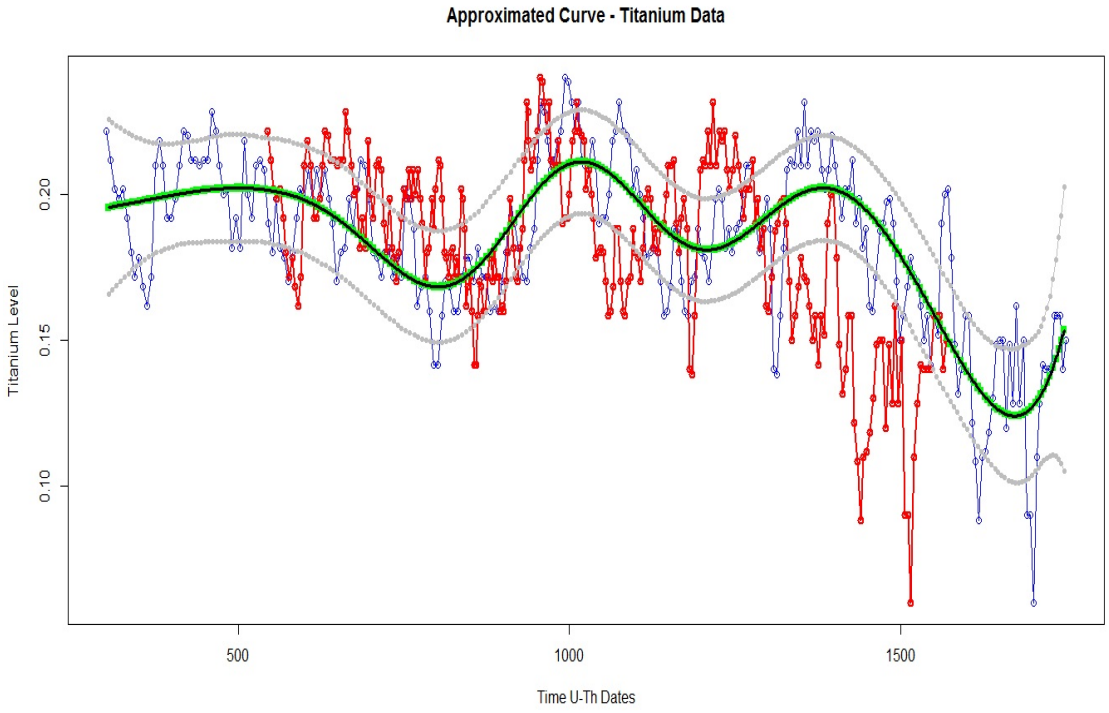

Figure 4.21: Approximated Curve for Noisy Titanium Data:

Figure provides a approximated curve for Titanium data. The noisy data is given in "Red". "Green" points are the point wise estimates for Titanium data. The "Black" curve is the curve which is obtained by evaluating the estimated coefficients on a fine grid. The approximation was carried out using posterior means. Estimation uncertainty is indicated by the two "Gray" lines which are 90% point wise confidence intervals.

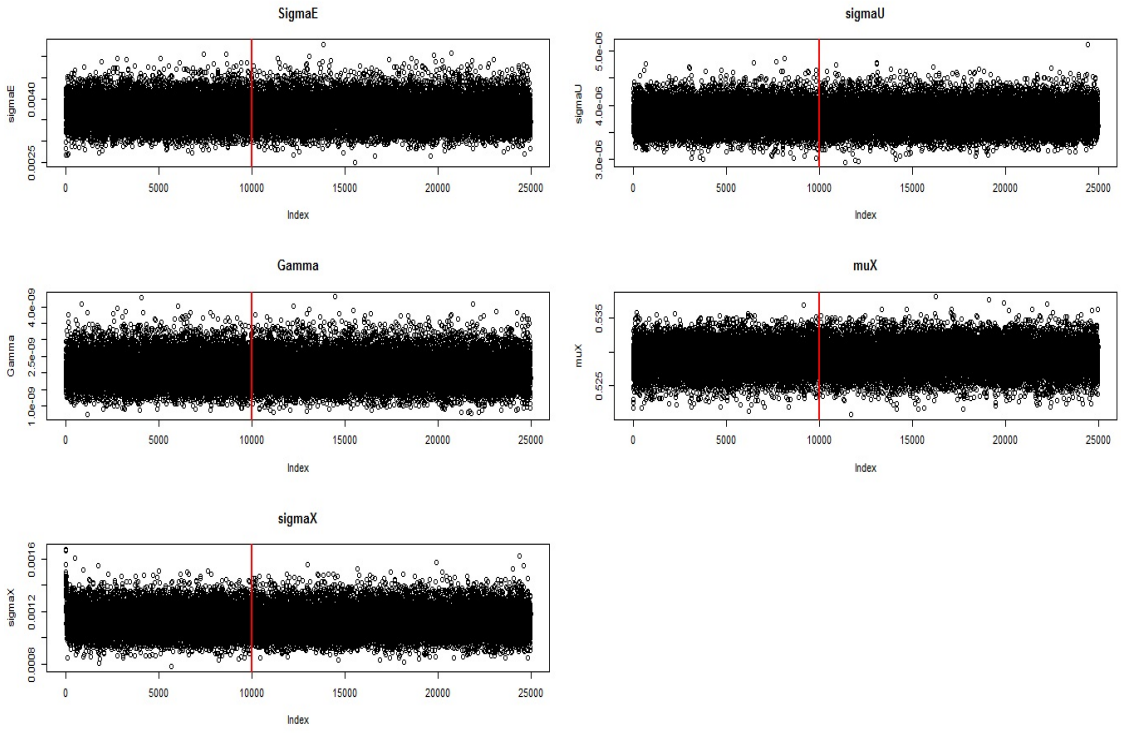

Figure 4.22: Trace Plots - Full Conditional Distributions from Gibbs Sampling:

Five plots on the figure represent each of the full conditional distributions sampled during the MCMC iterations. The vertical lines on each of the five figures give an idea on burn-in. For estimation we use the samples obtained beyond the vertical line as previous samples were discarded as burn-in.

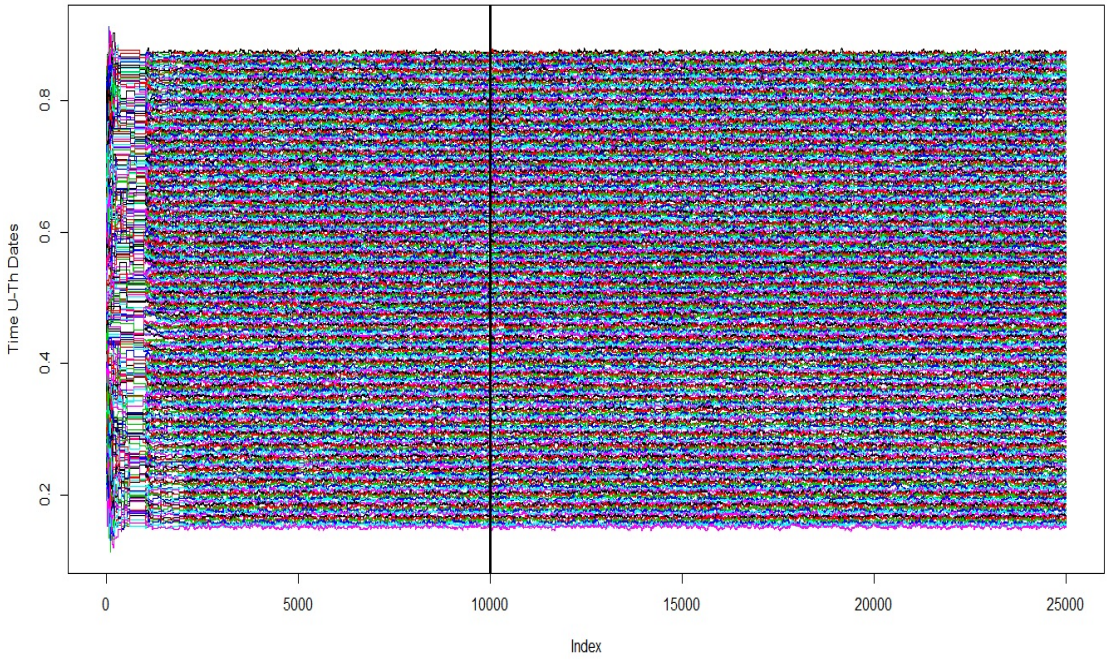

Figure 4.23: Samples of Unobservable time ”t” from Metropolis Hastings algorithm:

Figure represent the sampling of noisy covariate(time points) using the M-H algorithm. The vertical line at iteration 10,000 indicate the stoppage for adapting the M-H algorithm for variance adjustment and also the point for burn-in. Thus the samples prior to vertical axis were discarded as burn-in and the samples after vertical axis were only considered for estimation purposes

Figure 4.21 approximates noisy Titanium data and it indicates that the approximated curve provides a reasonable fit even though it does not capture all the variations in the noisy data. At the same time it should be mentioned that we do not wish to capture all the variation as the data comes with error. We only look for a reasonable fit as our main objective is to identify whether the two data sets correlate or not. Similar to Oxygen data Figures 4.22 and 4.23 indicate that, 25,000 samples obtained from each of the full conditional distributions are adequate to obtain our desired posterior distribution to approximate the Titanium data. The 90% posterior confidence intervals given on figure 4.25 in Gray indicate that they capture large proportion of data within them.

4.3.3 Identification of Dependency between Oxygen Isotope and Titanium Data

This section discusses the dependency between the two time series data sets, Oxygen Isotope and Titanium data using the approximated curves in chapters 4.3.1 and 4.3.2. The concept of windowed moving correlations which was discussed in chapter 3.2 will be used for this purpose. Windowed moving correlations between the two approximated curves were calculated at window sizes of 10%, 25% and 50%. The estimated basis coefficients from Oxygen Isotope and Titanium data were obtained and evaluated them on a fine grid of 1000 point to obtain two vectors of same length. In order to asses significance of the windowed correlations the posterior correlations was obtained. For each sample of smooth functions for Oxygen and Titanium obtained above, the windowed correlations were obtained. The highest density interval estimates for the correlation distributions were then obtained. Results of the windowed moving correlations are provided in the following figures.

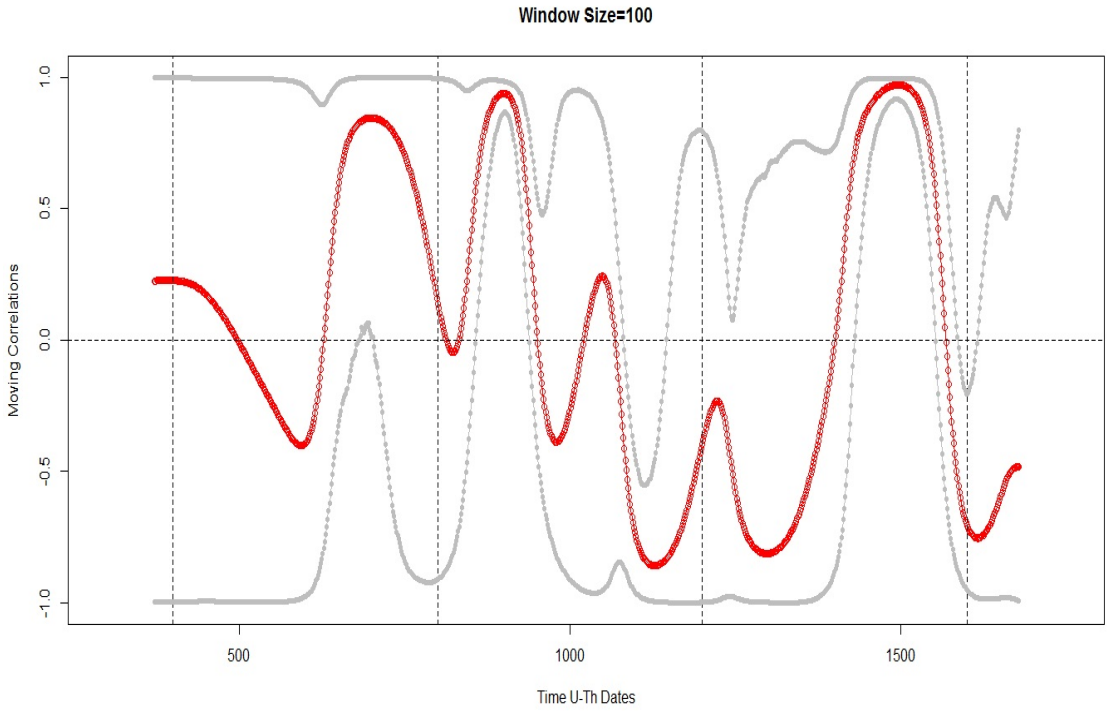

Figure 4.24: Windowed Moving Correlations between Oxygen Isotope and Titanium data:

Figure provides the moving correlations calculated at a window size of 10%. The vertical lines on the figure represent years 400, 800, 1200 and 1600 CE. The horizontal line represent the zero correlation between the two data sets. The Gray lines indicate the 90% point wise confidence intervals of moving correlations.

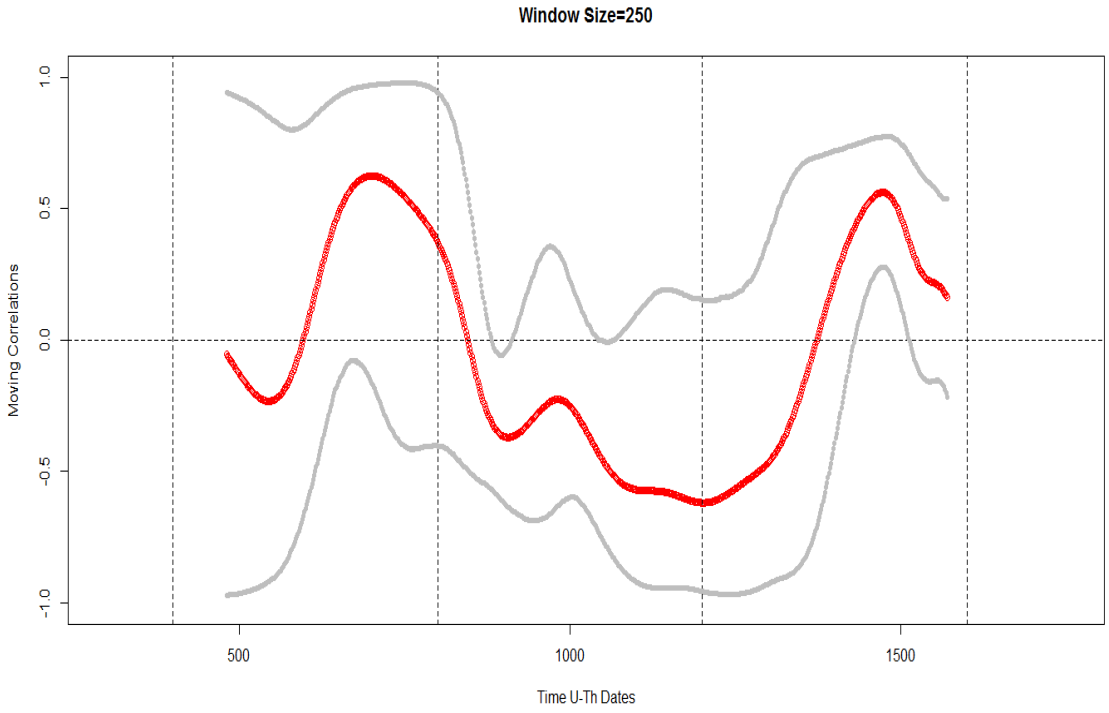

Figure 4.25: Windowed Moving Correlations between Oxygen Isotope and Titanium data:

Figure provides the moving correlations calculated at a window size of 25%. The vertical lines on the figure represent years 400, 800, 1200 and 1600 CE. The horizontal line represent the zero correlation between the two data sets. The Gray lines indicate the 90% point wise confidence intervals of moving correlations.

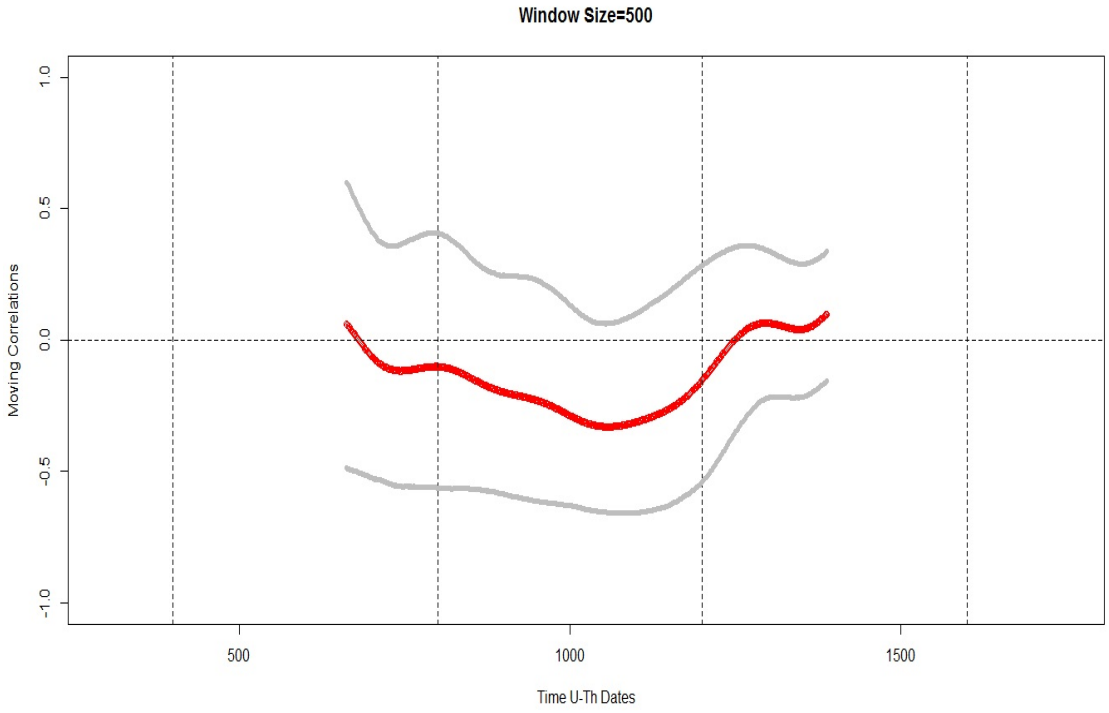

Figure 4.26: Windowed Moving Correlations between Oxygen Isotope and Titanium data:

Figure provides the moving correlations calculated at a window size of 50%. The vertical lines on the figure represent years 400, 800, 1200 and 1600 CE. The horizontal line represent the zero correlation between the two data sets. The Gray lines indicate the 90% point wise confidence intervals of moving correlations.

Figures 4.24, 4.25 and 4.26 indicate that most of the point wise moving correlations are close to zero. This means that the chance of having any association between Oxygen Isotope data and Titanium data is much less. Our primary goal is to identify whether there are any significant correlations between the two data sets or not. We could get an idea on significance by looking at the 90% point wise confidence intervals and hence we can observe that a large proportion of point wise moving correlations are insignificant as point wise confidence intervals have value zero throughout the range. Therefore, we can conclude that these two time series data sets do not correlate each other.

In [1], the two data samples (Oxygen Isotope and Titanium) were obtained from two different geographic locations. By looking at three moving correlation plots with non-significant correlations, what we can say is, there could be some geographical factors that could have had an impact on

rainfall. In other words, such geographical factors may have resulted in non-significant correlations which were observed in figures 4.24, 4.25 and 4.26.

As a final remark, it should be mentioned that the methodology suggested in this project could be used to find statistically significant correlations between two data sets which are irregularly spaced and subjected to measurement errors even though these specific data sets resulted in non-significant correlations.

Chapter 5

Further Improvements to the Study

This project is a methodological development and during this development we can identify three key stages. They are developing a measurement error model to incorporate the measurement error of the time scale, modelling irregularly spaced time series data via regressions P-splines and Bayesian sampling mechanism and identification of dependency between two data sets either via functional regression or using windowed moving correlations.

The first limitation of the study can be found at the development of measurement error model for time series data. If we look at our data in chapter 4.3, we can identify that they are time dependent measurements because they come from sediment cores. Thus, there is an important ordering of points. This ordering was ignored during the estimation process and due to error imposed in time scale (σ_U), the time points could move around ignoring the order of the time points. This is an area that we need to rectify during the estimation process in a potential future study and one possible remedy is to have a small error (σ_U) on top of time points in our measurement error model. Alternatively we could use a likelihood function with an indicator or a Dirichlet sorting process.

The second limitation can be observed at the approximation of Oxygen Isotope data. As seen in figures 4.15 and 4.17 it fails to give an approximation for the full range of Oxygen data. A possible remedy for this could be to have an observation specific μ_x (prior), which makes Xs evenly spaced rather than pulling them to a single value.

Bibliography

[1] Kennett D. J., Breitenbach S. F. M., Aquino V. V., Asmerom Y., Awe J., Baldini J. U. L., Bartlein P., Culleton B. J., Ebert C., Jazwa C., Macri M. J., Marwan N., Polyak V., Prufer K. M., Ridley H. E., Sodemann H., Winterhalder B., and Haug G. H. Development and disintegration of maya political systems in response to climate change. *Science*, 338:788–791, 2012. 4, 5, 33, 47

[2] Erdogan E., Ma S., Beygelzimer A., and Rish I. *Statistical Models for Unequally Spaced Time Series*, chapter 74, pages 626–630. 2

[3] Ramsay J., Hooker G., and Graves S. *Functional Data Analysis with R and Matlab*. Number 2009928040 in Use R! Springer Science+Business Media, LLC, 233 Spring Street, New York, NY, 10013, USA, 2009. 21

[4] Zhang J., He W., and Li H. A semi-parametric approach for accelerated failure time models with covariates subject to measurement error. *Computational Linguistics*, 2012. 7

[5] Schulz M. and Mudelsee M. Redfit: Estimating red-noise spectra directly from unevenly spaced paleoclimatic time series. *Computer Geosciences*, 28:421–426, 2001. 2

[6] Steele R., Platt R., and Ross M. Modelling birthweight in the presence of gestational age measurement error - a semi-parametric multiple imputation model. 6

[7] Caroll R. J., Roeder K., and Wasserman L. Flexible parametric measurement error models. *Biometrics*, 55:44–55, 1999. 6

[8] Maller R.A., Muller G., and SZIMAYER G. Garch modelling in continuous time for irregularly spaced time series data. *Bernoulli*, 14:519–542, 2008. 3

[9] Hossain S. and Gustafson P. Bayesian adjustment for covariate measurement errors: A flexible parametric approach. *Statistics in Medicine*, 28:1580–1600, 2009. 6

[10] Berry S.M., Carroll R.J., and Ruppert D. Bayesian smoothing and regression splines for measurement error problems. *Journal of the American Statistical Association*, 97:160–169, 2002. 7, 8, 10, 11, 12, 14, 21, 36, 39