Inference in Discrete High Dimensional Space: An Exploration of the Earth's Ice Sheets through Change Point and Variable Selection Techniques

By Eric Ruggieri

B.A. Providence College 2005

Sc.M. Brown University, 2006

A Dissertation Submitted in Partial Fulfillment of the

Requirements for the Degree of Doctor of Philosophy

in the Division of Applied Mathematics at Brown University

Providence, Rhode Island

May 2010

This dissertation by Eric Ruggieri is accepted in its present form

by the Division of Applied Mathematics as satisfying the

dissertation requirement for the degree of Doctor of Philosophy.

_____	_____
Date	Charles Lawrence, Director

Recommended to the Graduate Council

_____	_____
Date	William Thompson, Reader
_____	_____
Date	Steven Clemens, Reader

Approved by the Graduate Council

_____	_____
Date	Sheila Bonde, Dean of the Graduate School

The Vita of Eric Ruggieri

Born December 6, 1982 in Endicott, New York

Education

- Sc.M. Applied Mathematics, Brown University, May 2006
- B.A. Mathematics and Computer Science, Providence College, May 2005, *Summa Cum Laude*

Research Interests

- Statistical Inference in Discrete High Dimensional Spaces, Variable Selection Techniques, Statistical Modeling
- Geological Time Series Analysis
- Computational Biology – specifically Disease Association Studies, Phasing Algorithms, and Measures of Linkage Disequilibrium

Publications

- E. Ruggieri, T. Herbert, K. Lawrence, and C.E. Lawrence. Change Point method for detecting regime shifts in Paleoclimatic time series: Application to $\delta^{18}O$ time series of the Plio-Pleistocene. *Paleoceanography,* **24:** PA1204, 2009. doi:10.1029/2007PA001568
- E. Ruggieri and S. Schreiber. The Dynamics of the Schoener-Polt-Holis Model of Intra-Guild Predation. *Math. Biosci. Eng.*, **2**(2), 2005.

Published Software

- *ChangePoint* –Programmed and developed in Matlab with Graphical User Interface, 2009.

Presentations and Invited Talks

- *A Mathematical Exploration of the Earth's Glacial System,* Pi Mu Epsilon Induction Ceremony, Providence College, Providence, RI, April 2010.

- *The Frequency of Glacial Events,* University of Massachusetts at Amherst, Amherst, MA, February 2010
- *The Frequency of Glacial Events,* Kenyon College, Gambier, OH, February 2010
- *The Frequency of Glacial Events,* Niagara University, Niagara, NY, February 2010
- *The Frequency of Glacial Events*, Duquesne University, Pittsburgh, PA, February 2010
- *The Mid-Pleistocene Transition: From Forcing to Pacing of Ice Sheets,* AMS Session on Probability and Statistics, Joint Mathematics Meetings, San Francisco, CA, January 2010
- *The Mnemonics of Mathematics,* presented at the 6th annual New England Peer Tutoring Association Meeting, Bryant University, Smithfield, RI, April 2005
- *Effective Use of the Algorithms in Action Website,* presented at 8th Annual Consortium for Computing Sciences in Colleges Northeastern Conference (CCSCNE), CCRI, Providence, RI, April 2003

Academic Experience

Providence College

Special Lecturer of Mathematics - *Calculus and Analytical Geometry II,* September 2008-May 2010

Brown University

Special Lecturer of Applied Mathematics, *Information Theory*, September-December 2008.

Guest Lecturer

- *Statistical Inference*, November 2009
- *Information Theory*, October 2007, October-November 2009
- *Inference in Genomics and Molecular Biology*, February 2008, September 2008, March 2009

Teaching Assistant

- *Essential Statistics*, January-May 2007

- *Statistical Inference*, September-December 2006
- *Topics in Chaotic Dynamics*, January-May 2006

Professional Development

Sheridan Center for Teaching and Learning Certificate Program

- I - Teaching Seminar (2006)
- II – Classroom Tools Seminar (2008)
- III- Developing a Teaching Portfolio (2009)

Honors and Awards

- Brown University Dissertation Fellowship, September 2009 – May 2010
- Nominated for Presidential Award for Excellence in Teaching, May 2009
- Brown University Fellowship, September 2005 – May 2006

To my wife, Michelle, and my family, who never stopped believing in me

Acknowledgements

There are many people who have helped me get to where I am today. In my five years at Brown University, I was blessed to know and work with not only faculty and staff of the Division of Applied Mathematics, but the Center for Computational Molecular Biology and the Department of Geological Sciences as well. Your support has been invaluable to me. In particular, I would like to thank the following people:

First and foremost, I would like to thank my advisor, Chip Lawrence, for his incredible guidance over the last five years. He took me under his wing during my first year and together we have explored ice sheet dynamics, a topic foreign to both of us at the start of this project. Difficulties arose, but he was never short on ideas for how to move forward. I want to thank Chip for his patience and understanding and for all of the opportunities that he has provided me. I could not have done this without him.

I want to thank the members of my Lab group whom I worked with so closely during my time here at Brown. Each week, this Computational Biology group was thoughtful and offered suggestions for my Geology problem, while at the same time keeping me current on some of the research in our field. Most especially, I want to thank Luis Carvahlo, who provided much advice over this final year and Bill Thompson, who was always available to help me with computer issues and who agreed to be a reader of this dissertation.

I also want to thank Tim Herbert, Kira Lawrence, and Steve Clemens, our colleagues in the Department of Geological Sciences. Tim and Kira introduced Chip and I to the ice volume time series that became the foundation of this dissertation. They have been a great resource for the multitude of questions we had about the Geological background to the waxing and waning of ice sheets. Steve and Jim Russell welcomed me into their graduate seminar during my third year at Brown. It was the first Geology course I had ever taken and gave me a huge boost in confidence when talking about Geological

topics. Additionally, I want to thank Steve who graciously agreed to be a reader of my dissertation and who has been a great source of insight as this project has moved forward.

My gratitude also goes out to Liam Donohoe, Richard Connelly, and Jim Tattersal at Providence College, who strongly supported me even after I left PC, especially when I came back to teach a section of Calculus. Most importantly, I want to thank my former advisor, Joanna Su. Her advice and guidance over the years has been invaluable. She has been a truly great friend to me.

Last and certainly not least, I would like to thank all of my family and friends, but especially my wife Michelle and my parents, Randy and Cindy Ruggieri, for always believing in me. In my moments of weakness and frustration, your encouragement guided me back on path. Your constant support of me and my schooling has been incredible. Looking back on the years since I left for Providence College, it is pretty amazing how far all of us have come.

Table of Contents:

List of Tables

List of Figures

Abstract of "*Inference in Discrete High Dimensional Space: An Exploration of the Earth's Ice Sheets through Change Point and Variable Selection Techniques*"

by Eric Ruggieri, Ph.D. Brown University, May 2010

Glaciers have been melting and reforming on the Earth for millions of years. Over the last several decades, Geologists have created $\delta^{18}O$ proxy records that measure the amount of ice on the Earth's surface over the last 5 million years. The proxy records provide evidence for two major changes in the Earth's ice sheet dynamics. The first is an increase in the magnitude of glacial events around 2.7 Million years ago (Ma), which coincides with the intensification of glaciations in the Northern Hemisphere; the second, around 1Ma, is known as the Mid-Pleistocene Transition and represents not only a change in the magnitude of glacial events, but their frequency as well. Thus, two of the most important questions that can be asked about the $\delta^{18}O$ proxy record are: 1) When (exactly) have changes occurred? 2) Which mechanisms are operating in each of the different glacial regimes?

This dissertation is concerned with statistical inference in discrete high dimensional space, in particular on the $\delta^{18}O$ proxy record. We begin with the simplest of models, the Least Squares Change Point algorithm, which aims to optimally partition a time series into *k* regimes, fitting each with a different regression model. However, there is no way to characterize the uncertainty surrounding the optimal solution. To deal with this limitation, we introduce the Bayesian Change Point algorithm, which creates a probabilistic model for the $\delta^{18}O$ record, yielding uncertainty estimates on the number of change points, their locations and the regression parameters. In addition, to address the wide range of mechanisms proposed for glacial dynamics after the Mid-Pleistocene Transition, we introduce a novel variable selection technique (EBMA or Exact Bayesian Model Averaging) that has a smaller time complexity than existing algorithms. By combining EBMA with the Bayesian Change Point algorithm, we produce a highly flexible statistical model that can search through an enormously high dimensional space in a practical amount of time.

Chapter 1

Motivation

1.1 Introduction

Glaciers on the Earth have been melting and reforming for millions of years. A $\delta^{18}O$ proxy record derived from ocean sediment cores allows geologists to quantify the amount of ice on the Earth at a specific time in our past and thus provides a way to study the patterns of glacial growth and destruction through time. Obvious changes in the $\delta^{18}O$ record exist [Figure 1.1].

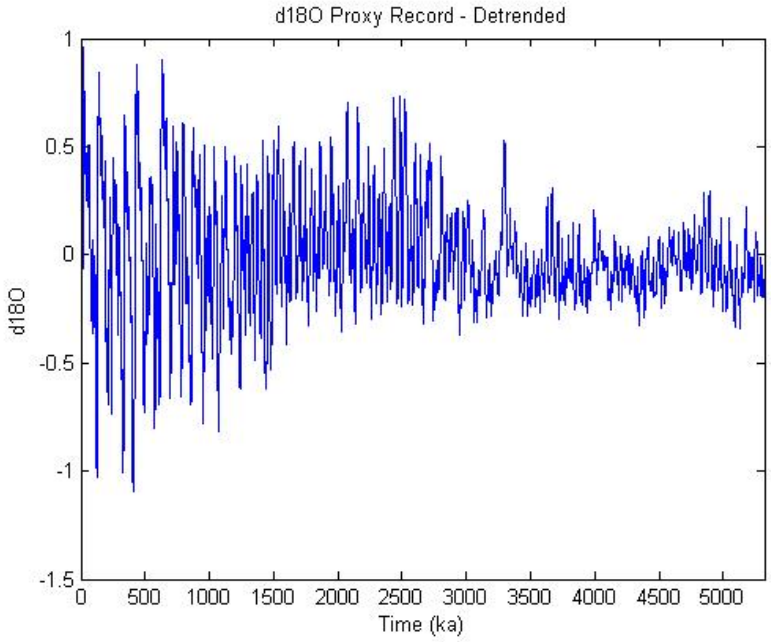

Figure 1.1 The $\delta^{18}O$ proxy record. The $\delta^{18}O$ proxy record [86] after using an exponential function to remove the long-term cooling trend. Present day is on the left and time moves backwards in thousand year steps (ka) as you move right along the x-axis. Larger $\delta^{18}O$ values indicate increased ice sheet cover.

Around 2.7Million Years Ago (Ma), at a time of intensification of glaciers in the Northern Hemisphere, a change in amplitude, or amount of glacial cover, is clearly seen in this data set. A more dramatic change occurred around 1Ma, known as the Mid-Pleistocene Transition. Here,

the change was not only in the magnitude of glacial events, but in their timing, changing from melting events occurring every 41 thousand years (kyr) to every ~100 kyr.

Today, most theories of ice sheet dynamics are a variation of Milankovitch Theory [99] which loosely states that ice sheets respond linearly to the amount of solar insolation (i.e. energy) received at the top of the Earth's atmosphere at 65°N Latitude during the summer. Changes in the amount of solar insolation received by the Earth are caused by variations in the Earth's orbit around the Sun. This motion can be described by three parameters: precession (wobble), obliquity (tilt), and eccentricity (ellipticity) of the Earth's orbit, each of which is nearly periodic. Precession varies at 19 and 23 kyr, obliquity varies primarily at 41 kyr and the eccentricity of the Earth's orbit changes at a triplet of frequencies 95, 124, and 404 which average out to ~100 kyr and 404 kyr. Eccentricity controls the absolute amount of insolation, whereas obliquity and precession control the distribution of insolation on the Earth's surface. Accordingly, solar isolation at any latitude and season can be represented as a function of obliquity and precession [72]. Thus this problem can be thought of in terms of a regression, matching the solar insolation input to the glacial output as represented by the $\delta^{18}O$ record.

At the Mid-Pleistocene Transition, the periodicity of the glacial cycles changes from 41 to 100 kyr. Few dispute obliquity forcing of glaciations prior to the Mid-Pleistocene Transition, but a linear response to Eccentricity forcing after this time has been highly contentious. One of the major drawbacks to the claim of eccentricity forcing is that the direct contribution of orbital eccentricity to solar insolation is too small (<0.1%) to be the force behind the major ice volume changes on the Earth [28, 59, 66, 71, 72, 121 and others]. Whereas variations in Precession and Obliquity can alter the amount of solar insolation received during the summer at 65°N Latitude by 30W/m^2 and 15W/m^2 respectively, variations in Eccentricity alter the solar insolation budget

by less than 1W/m^2 (out of ~500W/m^2). Additionally, questions arose as to why eccentricity forcing was not realized before that time.

To study the Earth's dynamic ice sheets, we introduce both efficient and exact Change Point and Variable Selection algorithms. We will use simplified ice sheet models to test some of the existing theories of ice sheet dynamics and their ability to fit the data at various points in time. Change Point can attempt to answer the question of when changes in the data set have occurred, while variable selection can help to detect which of several proposed models best fits the data in each interval of time. Brute force attempts to study the $\delta^{18}O$ proxy record are futile, as the solution space is enormous – there are >3.5×10^{14} ways to place 5 change points in this data set, before any considerations for variable selection are made. These algorithms make the solution space manageable. Furthermore, we know of no other approach that combines these two types of statistical models.

1.2. Background

1.2.1 The Change Point Problem

The analytical methods that exist to model non-stationary environmental records do not adequately represent a complex climate system. A common procedure is to parse paleoclimatic time series into shorter time windows of fixed length and analyze these using modifications of conventional spectral analysis [13, 77, 109, 154]. A series of Fourier spectra is sufficient to characterize intervals of time in terms of first-order changes in dominant frequency. For the glacial system, this method is used to detect changes in the sensitivity of the climate to different orbital functions. However, the series of Fourier spectra is inherently low-resolution as the boundaries of change are blurred by the moving windows of fixed size. Wavelet analysis [15, 82, 87, 142] has been applied to paleoclimatic time series as a means to better localize the spectral

properties of non-stationary time series, but like spectral methods, it too suffers from the appearance of side bands of fundamental frequencies during intervals of rapid spectral change [15] thus obscuring the underlying dynamics of the time series. The ability to jointly illustrate contributions in both the time and frequency domain is an advantage, but wavelet analysis lacks an obvious procedure for objectively determining when and how many changes have occurred in a time series for a single frequency, let alone more complex model functions. Singular Spectrum Analysis (SSA) is useful for decomposing a time series into trend (not necessarily linear), oscillatory, and noise components [53] by creating a series of orthogonal functions that can be ordered by the variance, similar to Principal Component or Empirical Orthogonal Function techniques. Like wavelet analysis, SSA can discover changes in the amplitude of the oscillatory components [53, 146, 147]. However, because the spectral resolution of oscillatory components is often rather broad [146], SSA is not well suited for detecting regime changes on glacial-interglacial variations as this orbitally-resolved time series is suspected (known) to contain strong periodic components.

At present, the unsatisfactory state of break point determination is highly evident in the studies of the transition from 41 kyr to 100 kyr glacial cycles in the Pleistocene - different approaches have variously put the onset at 1.5 Ma [130], ~0.90 Ma [92, 124], or 0.64 Ma [104] and determined that it was both gradual [110] and abrupt [92, 104]. Time series methods, like those described above, which inadequately trade off frequency and temporal resolution, may have triggered this unsettled debate. Needed is an algorithm that removes all restrictions on window size where the spacing between breaks in behavior and the length of segments depends only on the underlying behavior of the time series and the model chosen for fitting the data.

Given N data observations $Y_1 \ldots Y_N$ and a belief that k 'change points', or regime boundaries, exist, there are $O(N^k)$ distinct placements of the k change points – a combinatorial

nightmare. Various methods have been developed to try and solve the 'change point' problem in the context of climate issues. They range from methods that visually identify break points followed by further refinement [133], a Hidden Markov Model that claims only a locally optimum solution [78], Branch and Bound algorithms [1, 65], a penalized log likelihood procedure [21], Minimum Description Length algorithms [34, 91], and Dynamic Programming algorithms that can promise globally optimal solutions [1, 3, 4, 58, 129]. The Dynamic Programming change point methods have been shown to reduce the $O(N^k)$ calculation on the location of change points to a quadratic calculation, $O(N^2)$ [1, 3, 4, 57, 88, 129], making the compute time more practical. By breaking down into a problem into a set of progressively smaller sub-problems, the smallest of which can easily be solved, the full solution can be efficiently found by piecing together the sub-problems.

In Chapter 2, we detail our first approach to this problem, the Least Squares Change Point algorithm. The approach has aspects in common with Tome and Miranda [141], who described a least squares fitting procedure for fitting linear models to paleoclimate time series. However, their work is limited to piecewise linear fits and the time complexity of their algorithm grows exponentially with the length of the time series, thus becoming computationally intensive with long series. Our algorithm approaches a problem that is a generalization of the problem addressed by [141], but differs from it in two important ways: 1) Dynamic programming [7] is employed for optimization; and 2) each sub-interval can be described by linear combinations of forcing functions. Specifically, we describe an algorithm that partitions a single time series at *k* 'change points', subdividing a time series into (*k*+1) sub-intervals, each of which is characterized by a specified function using a set of parameter values. The model proposed to fit the data can take many forms, from polynomial to sinusoidal and if linear, standard least squares solutions for a regression problem can be quickly obtained. In this chapter, we illustrate the versatility of

the algorithm by looking at examples where there are changes in the mean, in the dominant (climatic) forcing function, or in linear combinations of forcing functions.

The downside to all of these algorithms is that they are unable to quantify the uncertainty associated with their solutions, both in the number and locations of the change points. Both Bai and Perron [4] and Lu et al [91] attempt to measure uncertainty by utilizing a large number of samples. However, if only a single time series exists, as is the case with a globally integrated $\delta^{18}O$ ice volume record [86], then both of these methods would fail to be able to address the uncertainty associated with a given solution. Several Bayesian Change Point methods have been developed including an algorithm for finding a single change point [20, 112], MCMC approaches [5, 24, 55, 83], Gibbs Sampling approaches [112, 137, 151], Jump Diffusion Sampling [114], and a Dynamic Programming procedure that utilizes calculations similar to the Forward/Backward calculations of a Hidden Markov Model [47]. However, Perreault et al [112] does not generalize to multiple change points and both MCMC and Gibbs Sampling approaches only approximate the posterior distributions and leave open the difficult question of assessing convergence.

In Chapter 3, we describe the Bayesian Change Point algorithm which builds off the work of Liu and Lawrence [88], to address the shortcomings of the previous methodologies. This approach to the change point problem is one of the few that uses dynamic programming to obtain an exact representation of the posterior space by reducing the complexity of the problem down to a more manageable size, $O(N^2)$. This probabilistic algorithm works in the time domain to create a probability distribution that includes the parameter values for the model and the locations of the change points; the algorithm returns a posterior distribution on the number of change points. In addition, samples drawn from the exact posterior distributions on the locations of the change points will clearly show the uncertainty inherent to the solutions. To

illustrate the usefulness of the Bayesian Change Point algorithm, we repeat the analysis of the data set containing a change in mean from Chapter 2 and explore applications to discrete data sets.

1.2.2 A Question of Variable Selection

Variable selection is a well-studied problem. Given *m* possible variables, there are 2^m distinct sub-models. Brute force attempts at variable selection quickly become intractable. Numerous methods exist for selecting the 'best' subset of regression terms. Miller [100] and references therein provide a good overview of some of the existing step-wise regression methods. Nonnegative garrote [17, 100] and Lasso techniques [40, 111, 140] have also been used in the context of subset selection. As an alternative, Ridge Regression [62, 63] has been studied extensively and has been reviewed in [36, 37, 136]. Zou and Hastie [156] developed the elastic net, a combination of Lasso and Ridge Regression.

Bayesian approaches focus on finding the posterior distribution across the ensemble of candidate sub-models. Approximations of the posterior space by stepwise regressions [8, 100, 132] can leave open the question of local versus global optima and often do not address the uncertainty of inclusion of specific variables. 'Spike and slab' was first introduced by Mitchell and Beauchamp [103]. Here, the MAP estimator could be viewed as the 'best' sub-model, but the entire, exponentially growing space needs to be searched to find this estimator. Stochastic methods such as Markov Chain Monte Carlo (MCMC) approaches [48, 64, 93, 118] as well as Gibbs Sampling, or Stochastic Search Variable Selection (SVSS) [35, 52] have been introduced to address this computational challenge. By approximating the posterior space, these approaches can address the issue of uncertainty of inclusion of specific regressors, but leave open the question of the number of samples or length of the chain needed for convergence. As

Fernandez et al [49] points out, the 'largest computational burden lies in the evaluation of the marginal likelihood of each model'. Therefore, the ability to make these calculations efficiently is of the utmost importance.

An exact representation of the posterior space for variable selection will always be exponential, but an efficient calculation for each of the possible sub-models can reduce the time complexity to be on par with the stochastic approximation methods. In Chapter 4, we describe Exact Bayesian Model Averaging or EBMA, a method to address uncertainty in variable selection without resorting to approximation techniques. We reduce the complexity of the most costly pieces of the probabilistic calculation, matrix inverses and matrix determinants, to be the same order as calculations utilized by approximation techniques. We also show how this algorithm can improve upon the existing BMA [117] and Iterative BMA [155] algorithms that have already been used with success on large problems.

As for the glacial proxy record, debate has raged within the Geological community not only with the abruptness and timing of the Mid-Pleistocene Transition, but its cause as well. Several models have been developed to try and recreate the 100 kyr glacial cycles. Hayes et al. [59] was one of the first to directly test Milankovitch Theory and concluded that the 100 kyr glacial cycle was non-linearly paced, rather than directly forced by eccentricity. Imbrie and Imbrie[72] as well as Pisias et al. [115] developed nonlinear differential models using solar insolation as input. Imbrie and Imbrie [72] were likely the first to note that the parameters of any model may have to change in time – indicating the existence of change points in the system. A footnote to the paper stated: "Although the character of the climate record is fairly constant over the last 600,000 years older Pleistocene records are quite different. For example, isotopic records in the interval from 600,000 to 2 million year ago have much reduced amplitudes and lack the 100,000 year cycle. Because the nature of orbital variation is thought to have remained

constant over the past 2 million years, we conclude that to understand these long climatic records, it may be necessary to use models whose parameters vary with time."

Some scientists have tried to abandon the orbital theory in order to explain the 100 kyr glacial cycle, attributing it instead to changes in the inclination of the Earth's orbital plane [105, 106, 107] or to changes in bedrock structure [26]. The orbital inclination of the Earth does not affect the solar insolation budget, but instead affects the climate through extra-terrestrial accretion of meteoroids and dust [105]. The authors argued that the spectral fingerprint of orbital inclination was a better match to that of $\delta^{18}O$ than is eccentricity. Changes in bedrock structure through erosion allow for the transformation of wet-based to dry-based ice sheets, which can allow larger glaciers to form. However, these theories have not gained much traction.

In the last decade, the orbital theory has been resurrected through the ideas of sea ice feedbacks and nonlinear phase locking. The 'sea ice' models envision that the long-term cooling of the Earth led to the establishment of Northern Hemisphere ice sheets large enough to survive some of the weaker summer insolation maxima [12, 54, 122, 143] with marine-based ice sheet margins replacing terrestrial-based ice sheet margins. Sea ice can act as a rapid feedback to the climate system. Along the same lines is the idea that deglaciations have been 'quantum' in nature, occurring every fourth or fifth precessional cycle [121, 126], or every second or third obliquity cycle [66, 68]. Before the Mid-Pleistocene Transition, Raymo [121] suggests that it was never cold enough for long enough at boreal latitudes to grow the large 100 kyr ice sheets. Tziperman et al. [144] further extended this idea by discussing nonlinear phase locking where they claimed that the resonance ratio could change with every glacial cycle and cautioned against using the fit of a model to prove a glacial mechanism. The question could then be asked: "Which model(s) are best supported by the $\delta^{18}O$ proxy record?" Change point in conjunction with variable selection is therefore ideally suited to answer the questions posed by this data set.

In Chapter 5, we combine EBMA with the Bayesian Change Point algorithm to analyze the $\delta^{18}O$ proxy record. Bayesian Change Point will answer the questions surrounding the timing of the changes in the glacial record while variable selection will answer the question of *which* model is best at each point in time. Together, they can begin to unravel the mystery surrounding the Mid-Pleistocene Transition.

Finally, in Chapter 6, we provide a summary of the results and conclusions from the previous chapters as well as a glimpse into future directions of study.

Chapter 2

The Least Squares Change Point Algorithm

2.1 Introduction and Background

A very large fraction of paleoclimatic research involves producing and analyzing time series. Instrumentation and laboratories have become progressively more efficient at generating long, well-resolved representations of the earth's past climate. Examples include an effort to characterize global temperatures over the past millennium from a variety of proxies [18, 44, 76, 95], millennium-long reconstructions of equatorial Pacific surface oceanography using stable isotopes from corals [29, 30, 116], high-resolution stable isotope and trace gas records from polar ice cores [22, 33, 98, 113] and records of glacial-interglacial climate cycles derived from ocean sediment cores [14, 60, 70, 73, 77, 86, 128].

Although these different time series resolve past climatic change at different time scales, nearly all share one characteristic: they are non-stationary over the length of the record sampled. Changes routinely occur in the mean (trends), amplitude of variability, and in dominant spectral frequencies over the course of time [29, 53, 76, 92, 94, 108, 119, 131, 146]. One might then idealize paleoclimatic time series as a succession of a number of segments with internally homogeneous statistical properties, bounded by abrupt or gradual shifts to subsequent or antecedent regimes. Such a view might lead to progress in understanding climate evolution in a number of ways. It might help optimize sampling strategies, by allowing paleoclimatologists to improve the temporal resolution or spatial coverage of "model" segments of time, which represent much longer spans of earth history adequately from a statistical point

of view. Identifying the duration of "regimes" and determining the timing of their transitions might also lead to a more fundamental understanding of the underlying forces in the climate system that lead to the non-stationarity observed at many time scales. However, Wunsch [153] cautions that purely random fluctuations of a stationary time series may yield a time series that appears non stationary.

Since we do not have theories that confidently predict when and how often the climate system went through transitions, the search at present is empirical. We introduce here the change point method to the study of one particular aspect of the earth's past in which non-stationarity is already well known: the evolution of oxygen isotopes measured from benthic foraminifera, a proxy for high latitude climate change (high latitude ice sheet growth and decay, and variations in temperatures of deep water masses fed by the high latitude surface oceans). To represent past ice sheet variations, we use a 5 Myr long orbitally tuned, benthic $\delta^{18}O$ stack [86], which was derived by compiling 57 existing oxygen isotope records from around the world, as well as a shorter 2.5 Myr-long non-orbitally, tuned stack [66]. Much previous work has established that the benthic $\delta^{18}O$ record contains a number of periodic components produced or paced by cyclic variations in the Earth's orbit and thus, the amount of solar insolation received by the Earth.

Variations in the amount of solar insolation received at any latitude and season at the top of the Earth's atmosphere are caused by changes in the position of the Earth in its orbit around the Sun. This motion can be described by three parameters: obliquity (tilt), precession (wobble), and eccentricity (ellipticity) of the Earth's orbit. Eccentricity controls the absolute amount of insolation, whereas obliquity and precession control the distribution of insolation on the Earth's surface. Accordingly, solar isolation at any latitude and season can be represented as a function of obliquity and precession [72].

Even a simple visual inspection of the $\delta^{18}O$ proxy record [92, 108, 120, 130] illustrates that over the last 5 million years the earth system has experienced intervals of markedly differing response. We know from previous work that the oxygen isotope record contains at least two first-order transitions in behavior: the shift from earlier 41 kyr-dominated glacial cycles to later 100 kyr dominated glacial cycles, the so-called "Mid Pleistocene Transition", which occurred sometime between 0.8 and 1.2 Ma [11, 128] and the earlier intensification of large-scale northern hemisphere glaciation at around 2.7 Ma [120, 134]. We evaluate the strengths and weaknesses of the change point method in identifying both of these major changes in paleoclimatic behavior and in deducing more subtle aspects of climate evolution over the past 5 Ma. The results are sufficiently positive to suggest that the change point method may be applied to many other time series in which changes in behavior are suspected.

2.2 The Least Squares Change Point Algorithm

Since we want to consider all possible locations of change points and lengths of intervening segments (i.e. climate regimes), change point analysis of the $\delta^{18}O$ time series presents a huge number of possible solutions. Dynamic programming provides a practical path through the forest of possibilities [7]. It guarantees globally optimal solutions of problems that can be divided into progressively smaller sets of sub-problems, the smallest of which can easily be solved. The method then builds solutions to a progressively growing set of sub-problems until the original full problem has been solved. Dynamic programming has become a well established means of solving problems that can be partitioned in this way and is described in many operations research textbooks [61, 85, 149, 152]. This methodology has been widely and successfully applied in many fields, including extensive application in the emerging fields of genomics and bioinformatics [38, 75, 80]. Dynamic programming was first applied to the change

point problem by Hawkins [57]. Change point analysis was first applied to biopolymer sequences by Auger and Lawrence [3], and Liu and Lawrence [88] show how Bayesian inference procedures allow for assessment of uncertainties in the unknown parameters in a change point analysis.

In this section, we address the following problem: given a time series, (for example, the benthic $\delta^{18}O$ time series), and a proposed set of models, we seek to minimize the sum of squared error over all possible sets of parameters and over all possible locations of change points. We impose the constraint that no two change points are within a minimal distance of each other. [This constraint arises so as to avoid aliasing of frequencies due to sparse data points in a time series when using sinusoidal models or more generally to ensure that enough data exists to make an appropriate inference on the parameters of the model]. For an initial choice of model (constant, polynomial, sinusoidal, etc.) and number of sub-intervals (change points), the algorithm produces the best possible fit to the entirety of the data. It allows us to relax the requirement of a fixed window that has been used in past paleoclimatic analyses and replace it with the ability to use windows of any size. The use of sinusoidal functions is not new. Fourier based methods depend upon the sinusoid and these functions have been shown to give a reasonable fit to the data.

The algorithm has three steps: 1) Parameter Fitting: for each possible model (in this case, sinusoids) in each of the $N(N+1)/2$ sub-intervals of a time series of length N; 2) The Forward Step: find the minimum sum of square residuals considering all possible positions of k change points for the full time series and for all sub-intervals that begin at time zero; and 3) The Backtrace Step: identify the solution (parameter set or argument) minimizing the sum of square residuals, i.e. the optimal set of change points. We begin with a series of time points t_n, $n=1, 2, \ldots, N$ that are not necessarily equally spaced. At each of these points, the measured value of the

proxy of interest Y_{t_n} is available. We also assume that we have *M* different models that are proposed to fit the data. The following is a summary of the computer code, the so called pseudo code, used in each of these steps.

Step 1: **Parameter fitting for *M* periodic functions with given frequency components**

for *i = 1,…,N*

for *j = i +d,…,N*

for *m=1,…,M*

$$SS_{i,j}^{m} = \underset{\substack{\alpha_0 \ldots \alpha_p \\ \beta_1 \ldots \beta_p}}{Min} \left\{ \sum_{t=t_i}^{t_j} \left(Y_t - \alpha_0 - \sum_{p=1}^{P} \alpha_p \sin(\omega_p t) + \beta_p \cos(\omega_p t) \right)^2 \right\}$$

end for

$$SS_{i,j} = Min\left\{SS_{i,j}^{1}, \ldots, SS_{i,j}^{M}\right\}$$

end for

end for

where ω_p is a specified frequency, SS_{ij} is the sum of square residuals in the sub-interval beginning at position t_i and ending at position t_j, and $d \geq 0$ is the size of the smallest permitted window. In cases where we want to consider *M* different, distinct models, the inner loop will allow the algorithm to find the minimum sum of square residuals for each model individually before choosing the minimum among all competing models.

While our focus here is on Milankovitch cycles, this procedure is applicable to a broad spectrum of fitting functions. For least squares fit to a general fitting function $f(\cdot\,|\Theta)$ with parameters Θ step one becomes:

for *i = 1,…,N*

for *j = i +d,…,N*

for *m=1,…,M*

$$SS_{i,j}^{m} = \underset{\Theta}{\text{Min}} \left\{ \sum_{t=t_i}^{t_j} \left(Y_t - f\left(X_{1,t}, \ldots, X_{p,t} \middle| \Theta \right) \right)^2 \right\}$$

end for

$$SS_{i,j} = Min \left\{ SS_{i,j}^{1}, \ldots, SS_{i,j}^{M} \right\}$$

end for

end for

where $X_{i,t}$ are a covariates, or "predictive variables". As indicated next, the remaining two steps of this algorithm can be completed for any function for which the necessary minimizations can be completed to obtain $SS_{i,j}$, *i=1, …, N & j > i*. Thus, $f(\cdot|\Theta)$ may be nonlinear. When a linear function is employed, the well known explicit solutions of linear regression equations can be employed to obtain the required minima.

The step just described fits the parameters and finds the minimum sum of square residuals for every possible sub interval. To find the optimal *k* change points, we find the overall minimum sum of square residuals by optimally piecing together *(k+1)* sub- intervals that include each of the *N* data points only once. There are a large number, $O(MN^K)$, of ways to piece together sub-intervals [for each value of *k*], so except for a short time series, this optimization cannot be completed by enumerating all of the possibilities [A total of $\sum_{k=1}^{k_{max}} MN^k$ solutions exist]. Instead, dynamic programming, a recursive algorithm, addresses this combinatorial explosion by breaking the problem into a progressive set of smaller problems. In change point analysis, this progressive set includes all substrings of the sequence that begin with the first data point and end at all time points from time 2 to time *N*. Optimal values, here the minimum

squared errors, are first obtained for the simplest of these, the one change point problem. The stored values from this first pass are then used to find the minimum squared errors for all solutions with two change points and these in turn for three change point models, and so on, until the maximum number of change points, K_{max}, has been included. Upon completion of the forward step, minimum squared errors for all sub-problems have been obtained and stored. Let $f_k(v)$ be the minimum sum of square deviations for the sub-interval from 1 to v with k change points. Note that $f_0(v) = SS_{1,v}$.

Step 2: Forward step

for $k=2,\ldots K$

for $n = k, \ldots, N$

$$f_k(t_n) = \min_{v< t_n}\{f_{k-1}(v) + SS_{v+1,t_n}\}$$

end for

end for

The forward step finds the minimum sum of square residuals with up to K_{max} change points, for the entire time series and all possible sub-intervals. To find the locations of the change points, we work backwards. On the backward step, the results from the forward step are employed to find the optimal solution. The location of the K^{th} change point is found by using the minimum squared error for the set of sub-problems with K-1 change points. The remaining change points are found recursively by stepping backwards in a similar manner. This means that we begin at the end of the time series, find the location of the last change point, and then recursively repeat this process with smaller values of k and progressively shorter sequences using the stored values from the forward procedure at each stage of the backwards procedure. Let c_k be the location of the k^{th} change point

Step 3: Backtrace

$c_{K+1} = t_N$

for *k = K,K-1,…1*

$$c_k = \text{argmin}_{v\epsilon\,[k-1,c_{k+1})}\{f_{k-1}(v) + SS_{v+1,c_{k+1}}\}$$

end for

Alternatively, if a pointer is stored on the Forward Step indicating which '*v'* satisfies the minimization, then the Backtrace can be accomplished using a set of these pointers.

These steps reduce the time requirements for this computation to $O(MN^2)$ for the parameter fitting step and *O(KN)* for the forward and backward steps.

2.2.1 Selecting the Number of Change Points

There is no statistically rigorous method to determine the appropriate number of change points in a least squares analysis. As an alternative, we examine changes in squared error as a function of the number of change points in an effort to identify the 'break in the curve', beyond which adding more change points yields relatively little improvement in fitting the data.

The first change points added to a model will provide a more significant reduction in squared error than the addition of later change points, as the initial change points highlight the most important distinctions between segments of the data set. Suppose that a data set contains a single change in mean with random noise. The most important change, and therefore the first identified by the algorithm, will be the change in mean rather than, for example, identifying a few 'noise' points as distinct from the rest of the data set. The identification of the signal will improve the fit of (and reduce the squared error for) more points than the identification of small groups of outliers. Therefore, the cumulative effect of

identifying the signal will be larger than the cumulative effect of identifying the noise. Thus, a (generally) downward sloping histogram containing the reduction in squared error as a function of the number of change points can be produced. One important caveat to this argument is that if an interval is distinct from the rest of the data set, then a pair of change points will need to be added (one at each end of the interval) to be able to isolate that interval from the rest of the data set. Thus, the histogram may not monotonically decrease in the number of change points

The histogram begins to flatten out at the point when additional change points begin to fit the noise rather than the signal itself. While the model's fit to the data can always be improved by adding more change points, there comes a point where all the major features have been identified and only small, incremental improvements remain. In essence, a constant amount of squared error is removed with each additional change point. The location in the histogram where the slope changes from downward to more flat is known as the 'break in the curve'. Therefore, an appropriate number of change points to select would be the last change point which removed a significant amount of squared variation. Figure 2.2 provides an example of this histogram using the example in Section 2.3. Here, the caveat regarding the addition of a pair of change points described in the previous paragraph is clearly seen. A researcher should select 4 change points for analysis, as the addition of the fourth change point is the final time that a significant amount of variance is removed by the addition of a change point. A second example can be found in Figure 2.5, the analysis of the $\delta^{18}O$ proxy record. An argument could easily be made that the 'break in the curve' occurs at either 4 or 6 change points in the top pane of the figure and either 5 or 7 change points in the bottom pane, with 4 and 5 being the most obvious choices, respectively.

Since this selection is subjective, all conclusions should be checked for robustness against variations in the number of change points. A Bayesian approach to this problem,

outlined in Chapter 3, will provide a more rigorous way to determine the number of change points.

2.2.2 Variations in the Error Variance - Normalized Regression

A weighted least squares approach should be considered if the error variance is not believed to be constant across the entire data set. The change point algorithm will find the optimal placement of change points so as to minimize the squared error over the entire data set. However, if one segment of the data has a larger error variance than another segment of the data, then the former data points will 'cost' more (in terms of squared error) to misfit than the latter ones. The algorithm may try to reduce this cost by placing additional change points in the more 'costly' section of the data at the expense of a less optimal fit in other parts of the data set. To circumvent this problem, one can use a weighted least squares approach (i.e. a normalized regression) where the squared error for each segment of the data is weighted by the inverse of the variance of the data in that segment before completing the forward step of the recursion. The pseudo code for the squared error calculation would now be:

for *i = 1,…,N*

for *j = i +d,…,N*

for *m=1,…,M*

$$SS_{i,j}^{m} = \underset{\Theta}{\text{Min}} \left\{ \sum_{t=t_i}^{t_j} \left(Y_t - f\left(X_{1,t}, \ldots, X_{p,t} \middle| \Theta \right) \right)^2 \right\}$$

end for

$$SS_{i,j} = {Min\left\{SS_{i,j}^{1}, \ldots, SS_{i,j}^{m}\right\}} \Big/ {var(y_{i,j})}$$

end for

end for

An example of normalized regression will be presented in Section 2.4

2.3 Application: Finding a Change in Mean - Copy Number Variation

We begin with a simple example to illustrate how the least squares change point algorithm works.

The human genome consists of 6 billion bits of information, commonly referred to as DNA, which determine everything about who we are. Our DNA is packaged into two sets of 23 chromosomes, one set inherited from each of our parents. Roughly 30,000 protein coding genes are encoded in these 6 billion bits of information and one could safely assume that since we have two sets of chromosomes, then we have two copies of each gene. But this does not always have to be the case.

Copy number variations are widespread in the human genome and represent a significant source of genetic variation [125]. Instead of having the standard two copies of each gene, a person may have one or both copies of the gene either deleted or duplicated. Thus, when comparing the genomes of two people, the number of copies of a given gene may vary. Copy number variations have been linked to many different types of diseases including cancer, HIV acquisition and progression, autoimmune diseases, and Alzheimer and Parkinson's diseases [45].

To keep the example simple, suppose that we can measure the amount of all genes present in a cell. Like all measurements, some random error will exist. Most genes will be present in the same quantity (2), while others will be of either a lesser (0, 1) or greater (≥3) quantity. Therefore, we will have three levels, 'loss', 'gain', or 'neutral'. By identifying changes

from one level to the next, we can discover where copy number variations are occurring and in what magnitude.

To simulate this phenomena, a sequence of 1000 data points has been generated at three levels, 1, 2, and 3, with added white noise ~N(0,1) [Figure 2.1]. Data points 1-350 will be generated as '2', 351-425 as '1', 426-700 as '2', 701-815 as '3', and 816-1000 as '2'. Thus, four change points exist in the data, at positions 350, 425, 700, and 815. Since our goal is to model changes in the mean of the data, our regression model takes the simple form $f(x) = C$, where C is a constant. The three constants can either be specified by the user or selected by the algorithm. In this example, we will let the algorithm select the constants in order to showcase its ability to make inferences about the parameters of the regression model.

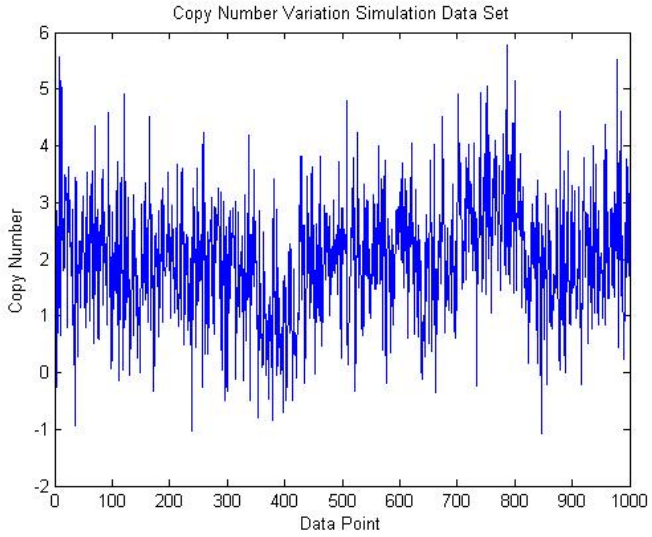

Figure 2.1: Copy Number Variation Simulation Data Set. 1000 data points were simulated at one of three baseline levels (1, 2, or 3). White noise was added to simulate instrumental or measurement errors. Four change points exist in the data set at positions 350 (from 2 to 1), position 425 (from 1 to 2), position 700 (from 2 to 3), and position 815 (from 3 to 2).

Before obtaining results from a change point analysis, a decision must first be made on the appropriate number of change points to analyze. As indicated in Section 2.2.2, there is no rigorous way to determine the 'correct' number of change points, so we instead have to look for

the 'break in the curve' in the histogram comparing the reduction in squared error with each additional change point. Figure 2.2 shows that the last significant reduction in squared error occurred with the addition of the fourth change point. Therefore, we run the change point analysis using four change points. The global optimal solution to this problem places the four change points at positions 351, 424, 700, and 811 [Figure 2.3]. This matches well with what we know to be their true locations [350, 425, 700, and 815]. Next, given the location of the change

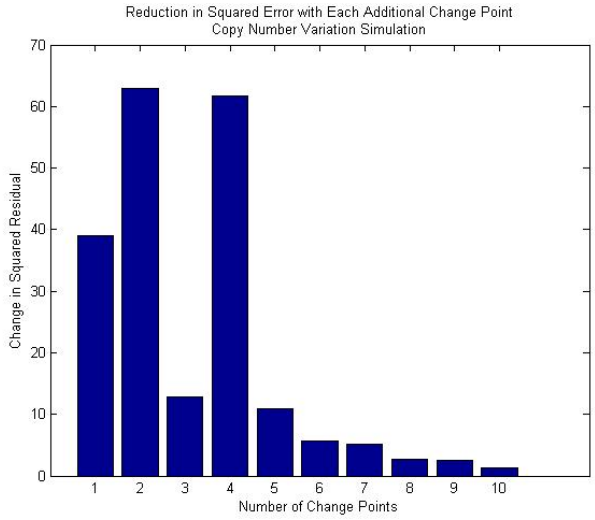

Figure 2.2 Reduction in Squared Error with each Additional Change Point – Copy Number Variation Simulation Data. The slope of the squared error curve as a function of the number of change points is indicated for the Copy Number Variation simulation data set. The 'break in the curve' or the last significant reduction in squared error occurs with the addition of the fourth change point.

points, we want the model that best fits the data in each interval. Using the standard least squares solution to a linear regression problem, we find that in the interval 1-351, the best fitting model is f(x) = 1.972, which again corresponds well with what we know to be the true model, f(x) = 2. In the other four intervals, the inferred model also matches the true model, picking constants of 0.917, 2.012, 2.903, and 1.961, compared to the true values of 1, 2, 3, 2 respectively [Table 2.1].

We note that a Hidden Markov Model (HMM) can also be used to study a change in mean [78]. However, HMMs can get trapped in locally optimal solutions and they require an iterative approach due to their use of the Baum-Welch algorithm to estimate the parameters of the regression model, whereas the Change Point algorithm obtains the global optimum solution in a single pass through the data.

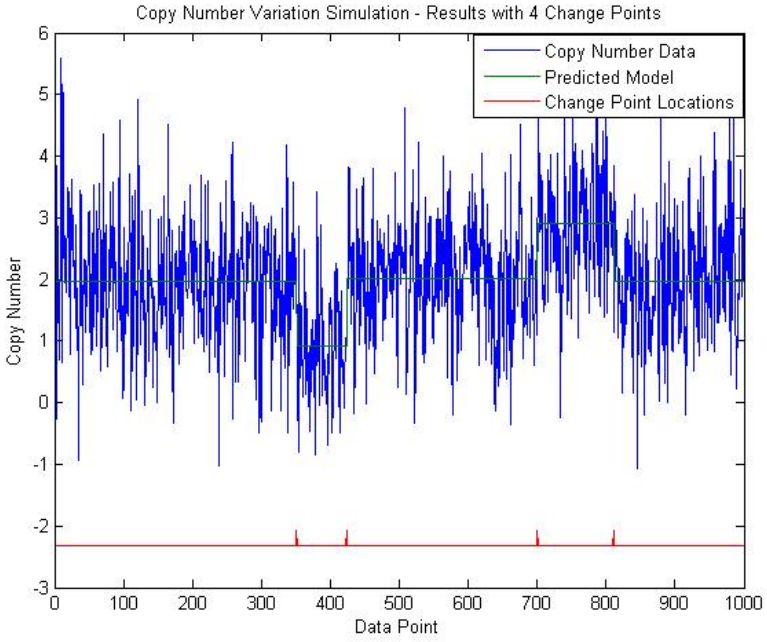

Figure 2.3 Copy Number Variation Simulation Results with Four Change Points. The locations of the four change points are indicated as red spikes at the bottom of the figure, corresponding to locations 351, 424, 700, and 811. The blue curve represents the simulated data while the green curve represents the inferred (constant) model for each sub-interval created by the placement of the change points.

Interval Start	Interval End	Predicted Copy Number	True Copy Number
1	351	1.972	2
352	424	0.917	1
425	700	2.012	2
701	811	2.903	3
812	1000	1.961	2

Table 2.1 Simulation of Copy Number Variation Data. The optimal placement of four change points creates the five sub-intervals of the data indicated in the table. The third column gives the optimal value for the constant model (in terms of squared error) in each sub-interval. The final column provides the constant used to generate the data in that sub-interval.

2.4 Application: The $\delta^{18}O$ Record of the Plio-Pleistocene

The change point algorithm was applied to two different types of $\delta^{18}O$ proxy records. One data set is the orbitally tuned 5 Myr-long isotope stack of Lisiecki and Raymo [86] [hereafter LR05]. Long climatic records have poorly constrained time scales, but are assumed to have fluctuations associated with solar insolation. As a result, the timing of features in a record may be adjusted to match predictions of the orbital theory in the hopes of improving dating accuracy. Thus, after aligning the 57 cores from around the world, Lisiecki and Raymo [86] adjust their age model so that the observed fluctuations in the $\delta^{18}O$ record correspond to the Milankovitch cycles. Their tuning target is a simple nonlinear model of ice volume [72] which takes the June 65°N insolation curve as its forcing function. Because orbital tuning may induce a bias in the locations of the change points, a second analysis was also done on a 2.5 Myr-long data set that is free from orbital tuning [66] [hereafter H07], an extension of Huybers and Wunsch [67]. The H07 approach pays the price of its freedom from orbital tuning by adding considerable age uncertainties in regions distant from absolute age control points. To address this uncertainty, several realizations of plausible age models would have to be simulated and independently tested by the change point algorithm to ensure that results are not due to stochastic fluctuations. This type of analysis is beyond the scope of this application and we chose not to weigh in on the merits of one type of data set over the other.

The two data sets are quite similar. As Huybers [66] points out, the differences between the ages of the two models has a standard deviation of only 6 kyr, which is less than the expected uncertainty for the depth-derived ages. There are however, two important differences. Prior to 600 ka, the sampling densities of the two data sets are different. H07 has data points corresponding to every 1 kyr, while LR05 has data points spaced every 2 kyr from 600 ka to 1500 ka, and then every 2.5 kyr from 1500 ka to the end of the H07 data set.

Secondly, a circa 100 kyr pulse is seen in the H07 data at 2 Ma that is not found in LR05. Accordingly, the change point algorithm will bound this interval with a pair of change points when the 100 kyr sinusoid is utilized in a model. A comparison of the results from the algorithm on the two published data sets is given below in Section 2.4.2 'Linear Combinations of Milankovitch Frequencies'. This type of comparison provides insight into the most important changes of the climate system that are visible across different models and across different data sets.

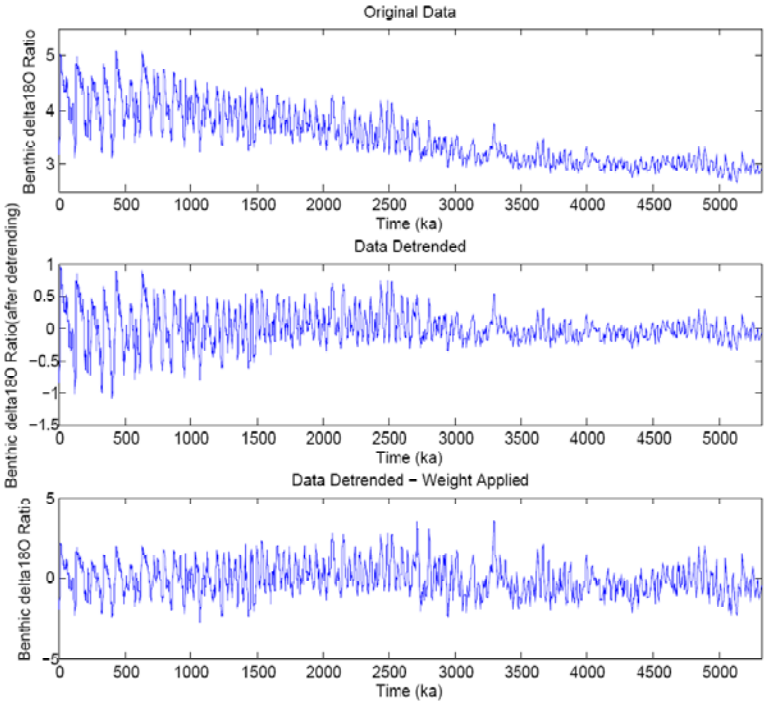

Figure 2.4 Benthic $\delta^{18}O$ data. a) The original $\delta^{18}O$ record [86]; b) $\delta^{18}O$ data after the overall trend was removed using an exponential function; c) The detrended data after a weight was applied to each sub-interval. The sub-intervals were selected by the change point algorithm when the input was the linear combination of the Milankovitch frequencies [Actual change points given in Table 2.3]. The weight used was 1/(standard deviation) of that sub-interval.

The first step for either data set is to remove the overall trend from the data via an exponential function so that the change points represent changes in the Earth's response to

orbital inputs rather than the cooling and long-term Plio-Pleistocene trend towards increased ice volume (See Figures 2.4a-b).

All analyses described here include change points associated not only with changes in the dominant frequencies, but also changes in the amplitude and phase parameters for a given frequency. The phase parameter takes values between -180 degrees and +180 degrees and gives the phase shift of the sinusoidal functions relative to time 0 (present day). This choice was arbitrary and has no bearing on the final results. The results are identical if we choose t_1 to be the most recent or the most distant time point. The amplitudes are in $\delta^{18}O$ units. Here and in the remainder of Section 2.4, we constrained the minimum interval length to be at least as large as the longest wavelength included in the regression model. Due to the nature of the algorithm, all change points will represent an abrupt or discrete change in the system when considered in isolation. However, gradual changes can be realized when the change points and resulting model parameters are viewed in context. Several nearby change points that show slowly varying or step-wise changes in parameter values are indicative of a gradual change in the system.

There is a marked decrease in the variability prior to about 2500 ka (Figure 2.4b). Accordingly, change points before 2500 ka are less likely since there is less squared variation to account for. Therefore, in addition to an analysis directly on the detrended data shown in Figure 2.4b, we conducted a 'weighted' least squares analysis as described in Section 2.2.2. After completing step one on each sub-interval, we weighted sub-intervals by 1/variance around the mean of that interval prior to application of Steps 2 and 3 (Figure 2.4c). In general, we found that this weighting made little difference except between ~2400 and ~2700 ka, where an additional change point was located in this region in the weighted analysis compared to the un-weighted analysis, and within the 'transition interval' ~800 to ~1200 ka discussed below, where

we saw a shift in the location of the change points. Here, we present the weighted analysis while referring the reader to the un-weighted analysis located in the supplement to [129].

2.4.1 Dominant Frequency Analysis

We begin our analysis by examining the LR05 time series and addressing the traditional problem of identification of the single dominant orbital frequency in each sub-interval, selecting in each sub-interval only one of the three sinusoidal functions: "eccentricity" (100 kyr), obliquity (41 kyr), or precession (23 kyr). This is an example of the pseudocode in Section 2.2.2 with M=3. Strong support for the existence of these frequencies within at least some region of the data already exists through conventional Fourier analysis. Other sinusoidal functions could also have been chosen. The choice of 100 kyr to represent eccentricity was based on the empirical power of a circa 100 kyr cycle during the late Pleistocene. This model is a first attempt to explain the data and serves primarily as a comparison to spectral analysis techniques that have been used in paleoclimatology.

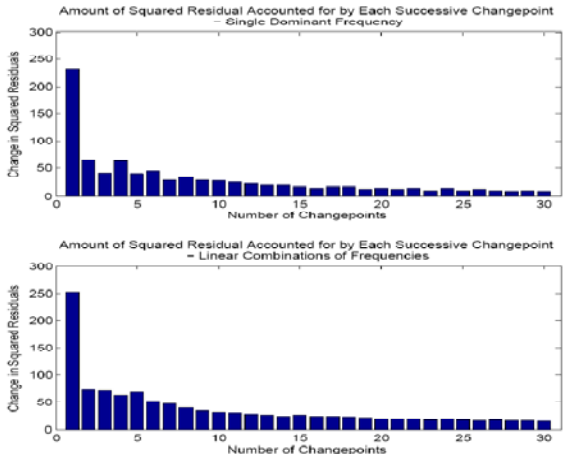

Figure 2.5 Amount of Squared Residual Error Accounted for by Each Successive Change Point The slope of the squared error curve as a function of the number of change points: a) the change in the residual sum of squares with the addition of each successive change point for the weighted analysis of the single dominant frequency b) the change in the residual sum of squares with the addition of each successive change point for the weighted analysis of linear combinations of the Milankovitch frequencies. The 'break in the curve' corresponds to a change in the slope of the squared error curve.

Recall from Section 2.2.1 that there is no statistically rigorous method to determine the number of change points required in a least squares analysis. As an alternative, we examined changes in squared error as a function of the number of change points in an effort to identify the 'break in the curve'. For the analysis of dominant frequencies, we used four change points (Figure 2.5a).

Single Dominant Frequency: Milankovitch Inputs (23, 41, 100 kyr), 4 Change points					
Constraint: Minimum interval length 100kyr			Phase in degrees (-180 to 180)		
Sub-Intervals	Dominant Period (kyr)	Amplitude ($\delta^{18}O$)	Phase	Constant (α)	R^2
1 to 113	100	0.4584	-150.78	0.1435	0.5368
114 to 424	100	0.5049	141.47	-0.0804	0.566
425 to 778	100	0.412	-160.20	0.021	0.5218
780 to 2713	41	0.2188	158.78	0.037	0.2895
2715 to 5320	41	0.0896	-156.12	-0.0468	0.1828
	Total R^2 = .4615				

Table 2.2 The Single Dominant Frequency [LR05]. The single dominant frequency in each of 5 sub-intervals using only the three Milankovitch frequencies: 23, 41, and 100 kyr, with the constraint that the minimal sub-interval length be at least as long as the longest wave in the analysis (100 kyr). The amplitude is given in $\delta^{18}O$ units, the phase in degrees from -180 to 180, and the R^2 values are the percent of the squared variation that a single wave is able to account for on its own.

Our dominant frequency analysis indicates that the most recent 1 Myr exhibit a more complex pattern requiring more change points than the previous 4Myr, even in the weighted analysis. Thus, as shown in Table 2.2, when four change points are selected, three are between 778 ka and the present (113 ka, 424 ka and 778 ka) and the regimes bounded by these change points are dominated by the 100 kyr cycle with changes in the amplitude and phase parameters. Figure 2.6 shows the observed data and the fitted functions. We find a change point near the minimum permitted length for the most recent interval, here at 113 ka. When we relaxed the minimum length constraint, we consistently found that the most recent 80 ka includes a change point that seeks to capture an unusual pattern of the recent past. The change point at ~780 ka persists throughout the subsequent analyses (Tables 2.3, 2.4 and 2.5). The two intervals prior to ~780 ka [780-2713 ka and 2715-5320 ka] are both best fit by the obliquity forcing function with a threefold change in the amplitude of the obliquity response at about 2713 ka (Table 2.2). Only

the influences of the 41 kyr (obliquity) and 100 kyr (eccentricity?) forcing functions are inferred over the entire time series. The un-weighted analysis with four change points yields similar results to the weighted analysis, with change points at 113, 425, 786, and 2703 ka, and an R^2 value that is less than half a percent greater than the weighted least squares, 0.465.

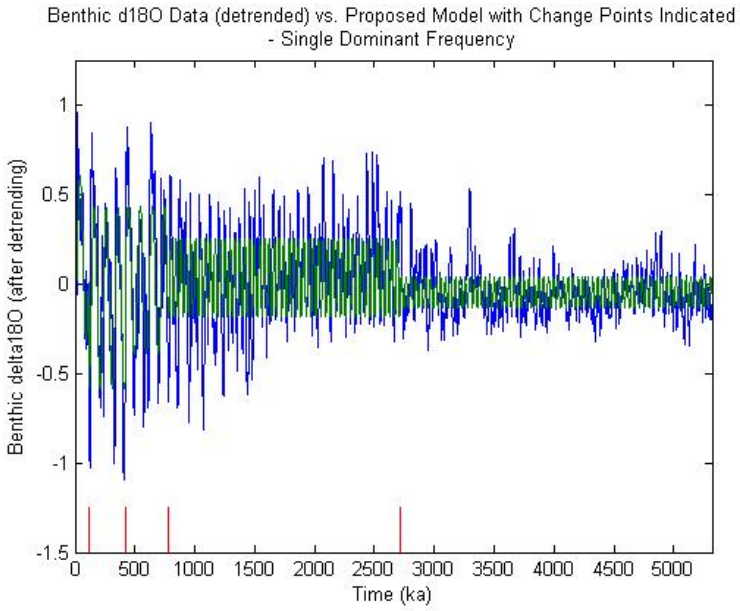

Figure 2.6 Benthic $\delta^{18}O$ Data Plotted vs. Proposed Model with Change Points Indicated – Single Dominant Frequency [LR05]. The least squares fit for the single dominant Milankovitch based cycle in each sub-interval. The data is plotted in blue, the proposed model in green, and the change points as red spikes at the bottom of the figure. The data shown is the original data after the trend was removed. The change points were selected using the weighted least squares approach, but are plotted against the unweighted data.

Over the full 5320 kyr interval, less than half [0.462] of the squared variation (R^2) in the ratio of oxygen isotope values is accounted for by the best fit of a single dominant "Milankovitch" frequency in 5 sub-intervals (Table 2.2). Furthermore, in the interval from 2715 ka to 5320 ka only 18% of the squared variation is accounted for by the best fitting of these functions (Table 2.2). Table 2.2 illustrates the ability of the algorithm to identify dominant frequencies without the required restrictions on window size. Increasing the number of change points provides a better fit, but doubling the number of change points (to 8) only increases the

total R^2 value to 0.521. Reducing the number of change points to 2, corresponding to the two well-recognized changes in the behavior of Plio-Pleistocene glaciation [the Mid-Pleistocene Transition (~900ka) and the intensification of northern hemisphere glaciation ~2.7Ma] returns changes at 692 ka and 2713 ka, but an R^2 value of only 0.357.

In the 2 change point analysis, the Mid-Pleistocene Transition occurs at 692 ka in comparison to its location at 778 ka in the 4 change point analysis. The movement of the location of change points when additional change points are added to the analysis is not unexpected. For example, suppose that the data contained more 'changes' than the number of change points the algorithm was asked to find. If this were the case, a single sub-interval may have to account for several inhomogeneous regions within its bounds and would thus average the parameter values across these inhomogeneous regions. With an appropriate number of change points, each sub-interval truly has homogeneous parameter values within its bounds.

We were also concerned that our results might be sensitive to the precise frequency chosen for the circa 100 kyr glacial cycle. To address this concern, the single dominant frequency analysis was also separately run using 95 kyr and 104 kyr cycles instead of the 100 kyr cycle. When the 95 kyr cycle was used instead of the 100 kyr cycle, the change points were located at time positions 112, 788, 916, and 2713 ka, with a total R^2 value of 0.449, a 1.5% reduction. As was true for the analysis using a 100 kyr sinusoid, the circa 100 kyr cycle, in this case the 95 kyr term, dominates [i.e. best fits] the three most recent intervals, while the 41 kyr cycle best fits the other two intervals. When using the 104 kyr cycle instead of the 100 kyr cycle, the change points occur at time positions 423, 792, 1216, and 2715, with a total R^2 value of 0.422, a 4.3% reduction. Again, the circa 100 kyr cycle dominates the most recent three intervals [0- 1216 ka], while the 41 kyr cycle dominates the other two [1218-5320 ka]. Importantly, throughout this range of candidate eccentricity frequencies, the change points at

~790 ka and ~2715 ka remain unchanged. Our modeling of circa 100 kyr behavior in glaciation over the past 5 Ma is therefore quite reliable, whatever our limitations on its interpretation may be.

2.4.2 Linear Combinations of Milankovitch Frequencies

Next we consider models with linear combinations of sinusoidal functions of the three classic Milankovitch frequencies (100, 41, and 23 kyr). This is a special case of the pseudocode in Section 2.2.2 with M=1. In this section, we analyze both the 5 Myr orbitally tuned benthic δ^{18}O LR05 data set as well as the 2.5 Myr non-orbitally tuned benthic δ^{18}O H07 data set. The goal of this section is not only to compare the results of the change point algorithm on the two data sets, but to be able to draw inferences that are independent of the age models used. We begin our discussion with an analysis of the LR05 data set.

In this analysis, we increased the number of change points to seven to show a more nuanced result (Figure 2.7 and Table 2.3). Change points occur at 102 ka, 380 ka, 786 ka, 1030 ka, 1198 ka, 2418 ka, and 2713 ka. Five of the seven change points occur between the early onset of the Mid-Pleistocene Transition identified in the literature ~1500 ka (Rutherford and D'Hondt 2000) and the present, while the remaining two change points occur near the intensification of northern hemisphere glaciation ~2.7Ma. As shown in Ruggieri et al [129], adding up to 10 change points does not relocate six of the seven change points and moves the seventh change point by only 2 kyr.

To better understand the contributions of each of the Milankovitch terms to each sub-interval (i.e. climate regime), we examine the contributions of all subsets of Milankovitch terms to the proportion of squared variation (R^2) accounted for in each of the predicted sub-intervals, with the seven change point locations fixed to those in Table 2.3 (Table 2.4). A '*' indicates the

subset of Milankovitch sinusoids that accounts for the most variance in each sub-interval. The contribution of the precession-based term is small compared to the obliquity and eccentricity terms (Table 2.4).

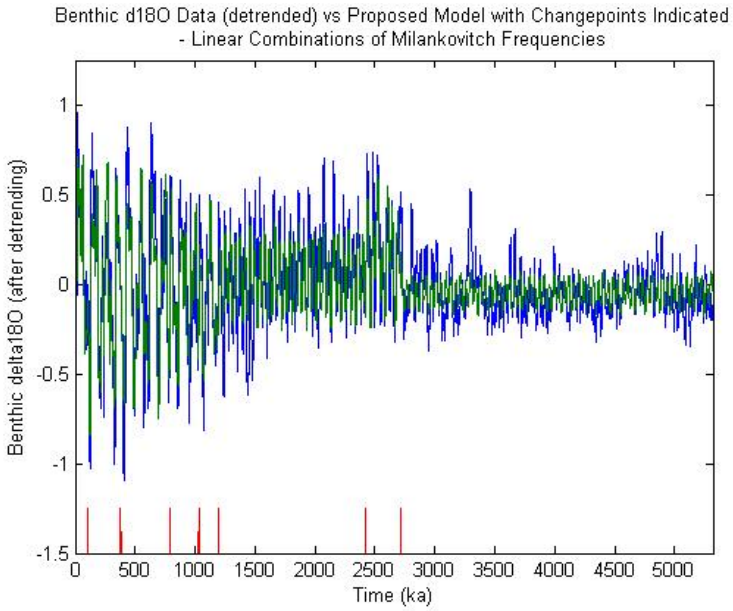

Figure 2.7 Benthic $\delta^{18}O$ Data Plotted vs. Proposed Model with Change Points Indicated – Linear Combinations of Milankovitch Frequencies [LR05]. The least squares fit for the linear combination of the three Milankovitch based cycles and a constant term with seven change points. The actual data is in blue, the proposed model is in green, and the red spikes at the bottom of the figure represent the optimal locations of change points selected by the algorithm.

The amplitude of the obliquity-driven glacial cycle has a roughly constant magnitude from 2715 ka to the present, even though it is not the dominant frequency for the past 1.2 Myr (Table 2.3). Prior to 1.2 Ma, the obliquity term dominates while the other two terms add little to the overall fit. This result is robust against the selection of the number of change points, as opting for fewer (5) or more (10) change points yields similar results. While the obliquity response dominates in the interval 2420 ka to 2713 ka, the amplitude of the eccentricity based function increases relative to adjoining intervals. It is perhaps not coincidental that this ~300 kyr interval of mixed spectral behavior follows immediately after the significant increase in the amplitude of glacial

cycles at about 2713 ka. In all cases, the amplitude of the obliquity based function was nearly three-fold lower before 2715 ka ago, consistent with the absence of large northern hemisphere glaciers and their associated feedbacks during this interval [134]. In this interval, we see that all three sinusoids together are only able to account for just over 20% of the squared variation of the $\delta^{18}O$ data. We concede that our statistics concerning the dominance of obliquity forcing may be biased by the obliquity-based tuning procedure used by Lisiecki and Raymo [86] to date the $\delta^{18}O$ record. However, despite the fact that Lisiecki and Raymo [86] also included precession in their tuning procedure, the 23 kyr cycle never plays an important role in the change points we identified, and never explains a large fraction of the variance in our models.

The Milankovitch Frequencies: 23,41,100kyr							
Constraint: Minimum interval length: 100kyr					Phase in degrees (-180 to 180)		
	23kyr		41kyr		100kyr		
Sub-Intervals	Amplitude	Phase	Amplitude	Phase	Amplitude	Phase	Constant (α)
1 to 102	0.2043	71.68	0.2986	146.18	0.376	-147.34	0.1797
103 to 380	0.1962	39.51	0.2916	-161.88	0.4354	133.52	-0.0209
381 to 786	0.0344	-158.14	0.2569	-164.03	0.4675	-169.64	-0.0303
788 to 1030	0.1431	28.07	0.2002	178.88	0.2761	13.61	-0.0492
1032 to 1198	0.0925	14.63	0.1281	139.93	0.3148	-82.71	-0.0154
1200 to 2418	0.0292	-129.26	0.225	152.78	0.0553	102.22	0.0402
2420 to 2713	0.0273	29.12	0.2588	162.55	0.184	-65.01	0.177
2715 to 5320	0.012	-172.05	0.0893	-156.22	0.0293	-15.29	-0.0467
	Overall R^2 = .6641						

Table 2.3: Linear Combinations of Milankovitch Frequencies [LR05]. Linear Combinations of Milankovitch frequencies: 23, 41, and 100kyr. The seven change points selected by the algorithm yield eight sub-intervals with corresponding amplitude and phase parameters for each of the three input waves in each sub-interval. The amplitudes are given in $\delta^{18}O$ units and the phase in degrees from -180 to 180. Once again, a constraint was imposed that the minimum sub-interval length be at least as long as the longest wave in the analysis (100 kyr).

In the last 1200 kyr, the eccentricity-based term is the 'dominant' cycle because it is able to account for the most squared variation in each of these sub-intervals. The 100 kyr based

cycle shows a marked increase in amplitude beginning 1198 ka ago (Table 2.3). The amplitude further increases at 786 ka. Thus, the time between roughly 1200 ka and 800 ka could be considered a 'transition interval'. When we increased the number of change points in the single dominant frequency analysis from four to seven, the 100kyr cycle dominates from 1200 ka to the present. Thus, the extension of the interval in which eccentricity is dominant, stems from the inclusion of three additional change points, again suggesting that the interval between ~780 ka and ~1200 ka is a transition interval.

Sub-Intervals	23	41	100	23,41	41,100	23,100	23, 41, 100
1 to 102	0.1384	0.3589	0.5207*	0.5013	0.7222*	0.6171	0.8233
103 to 380	0.0862	0.198	0.4908*	0.291	0.6952*	0.5818	0.7923
381 to 786	0.0044	0.1546	0.5367*	0.16	0.7*	0.5389	0.7029
788 to 1030	0.0934	0.2313	0.3557*	0.31	0.5356*	0.4548	0.6206
1032 to 1198	0.0561	0.0764	0.5055*	0.1405	0.5876*	0.5451	0.6356
1200 to 2418	0.0063	0.3739*	0.0211	0.3804	0.3969*	0.0273	0.4032
2420 to 2713	0.0086	0.4348*	0.1905	0.4392	0.6636*	0.2006	0.6688
2715 to 5320	0.0036	0.1828*	0.0204	0.1861	0.2023*	0.024	0.2056

Table 2.4: Percent of Squared Residual Accounted for by Each Subset of Inputs [LR05]. Given the location of the seven change points from the linear combination of Milankovitch frequencies [Table 2.3], we look at the amount of squared variation (R^2) that each subset of sinusoids is able to account for. For a given number of sinusoids included in the model, the subset that accounts for the most squared variation is indicated with a '*'. Note: the Total R^2 =.662 for the model using the three Milankovitch frequencies.

In the most recent 5 sub-intervals [0 -1198 ka], the addition of the obliquity term to the 100 kyr term improves the fit by over 16% in all sub-intervals with the exception of 1032-1198 ka where the fit is improved by 8%. Also, we note that the contribution to squared error is nearly additive [i.e. adding together the R^2 values for the 41 kyr and 100 kyr individually, gives the R^2 for the combination 41 kyr and 100 kyr subset], suggesting that the sinusoidal waves are nearly orthogonal [16]. When the frequencies are different, all sinusoids of infinite lengths will be orthogonal; in this case, we are dealing with sometimes short segments so that orthogonality cannot be assumed.

For 6 of the 8 intervals, the constant term is close to zero as expected after detrending, but not so for the other two intervals. For the most recent 102 kyr, the constant is 0.179. This is

consistent with higher $\delta^{18}O$ levels associated with the well reported greater glacial maximum in this interval. Also, in the interval 2420 ka to 2713 ka, the constant has a value of 0.177, perhaps indicating greater average ice volume in this interval relative to adjoining intervals.

In total, 66% of the total un-weighted squared variation in these data can be fit by the three sinusoidal functions in the indicated eight sub-intervals. As there is clear evidence that the climate behavior at the orbital periods is not linear [2, 79], our simple model undoubtedly underestimates the total fraction of climate variance paced by orbital forcing through the Pliocene and Pleistocene. At the same time, some of the variation explained by the model may be byproduct of the orbital tuning of the data set.

The non-orbitally tuned H07 data set provides us with a different set of change points than does the LR05 data set (Table 2.5). To aid the comparison, we sought the six optimal change points as the seventh change point in the LR05 data set falls outside of the timeframe of the H07 data set. In total, the change point algorithm with six change points for the linear combination of Milankovitch frequencies was able to account for 56.8% of the squared variation in the data, nearly 10% less than in the LR05 data set. This difference may be attributable to orbital tuning and the inclusion of a 41 kyr sinusoid in the model being tested. Two of the six change points are at nearly identical spots, ~788ka and ~1200 ka. The fact that these two change points appear in both data sets strongly indicates that major changes occurred at these times. The biggest difference between the locations of the change points in the two data sets was in their clustering. In the H07 data set, the majority of the six change points are located prior to 1200 ka (Figure 2.8); in the LR05 data set, a majority of the change points are located after 1200 ka. We caution that the two data sets have different sampling densities at 2 Ma and that we cannot rule this out as one possible cause of this discrepancy.

As with the LR05 data set, in H07, we see a step-wise increase in the amplitude of the 100 kyr cycle at 1200 ka and again at ~790 ka. In the H07 data set, after 1200 ka the 41 kyr cycle remains an important component to the system, albeit subordinate to the 100 kyr cycle. Because the data set is not orbitally tuned, we no longer see the phase locking of the 41 kyr cycle that was present in the LR05 data set. However, the change in phase remains small (less than 30 degrees) except between 1700 and 2103 ka. The change point at 1700 ka deserves special attention. It appears that this change point is a result of a phase shift rather than an amplitude change of any of the three model components. It is plausible that this change point arises from age model uncertainty. The interval from 1991 to 2103 ka is an anomaly compared to the sub-intervals that surround it. In this interval, we see the relative weakness of the 41 kyr signal and the emergence of a longer period signal, a combination that is absent from the LR05 stack at this same time period. Accordingly, the change point algorithm picks out this interval as distinct from its neighbors. Finally, as with the LR05 data set, the precessional component is only a minor contributor throughout the time series.

Linear Combinations of Milankovitch Frequencies							
Constraint: Minimum interval width: 100kyr					Phase in degrees (-180 to 180)		
	23 kyr		41 kyr		100 kyr		
Sub-Intervals	Amplitude	Phase	Amplitude	Phase	Amplitude	Phase	Constant (α)
1 to 791	0.0722	7.71	0.1467	50.38	0.434	13.50	-0.0106
792 to 1195	0.0584	72.26	0.1782	19.57	0.1793	2.96	0.0339
1196 to 1700	0.0291	-26.85	0.2521	39.84	0.0431	-52.44	-0.0264
1701 to 1990	0.0273	34.89	0.2003	-36.07	0.0445	37.32	0.0544
1991 to 2103	0.0805	26.96	0.0774	89.58	0.2362	51.11	0.0719
2104 to 2331	0.0146	45.11	0.176	29.45	0.0206	-77.50	-0.0877
2332 to 2580	0.0367	-59.36	0.2101	-6.77	0.1574	76.38	0.0189
	Total R^2 = 0.5680						

Table 2.5: Linear Combinations of Milankovitch Frequencies [H07]. Linear Combinations of Milankovitch frequencies: 23, 41, and 100 kyr. The six change points selected by the algorithm yield seven sub-intervals with corresponding amplitude and phase parameters for each of the three input waves in each sub-interval. The amplitudes are given in $\delta^{18}O$ units and the phase in degrees from -180 to 180. A constraint was imposed that the minimum sub-interval length be at least as long as the longest wave in the analysis (100 kyr).

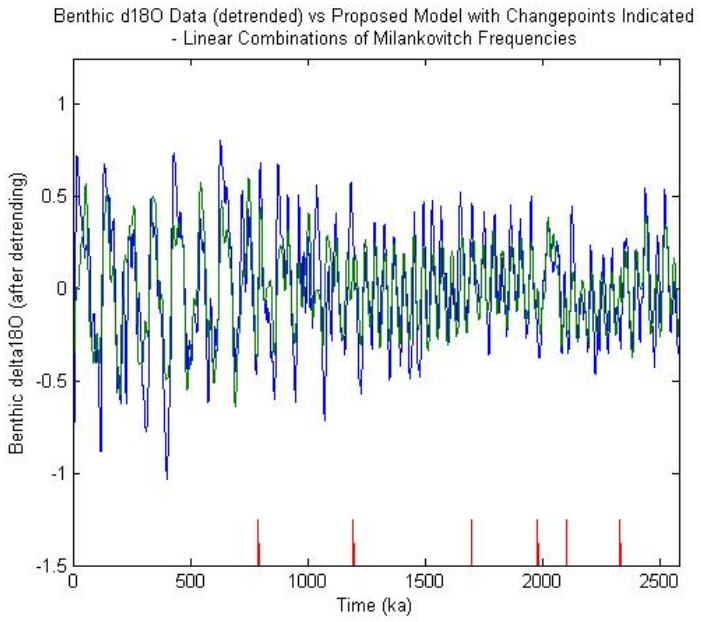

Figure 2.8: Benthic $\delta^{18}O$ Data Plotted vs. Proposed Model with Change Points Indicated – Linear Combinations of Milankovitch Frequencies [H07]. The least squares fit for the linear combination of the three Milankovitch based cycles and a constant term with six change points. The actual data is in blue, the proposed model is in green, and the red spikes at the bottom of the figure represent the optimal locations of change points selected by the algorithm.

Looking more closely at the change points obtained from the H07 data set, we see that half of them fall near magnetic reversals, boundaries which were used to define the H07 age model: 780*, 990, 1070, 1770*, 1950*, and 2580 ka. The only change point in the LR05 time series that coincides with a magnetic reversal is the timing of the onset of the 100 kyr regime at ~790 ka. It thus seems likely that in the case of the H07 data set, the change point method detects shifts in amplitude or phase of glacial cycles which are at least partly an artifact of how the time scale was derived.

2.4.3 The '100 kyr' Cycle Analysis

A recent study [68] suggests that the 100 kyr cycle conventionally attributed to orbital eccentricity (in some form) may stem instead from non-linear climate system responses to the

obliquity forcing. Non-linear interactions of an obliquity response with strong feedbacks, such as the adjustment of large ice sheets and/or the extent of sea-ice, might generate periodicities in glacial cycles near the observed 100 kyr period. In particular, the transition to circa 100 kyr glacial cycles might resemble the "Devil's staircase" transition in non-linear systems [74], where sub-harmonics (rational number multiples of the fundamental period) may dominate the overall response of the system [89].

To examine this question, we compare the ability of three different models to fit both the LR05 and H07 data sets: 1) 'Classic Milankovitch' - 23, 41, and 100 kyr; 2) 'Harmonics of Obliquity' – 41, 82, and 123 kyr; and 3) 'Alternate 100 kyr' – 41, 95, and 124 kyr. The 'Classic Milankovitch' model is the same model that was just discussed. The 95 kyr and 124 kyr components of the 'Alternate 100 kyr' model were chosen because they represent two of the major spectral components of the eccentricity function [9, 10]. The hypothesis that we wish to examine is whether or not the 'Classic Milankovitch' model fits the data as well or better than two alternative models for ice volume. In order to keep the model complexities the same, the number of change points was held fixed at seven and only three sinusoidal components were included in each model. As a result, the precession sinusoid was left out of both the 'Harmonics of Obliquity' and the 'Alternate 100 kyr' models. Our aim is twofold: 1) To determine which model provides the best overall fit to the data; and 2) To look for similarities in the change point locations across the three models in order to discover changes in the time series that are independent of the functions used to model the data.

We first compared the proportion of the variance (R^2) accounted for by each of the competing models on the LR05 data set. When seven change points were used, the 'Harmonics of Obliquity' model was able to account for 66.9% of the squared variation in the data, while the 'Classic Milankovitch' and 'Alternate 100 kyr' models were able to account for 66.4% and 65.6%

of the squared variation, respectively. In essence, all fit the data nearly equally well. From a practical standpoint, these models would appear to be nearly equivalent in their ability to represent the $\delta^{18}O$ data, even though two of the three models leave out precession, a minor but still important component. Inclusion of the precession term in the model increases the R^2 values for the 'Harmonics of Obliquity' and 'Alternate 100 kyr' to 0.699 and 0.684, respectively. One must caution reading too much into the improvement in fit. These models now have two extra parameters [amplitude and phase of the 4^{th} sinusoidal input] with which to fit the data.

We also examined these three models and their fit to the H07 data. As before, in order to keep the model complexities the same, the number of change points was held fixed at six as the seventh change point in LR05 falls outside of the temporal range of the H07 data set. When six change points were used, the 'Harmonics of Obliquity' model was able to account for 64.2% of the squared variation in the data. The 'Classic Milankovitch' and 'Alternate 100 kyr' models were able to account for 56.8% and 55.9% of the squared variation, respectively. In contrast to the LR05 data set, it appears that one of the models, 'Harmonics of Obliquity', substantially out-performs the other two models. Our analysis therefore supports the recent suggestion (Huybers and Wunsch, 2005; Huybers, 2007) that obliquity related terms describe the quasi-periodic components of glacial-interglacial change since the intensification of northern hemisphere glaciation at least as well as the traditional model.

The change point locations for each of the three models on the LR05 data set are as follows: 1) Harmonics of Obliquity – 263, 586, 1030, 1240, 1950, 2272, and 2713 ka; 2) Classic Milankovitch – 102, 380, 786, 1030, 1198, 2418, and 2713 ka; 3) Alternate 100 kyr – 118, 790, 1030, 1202, 1698, 2485, and 2878 ka. One of the most interesting details about the locations of the change points in the three models is the commonalities that exist between the three sets. There is one change point that appears exactly in all three models, 1030 ka, and others that are

similar in at least two of the three models - ~788, ~1200, ~2450, and ~2713 ka. The importance of these times is highlighted by their inclusion in the optimal change point sets across multiple models. Interestingly, these common change point locations occur during the Mid-Pleistocene Transition, between 0.8 Ma and 1.2 Ma, and at the onset of the intensification of Northern Hemisphere glaciation around 2.7 Ma. Again, complete details of these two models with corresponding parameter values can be found in the supplement to [129].

As previously mentioned, the change point locations for each of the three models on the H07 data set were somewhat different from their corresponding models on the LR05 data set: 1) Harmonics of Obliquity – 262, 532, 1104, 1224, 2090, and 2303 ka; 2) Classic Milankovitch – 791, 1195, 1700, 1990, 2103, and 233l ka and 3) Alternate 100 kyr – 794, 1026, 1195, 1699, 1977, and 2337 ka. The change points for the 'Classic Milankovitch' and the 'Alternate 100 kyr' models are very similar to each other on this data set [five of the six change points are within 13 kyr of each other], but different from those of the 'Harmonics of Obliquity' model and different from their counterparts in the LR05 data set. On the other hand, the change points for the 'Harmonics of Obliquity' are similar in both data sets. There are two similarities across the three models on the H07 data set, namely the change points at ~1200 ka and ~2330 ka. Also of note is the change point found ~790 ka in two of the three models. This change point was also found both in the analysis on the LR05 data set and in the analysis on the single dominant frequency.

2.4.4 Evaluation and Conclusions about the $\delta^{18}O$ record

Using the well-established "principle of optimality" [7], we describe a general change point algorithm for the characterization of paleoclimatic time series and illustrate its application to the analysis of $\delta^{18}O$ data for the identification of the earth system's response to Milankovitch forcing. The algorithm's novel characteristics are: 1) globally optimal least squares fitting to a

time series with windows of arbitrary size for a specified number of change points; 2) fitting of time series with combinations of predictor variables; and 3) applicability with any fitting function, linear or non linear. We presented an application of the change point method tailored to evaluating climate change on Myr timescales, but note that the method may be equally viable in the detection of significant changes in climate behavior on other timescales, including those induced by human activity.

The strongest evidence for the utility of this algorithm in paleoclimatology comes from its confirmation of previously identified first-order spectral changes in Plio-Pleistocene $\delta^{18}O$ records, identification of the consistent importance of obliquity forcing throughout the Plio-Pleistocene, and demonstration that obliquity and its sub-harmonics as well as an alternate representation of the empirically derived 100 kyr cycle fit the $\delta^{18}O$ record as well if not better than the traditional Milankovitch model [59]. This last result suggests that the use of the empirical 100 kyr cycle may no longer be required. Using the LR05 data set as an example, the most important change points correspond to the intensification of northern hemisphere glaciation at about 2.7 Ma [134] and to the Mid-Pleistocene onset of large 100 kyr glacial cycles (Tables 2.2-2.4). The 2.7 Ma change point comes not from a switch in the dominant periodicity (the 41 kyr cycle dominates the orbital component of $\delta^{18}O$ change for more than 2 Myr before the change point and for nearly 2 Myr after), but from a significant increase in the amplitude of the 41 kyr component in the $\delta^{18}O$ time series after 2.7 Ma and is associated with a large increase in the amount of variance explained by an orbital model for $\delta^{18}O$ change. Because the time series was weighted by the variance of each sub-interval, the 2.7 Ma change point arises largely from the increased signal-to-noise ratio of the 41 kyr component of benthic $\delta^{18}O$ after 2.7 Ma. Also, it appears that the 23 kyr precessional component of ice volume strengthened significantly

after 2.7 Ma, although it is always subordinate to the 41 kyr obliquity cycle and accounts for only a very small portion of the variance throughout the time series (Table 2.4).

Idealized as the transition from one dominant periodicity of glacial cycles to the next, the Mid-Pleistocene Transition occurred at about 0.788 Ma in both the LR05 and H07 analyses (Tables 2.2, 2.3, and 2.5), on the young side of most published analyses [110, 130]. However, a more complex representation of the $\delta^{18}O$ time series as sums of orbital terms (Table 2.4) displays interesting behavior during the transition from the "41 kyr" world to the "100 kyr" world. Two short (circa 200 kyr) intervals between 0.788 and 1.2 Ma emerge from the LR05 data. These relatively short intervals may correspond to the apparently stepped transition from 41 kyr to 100 kyr dominated cycles detected by Mudelsee and Schulz [104], in which the authors found that an increase in average ice volume preceded the spectral shift to 100 kyr glacial cycles by more than 200 kyr. In our analysis of the earlier (1.2 Ma) break point, the amplitude of the 41 kyr cycle drops significantly at the same time that the strength of the 100 kyr component of glaciation rises steeply. This dip in the obliquity component exists only for a single sub-interval, from 1032 ka to 1198 ka, before returning to its previous level (Table 2.3). The younger bound for the Mid-Pleistocene Transition, 0.788 Ma, corresponds to the time after which the strength of the 100 kyr cycle is nearly constant, a result echoed by the H07 data set. The Mid-Pleistocene Transition is also associated with a step-wise increase in the 23 kyr precessional component (Tables 2.3 and 2.5). However, the 23 kyr precessional sinusoid remains the least important of the three model components in terms of its amplitude [in both data sets] throughout the entire time period analyzed (Tables 2.3 and 2.5).

Change point analysis of LR05 also suggests that climatic break points become more frequent toward the present. If the LR05 age model is robust, our analysis suggests that the climate system lurched more often as the intensity of glaciations increased in the late

Pleistocene. The increasingly unstable pattern of glacial-interglacial cycles, evident by the large number of change points after the Mid-Pleistocene Transition in the LR05 data set, may indicate that the climate has drifted into a progressively more non-linear state over time, and/or that new feedbacks have arisen in the late Pleistocene that significantly alter the frequency and amplitude behavior of the $\delta^{18}O$ time series. Paradoxically, the amount of variance explained by a simple sum of "Milankovitch" sinusoids increases dramatically toward the present (Table 2.4). This observation suggests that the larger amplitude glacial cycles of the late Pleistocene are better fit by finite length sinusoids than their early Pleistocene and Pliocene counterparts. The tendency of internal feedbacks to resonate at or near orbital periodicities may therefore have grown over time (see [2] for examples of how ice age cycle could become phase locked to orbital forcing).

On the other hand, change point analysis of H07 suggests that the time around 2 Ma was more rugged. However, climate related inferences based on the H07 data set are limited because the model did not fit the data as well as it did the LR05 data set and because several of the change points may be artifacts of how the age model was determined (Tables 2.3 and 2.5). When comparing the results obtained from the orbitally tuned LR05 data set to the non-orbitally tuned H07 data set, two change points (788 ka and 1198 ka) appear in both data sets. Further comparisons between the two data sets are hindered by the differing sampling density around 2 Ma and the appearance of numerous changes in the H07 around magnetic reversals.

2.4.5 Caveats and Future Work

The approach described here does have some important limitations. Thus, some caveats are in order. All of the models that were used in this analysis were assumed to be sinusoidal functions based on the idea that climate change is the result of external forcing [i.e solar insolation and

Milankovitch Theory]. However, the use of sinusoidal functions is not new as it is used for Fourier based methods as well. As discussed earlier, the change point algorithm can be adapted to more complex model functions. With respect to the near 100 kyr cycle, it may be that this phenomenon originates from internal feedbacks in the climate system rather than external eccentricity forcing. The 100 kyr term was included in the analysis because of its empirical significance in the $\delta^{18}O$ record.

The change point algorithm in its current form is an example of interrupted regression [96] and thus does not have a continuity constraint. Therefore, the algorithm allows for both gradual and abrupt changes to take place. If continuity is desired, spline regression techniques can instead be used, but the problem then becomes non-linear in its parameters and approximation techniques must instead be used [96]. Typically, the optimization is a very hard problem [84] and dynamic programming is not available to deal with this circumstance.

One of the most important limitations of this algorithm is the lack of a rigorous method to infer the number of change points in a time series, leading to a need to examine the robustness of a study's conclusions against variations in this number. We find our major conclusions are insensitive to small variations in the number of change points [2-3 fewer/more]. Drastic differences in this number would inevitably alter the results. Another important limitation of the present analysis is the absence of information concerning uncertainly in change point locations. With the current algorithm, it is not possible to distinguish between change points whose specific location is strongly supported by the available data, from those in which there may be considerable uncertainty about the exact timing of a change. It is possible that another set or several other sets of change points are almost as good as the optimal set returned by the algorithm. As a consequence, the ability of the algorithm to explore the rapidity of change is limited. Fortunately, a Bayesian approach (described in the next chapter), similar to

that employed by Liu and Lawrence [88], is available to address this limitation, and to draw inferences from multivariate proxy data. By creating a probability distribution on the data and the parameters, we can draw sample solutions in proportion to their probability. Instead of returning the unique global optimum, the Bayesian approach returns an ensemble of solutions from which to draw inferences. The variability of these sampled solutions will give insight into the uncertainty in the number of change points, their locations, and the values of the parameters in the model.

Finally, it is clear that the results of the change point method will depend upon the choice of an initial model for ice age behavior; the present analysis is a point of departure, not a solution of the many enigmas of the ice ages. Our method does allow for a more formal and flexible study of transitions or breaks in time series than previously attempted, with the innovation of flexibility in window sizes and in subsequent modeling functions.

2.5 Special Topic: Testing Milankovitch's Theory

2.5.1 Introduction

More than three decades of research have sought to explain a fundamental observation ubiquitous in paleoclimate time series that span the last 2 Myr, the transition in cyclicity from a dominant 41 kyr beat to a dominant 100 kyr beat that occurs between 1.2 and 0.7 Ma, the so called Mid-Pleistocene Transition [11, 92, 124, 128]. The 100 kyr cycles that dominated the late Pleistocene are puzzling not only because they depart from the 41 kyr cycles that characterized climatic variations for millions of years leading up to the Mid-Pleistocene Transition, but also because their amplitudes and timing are much greater than can be linearly accounted for by the amplitude of the isolation variations that occur in the 100kyr frequency band [71]. This

discrepancy has resulted in a wide array of hypotheses that attempt to account for the striking appearance and amplitudes of the 100 kyr cycle. Some of the proposed mechanisms include forcing by changes in orbital inclination [105, 106, 107], changes in bedrock structure [26], non-linear amplification of eccentricity [59, 71, 72], feedbacks from sea ice [54, 143], Precession 'bundles' [124, 126], and Obliquity 'bundles' [66, 68], where 'bundles' mean that deglaciations occur after an integer number of precession or obliquity cycles respectively. In virtually all explanations for the Mid-Pleistocene Transition, ice sheet changes play a leading role. We aim to test aspects of the dominant theory for explaining the waxing and waning of ice sheets, Milankovitch Theory, which states that ice sheets respond to solar insolation at 65°N in the summer [99].

Depending on its parameters, a dynamic system may be dominated by outside forces applied to it (a forced system) [72, 99, 115] or these outside forces may provide a stimulus that primarily excite the natural frequencies of this system (paced system) [68, 121, 144]. If the Earth's dynamic glacial system is a forced system, then changes in ice volume as represented by the $\delta^{18}O$ proxy record should follow the direct orbital or solar insolation inputs more closely than a sinusoidal approximation to these inputs. On the other hand, a finding that the Earth's dynamic glacial system more closely follows the sinusoidal model indicates that the dynamic system is resonating at its natural frequency.

The forcing versus pacing question has been looked at before. Ruddiman [127] explores the mechanisms associated with obliquity and precession forcing of ice sheets in depth, including feedbacks associated with CO_2 and CH_4 that enhance insolation forcing of ice volume. He invokes a gradual global cooling [121], which allows ice sheets to survive during weak precession insolation maxima, to explain the emergence of 100 kyr glacial cycles at the Mid-Pleistocene Transition. Eccentricity, through its modulation of precession, then acts as a

pacemaker to the glacial system. Maslin and Ridgwell [97] echo the idea that the Mid-Pleistocene Transition was a shift to a glacial system paced, rather than driven, by eccentricity and outline several threshold events that could have caused this transition including: 1) A long term cooling trend caused by a reduction in the concentration of greenhouse gases [104, 124]; 2) The erosion of regolith under the Laurentide ice sheet [26], which would allow ice sheets to rest directly on bedrock, enabling them to take on an equilibrium form with a much greater height; and 3)Long-term cooling of the deep ocean during the Pleistocene, which alters the relationship between atmospheric temperature and the accumulation rates of snow on continental ice sheets, causing more extensive global coverage of sea ice [i.e. the 'sea ice switch'] [54, 143].

2.5.2 The Models

To represent past variations in insolation, we use a set of four plausible models (described below) that are based either directly on variations in Earth's orbital parameters or on the frequencies that dominate variations in the $\delta^{18}O$ record and some of their harmonics. The four models we seek to test are:

1) Direct forcing by solar insolation (Solar Insolation)– If Milankovitch Theory [99] is accurate, then the $\delta^{18}O$ ice volume record should match the June 65°N solar insolation curve after a suitable time delay. Here we examine this assertion by generating not only the June 65°N insolation curve, but insolation curves corresponding to every 5° of latitude and for every month of the year [66]. For each interval, the solar insolation curve which best matches the $\delta^{18}O$ proxy record (in terms of squared error) is chosen to represent that interval.

2) Direct forcing based upon the orbital parameters of motion (Orbital) – Precession, Obliquity, and Eccentricity [10, 69]. Instead of obliquity and precession acting in tandem to produce a specific insolation curve [72], this model finds the optimal linear combination of these three parameters directly.
3) Sinusoidal models representing the largest spectral components in the $\delta^{18}O$ record (Sinusoids). Previous analyses of the $\delta^{18}O$ record have been calculated through Fourier methods which break apart a time series into its periodic components; causality has been argued based upon the coherence (or lack thereof) between the Fourier spectra of the $\delta^{18}O$ record and the orbital parameters [59, 71, 105, 106, 107]. The variance of the $\delta^{18}O$ record is concentrated in three spectral bands: 23, 41, and 100kyr. Specifically we find the best fitting linear combination of sinusoids at these three wavelengths.
4) Sub-Harmonics of Obliquity (Harmonics of Obliquity) – Recent work [66, 68, 89] has suggested that ice sheet dynamics are quantum in nature, responding to 'bundles' of obliquity cycles, where deglaciations are triggered every second or third obliquity cycle. We test this theory by finding the best linear combination of sinusoids at 41, 82, and 123 kyr for each regime of the $\delta^{18}O$ record.

Our results have a single constraint: we require the distance between any two change points to be at least 100 kyr, equivalent to the duration of glacial cycles in the recent past. This constraint keeps the ratio of data points to free parameters from becoming too small. As in Section 2.4, we remove the long-term cooling trend via an exponential function before analyzing the $\delta^{18}O$ proxy record [86] [Figure 2.4] and employ the weighted change point analysis of Section 2.2.2 to account for the substantial increase in variance of the $\delta^{18}O$ record through time.

Finally, because the choice of the optimal number of change points is an unsolved problem in the least squares setting, our conclusions must be checked for robustness across a varying

number of change points. It is clear from the data that there are at least two changes in the $\delta^{18}O$ record over the last 5 Myr – the frequency and amplitude shift associated with the Mid-Pleistocene Transition and the cooling and ice growth during the late Pliocene which resulted in the onset of significant northern hemisphere glaciation at ~2.7Ma. Here, we consider up to 6 change points as this is the number after which the addition of further change points reduces the total squared error by a approximately constant amount (i.e. the 'break in the curve' of Section 2.2.1).

2.5.3 Results

The Least Squares Change Point algorithm was employed to find the best fitting model for each interval of time. Since the algorithm optimizes over all combinations of models and locations of the change points as well as the regression parameters for each model, every change point

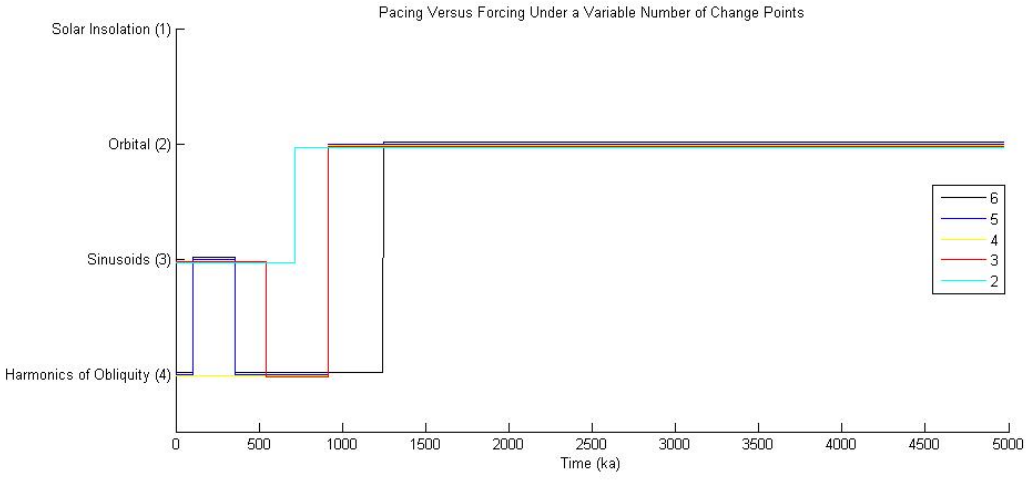

Figure 2.9 Pacing Versus Forcing. Squared error is minimized over the locations of the change points and parameters of the regression model. The specific model chosen at each time point by the change point algorithm is indicated at a varying number of change points. Change points may signify a change in parameter value (i.e. amplitude), a change in model (i.e. forcing function), or both. However, only the changes in model are visible in this figure. The switch from forcing to pacing at the Mid-Pleistocene Transition occurs at either 1.24 Ma, 0.91, or 0.71 Ma, depending on the number of change points.

solution will have a mix of models in the results. We find that the Orbital model (2) best fits the $\delta^{18}O$ record prior to the Mid-Pleistocene Transition over a range of change point values (Figure 2.9). Specifically, it fits the benthic $\delta^{18}O$ curve best prior to 715 ka when 2 change points are utilized, 915 ka when 3-5 change points are utilized and prior to 1.24 Ma when 6 change points are utilized, helping to put a constraint on the timing of the Mid-Pleistocene Transition.

To more fully explore the contributions of each of the four models individually, we examined the change in R^2 values through time for each of the four models tested separately with five change points (Figure 2.10). Since each model selects its own change points, the locations of the change points do not have to correspond. However, they often do occur at similar times (Figure 2.10). For example, the change associated with increased Northern Hemisphere glaciations during the late Pliocene is mutually identified by both the Sinusoids (3) and Harmonics of Obliquity (4) model ~2.6-2.7Ma, while the Solar Insolation (1) and Orbital (2) models jointly identify this change ~3.5Ma. This difference can be attributed to the regime boundary being identified as either the large protrusion in the record around 3.3Ma or the next significant glaciations, which occurred ~2.7-2.8 Ma [Figure 2.4]. Again, we see that the forcing models (1) and (2) fit better prior the Mid-Pleistocene Transition, while the pacing models (3) and (4), fit better after the Mid-Pleistocene Transition. Notice also that the Harmonics of Obliquity (4) model has its R^2 value increase significantly at the Mid-Pleistocene Transition prior to the other three models. Furthermore, the solar insolation model never explains more than 45% of the variation in the data and fits much more poorly than the other three models over the last 1.24 Myrs. Overall, the sinusoidal models, (3) and (4), provide a superior fit compared to the Solar Insolation (1) and the full Orbital model (2) to the entire $\delta^{18}O$ proxy record under a varying number of change points (Table 2.6).

The solar insolation change point analysis permits the identification of the best fitting month and latitude for each segment of the data. Table 2.7 shows that, in each interval, high latitude northern or southern winters best fit the $\delta^{18}O$ proxy record. This result is particularly relevant to the times when the glacial system appears to be orbitally forced, i.e. prior to the Mid-Pleistocene Transition. The structure of the winter insolation curves compared to the

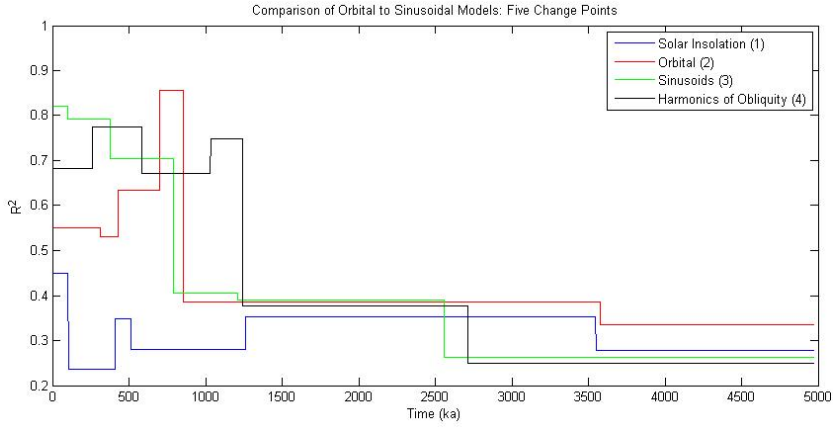

Figure 2.10 Comparison of Orbital to Sinusoidal Models. Each of the four models is individually fit with the change point algorithm. Shown are the R^2 values through time for each model with five change points. The orbital-based (forcing) models provide the best fit until around 1.24 Ma, at which time a transition begins to the superior fit of the sinusoid-based (pacing) models. The transition is completed by 0.74 Ma.

summer insolation curves, rather than their magnitudes is the primary reason for this superior fit. Winter insolation curves have less precession variance than the corresponding summer insolation curve at high latitudes. Since the $\delta^{18}O$ record lacks an obvious precession signal, this result is not surprising. Figure 2.11 compares the fit of the June 65°N insolation curve to the optimal insolation curve (in terms of squared error) during one of the intervals identified by change point analysis on the solar insolation model, 1262 – 3545ka [Table 2.7]. The conflict between precession variance and the $\delta^{18}O$ record is unambiguous [Figure 2.11]. A Partial F test shows that the difference in fit between the two insolation curves is highly significant, with a p-value $p < 10^{-61}$ [Table 2.7]. Each of the other intervals shows similar phenomena, with larger, but still highly significant p-values.

# Change Points	Solar Insolation	Orbital	Sinusoids	Harmonics of Obliquity
2	0.2535	0.3799	0.4762	0.5042
3	0.2854	0.4660	0.5061	0.5742
4	0.2958	0.5145	0.5559	0.6074
5	0.327	0.5380	0.6247	0.6421
6	0.3548	0.5822	0.6486	0.6555

Table 2.6 A Comparison of Overall R^2 Values. Shown is the fit over the entire 5Myr $\delta^{18}O$ proxy record in terms R^2 for each of the four models individually under a varying number of change points.

The $\delta^{18}O$ proxy record was orbitally tuned to June 65°N solar insolation. While orbital tuning can introduce a bias to this analysis, we find that our results were robust to the tuning procedure as the major conclusions about the transition from a forced to a paced glacial system as well as the superior fit of winter rather than summer insolation curves to the $\delta^{18}O$ record remain valid when a non-orbitally tuned version of Lisiecki and Raymo's $\delta^{18}O$ proxy record was analyzed.

Time (ka)	Month	Latitude	Lag (kyr)	R^2 for Optimal	R^2 for June 65°N	P-value
0-104	January	55N	4	0.4495	0.385	.0041
105-410	August	75S	10	0.2353	0.1614	$9.8\text{x}10^{-7}$
411-512	November	65N	1	0.3536	0.0975	$1.1\text{x}10^{-7}$
513-1260	October	75N	9	0.2786	0.1471	$1.0\text{x}10^{-15}$
1262-3545	December	65N	10	0.3517	0.0843	$<10^{-61}$
3550-4965	November	65N	1	0.2778	0.0846	$4.3\text{x}10^{-15}$
			Overall R^2:	0.327	0.1777	$<10^{-77}$

Table 2.7 The Best Insolation Curve for Each Sub-Interval of Time. Insolation curves for each month of the year and every 5° latitude were generated using Huybers program [66]. After fitting all possible insolation curves with a time lag of up to 10 kyr to each segment of the data, the change point algorithm pieced together these segments using dynamic programming. The lag parameter can be interpreted in terms of the lag in the climate system – the time difference between solar forcing and climatic response. Five change points produce six regimes. Indicated is the insolation curve which best fits the $\delta^{18}O$ proxy record in each interval of time and its associated R^2, along with the R^2 value for the June 65°N insolation curve of Milankovitch Theory. The final column gives the p-value for the significance of the improvement in fit. [Derived using a Partial F-Test]

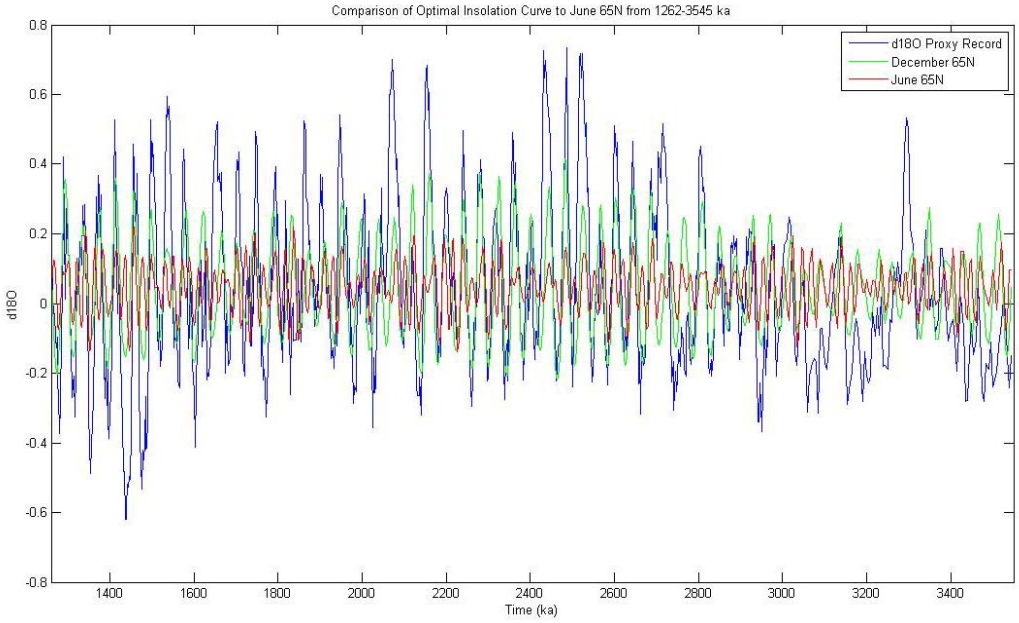

Figure 2.11 A Comparison of June 65°N to the Insolation Curve which Best Fits the $\delta^{18}O$ Proxy Record. Six intervals of time are created by the change point analysis of the solar insolation model (1) (Table 2.7). For one of these intervals, 1262-3545 ka, we compare the fit of the best insolation curve, here December 65°N (green) and the Milankovitch ideal of June 65°N (red) to the $\delta^{18}O$ proxy record during this time. The lack of a precession imprint on the winter insolation curves allow for a better fit (in terms of squared error) to the $\delta^{18}O$ Proxy Record. The difference is highly significant, with a p-value $p < 10^{-61}$. The other five intervals show similar phenomena.

2.5.4 Discussion

Glacial variations can be viewed as the output of a dynamic system whose outside forcing function is represented by solar insolation. Such dynamic systems often have natural frequencies at which they resonate. For example, the response to outside forcing may be similar to the way that a tuning fork responds to a singer's voice. A singer's voice emits a complex signal, but the tuning fork will only respond at its natural frequency. In much the same way, solar insolation is a complex signal to which glacial growth and decay may respond (resonate) primarily at natural frequencies. For many dynamic systems, the parameters of the system are such that the output will be dominated by these natural frequencies and their harmonics. Some types of resonators have the ability to respond to overtones, or harmonics of

their natural frequency, and some non-linear systems respond at sub-harmonics. In the Milankovitch setting, if the natural frequency of ice sheets matches that of obliquity or precession, then sub-harmonics can be represented as obliquity [66, 68] or precession [124, 126] bundles, respectively.

Our finding that the sinusoidal models (3) and (4) fit the $\delta^{18}O$ data substantially better than the full Orbital (2) or Solar Insolation (1) models since the Mid-Pleistocene Transition suggests that in this interval, the parameters of the Earth's glacial system strongly favored resonances at their natural frequencies. During this interval, the timing, rather than the structure of the input (solar insolation) became more important to this paced dynamic system. Several studies have explored this quantum nature of glaciations. Besides the precession and obliquity bundles described previously, Tziperman et al [144] discusses nonlinear phase locking which extends the idea of bundles by claiming that the resonance ratio can change with every glacial cycle. However, they caution against using the fit of a model to prove the validity of a glacial mechanism as non-linear phase locking can 'determine the timing of major deglaciations, nearly independently of the specific mechanism or model that is responsible for these cycles as long as this mechanism is suitably nonlinear'.

The apparent superiority of the orbital models (1) and (2) to the sinusoidal models (3) and (4) prior to the Mid-Pleistocene Transition may indicate the importance of obliquity modulation in the $\delta^{18}O$ proxy record. This would imply a forced dynamic system where the size of ice sheets in addition to the timing of deglaciations is set by solar insolation. This finding also indicates that any resonance phenomena for the dynamic system played a smaller role during this time. The poor fit of the four models to the $\delta^{18}O$ proxy record prior to the Mid-Pleistocene Transition suggests that other factors may exert a stronger influence on $\delta^{18}O$ variations in this interval, including internal ice sheet dynamics.

Given that a solar insolation curve can be approximated by a linear combination of obliquity and precession alone [72], it is not surprising that the richer Orbital model (2), which contains eccentricity along with precession and obliquity, produces a better fit than the Solar Insolation (1) model. Adding parameters to a model always improves its fit to the data. The fit of the Orbital (2) and Solar Insolation (1) models before the Mid-Pleistocene Transition is similar (Figure 2.10), consistent with the absence of 100 kyr glaciations from the $\delta^{18}O$ proxy record during this time, a frequency that the Solar Insolation model cannot linearly generate. The time at which their R^2 values diverge (~1.2Ma) can be viewed as the beginning of the 100 kyr glacial cycles, since the Solar Insolation model lacks the 100 kyr power drawn from the eccentricity function.

Our finding that high latitude winter insolation curves best fit the $\delta^{18}O$ proxy record is a challenge to Milankovitch Theory. The relatively poor fit of the June 65°N insolation curve to the $\delta^{18}O$ proxy record [Table 2.7] seems to be due to the strong presence of precession at this season and latitude compared to the relative absence of precession in both the $\delta^{18}O$ record and high latitude winter insolation curves. Thus, the solar insolation curve which best matches the $\delta^{18}O$ proxy record structurally (i.e. minimum squared error) is the one that contains a reduced precession component. As a physical mechanism, control of ice sheets by high latitude winter insolation seems implausible because the changes in solar radiation through time are minimal. Thus, we cite three possible explanations for the absence of a precession signal from the benthic $\delta^{18}O$ stack.

1. A cancellation of the out-of-phase glacial response to changes in precession [27, 122]. Precession insolation forcing is asymmetric across hemispheres. That is, warm Northern Hemisphere and Southern Hemisphere summers are 180° out of phase with each other. Since ice sheets are assumed to be controlled by summer insolation, this implies that ice

sheets are growing in one hemisphere while they are melting in the other at the precession band, resulting in a minimal net change in ice volume. On the other hand, obliquity insolation forcing is symmetric across hemispheres, with warm summers occurring in both the Northern and Southern Hemispheres at the same time. If the effects of precession cancel in the global $\delta^{18}O$ proxy record, then the only recorded contribution is that of obliquity forcing. Should this mechanism prove correct, then the superior fit of the winter insolation curves is merely an artifact of ice sheets responding to an out of phase precession signal whose effects are absent from the global record (i.e. winter insolation is not driving the system). Additionally, this mechanism could be testable with proxy records that respond to regional rather than global forcing, such as sea surface temperatures.

2. The $\delta^{18}O$ proxy record is a record of global ice volume, whereas Milankovitch was considering only the Northern Hemisphere ice sheets. More generally, Milankovitch Theory cites high latitude summer insolation as important. Perhaps ice volume is sensitive to insolation at more than one location on the globe or more than one time during the year.
    a. Average Annual Insolation – For any latitude, Insolation received over the entire year varies only with obliquity (Loutre et al 2004). Mid-Latitude (45°) average insolation best matches the $\delta^{18}O$ record and can produce an R^2 of 0.396, which is better than any individual insolation curve.
    b. Insolation Gradient – Variations in the summer insolation gradient between high and low latitudes are driven by obliquity [90, 123]. Decreased obliquity causes cooler summer temperatures (poleward of approximately 14° north and south)

which increases the insolation gradient and promotes a pole ward flux of moisture that encourages ice sheet growth.

3. Given that the insolation model is able to account for only a small amount of the variance in the $\delta^{18}O$ proxy record, other more powerful mechanisms for ice sheet dynamics must contribute to this signal. This could include feedback mechanisms [127] or internal dynamics of the ice sheets themselves such as a resonance response to obliquity forcing.

While Table 2.7 gives the best fitting month and latitude for the solar insolation analysis, results since the Mid-Pleistocene Transition should be interpreted with caution since there is clear evidence strongly favoring alternative models. Perhaps a model using multiple solar insolation curves in tandem can improve the fit of an insolation based model.

Chapter 3

The Bayesian Change Point Algorithm

3.1 Introduction

The Bayesian Change Point algorithm is built using the same recursive procedure as its Least Squares counterpart described in Chapter 2. However, a Bayesian approach to the Change Point problem allows you to address one of the major shortcomings of the Least Squares Change Point algorithm – its inability to assess the uncertainty surrounding the optimal solution [88]. Given the probabilistic framework of the Bayesian Change Point algorithm, inferences can now be made about the number of change points, their locations, and the parameters of the model being analyzed. These inferences replace the need for an arbitrary 'break in the curve' decision on the number of change points (Section 2.2.1) and the use of Maximum Likelihood estimates for the parameters of the model. While the chapter is mainly focused on a regression model, the Bayesian Change Point algorithm is applicable to any type of probabilistic model where the probability of the data can be calculated. For example, Liu and Lawrence [88] use a multinomial distribution to model the frequency of nucleotides in a DNA sequence.

Hidden Markov Models (HMMs) can also be used to solve the Change Point problem in a probabilistic setting [78]. The transition from one state to another in a HMM is similar to the placement of a change point, while the emissions of a HMM are representative of the predictions of a regression model. However, the use of HMMs will require an iterative approach due to their use of the Baum-Welch algorithm to estimate the parameters of the regression model, whereas the Bayesian Change Point algorithm can obtain exact posterior distributions in a single pass through the data.

Section 3.2 introduces the Bayesian Change Point regression model and then describes the algorithm in detail. In Section 3.3, we return to the Copy Number Variation simulation that was analyzed in Section 2.3. Finally, Section 3.4 will explore the multinomial model described by Liu and Lawrence [88] in order to highlight the versatility of the algorithm.

3.2: The Bayesian Change Point Algorithm

Given the dependent variable, *Y*, and *m* known predictor variables $X_1, \ldots, X_m$, linear regression methods are based upon the statistical model

$$Y = \sum_{j=1}^{m} \beta_j X_j + \varepsilon$$

where β_j is the j^{th} regression coefficient and ε is a random error term. Suppose that the data set contains *k* change points and let {C} be a set of change points whose locations are $c_0=1, c_1, \ldots, c_k, c_{k+1} = N$, where *N* is the total number of data points. Furthermore, define σ^2 to be the residual variance. The Bayesian Change Point algorithm is concerned with making inferences from the joint distribution

$$f(Y, \beta, \sigma^2, \{C\}|X) = \left[\prod_{i=0}^{k} f\left(Y_{c_i:c_{i+1}}|\beta, \sigma^2, c_i c_{i+1}, X\right) f(\beta|\sigma^2, c_i c_{i+1}, X) f(\sigma^2|c_i c_{i+1}, X)\right] * f(\{C\} = c_0, c_1 \ldots, c_k, c_{k+1})$$

A combinatorial problem arises in the number of possible change point solutions. If there are *N* observations of the dependent variable $Y_1 \ldots Y_N$, and *k* change points in the data, then there are $\binom{N}{k}$ combinations of change point solutions, *{C}*, rendering brute force attempts to compute the joint distribution futile. However, the dynamic programming associated with the Bayesian Change Point algorithm has the ability to reduce the $O(N^k)$ possible sets of change points to a more manageable $O(N^2)$ calculation. The key is to break down the enumeration of all

possible sets of change points into a series of progressively smaller sub-problems, the smallest of which, the placement of a single change point, can easily be solved. The full solution can then be found by efficiently piecing together the solutions to these sub-problems.

To solve the one change point problem, $f(Y_{i:j}) = f(Y_{i:j} \mid X)$ must first be calculated for each and every possible sub-string of the data, $Y_{i:j}$, $1 \le i < j \le N$. Starting from one end of the data set, we can find the probability of any prefix of the data containing one change point by multiplying together the probabilities of two non-overlapping substrings [already calculated] and summing over all possible placements of the change point. Let $F_k(Y_{1:j}) = F_k(Y_{1:j} \mid X)$ be the probability density of the first j observations of the data containing k change points given the regression model. Thus, $F_1(Y_{1:j}) = \sum_{v=1}^{j-1} f(Y_{1:v}) * f(Y_{v+1:j})$. The position of this change point has now been marginalized out; No further information about its location is needed to solve the full problem.

Next, to find the probability density of a prefix with two change points, $F_2(Y_{1:j})$, multiply together the probability density of a prefix containing one change point, $F_1(Y_{1:v})$, and a non-overlapping substring which fills out the rest of the prefix, $F(Y_{v+1:j})$ [both previously calculated] and then sum over all possible placements of the second change point: $F_2(Y_{1:j}) = \sum_{v=1}^{j-1} F_1(Y_{1:v}) * f(Y_{v+1:j})$. Again, the locations of both change points have now been marginalized out as no further information about their positions are needed. The process continues, $F_k(Y_{1:j}) = \sum_{v=1}^{j-1} F_{k-1}(Y_{1:v}) * f(Y_{v+1:j})$, until the full problem is solved. Inferences about the parameters, including the locations of the change points, can be made by sampling from the posterior distribution of the quantities of interest.

Thus, the Bayesian Change Point algorithm has three steps:

1. **Calculating the Probability Density of the Data $f(Y_{i,j} \mid X)$:**

Let X be the (sub)set of regressors included in the regression model. For now, assume that X is fixed (this assumption is relaxed in Chapter 5). Our regression model assumes that the error terms, ε, are independent, mean zero, normally distributed random variables. Therefore, the likelihood function for a substring of the data, $Y_{i:j}$, is $f(Y \mid \beta, \sigma^2, X) \sim N(X\beta, \sigma^2 I)$, where I is the identity matrix. Conjugate priors were chosen for the prior distributions on the vector of amplitudes, β, and the error variance, σ^2. Specifically, β is multivariate normal [$\beta \sim N(0, \sigma^2/k_0)$], and $\sigma^2 \sim Scaled - Inverse\ \chi^2(v_0, \sigma_0^2)$, where k_0 is a scale parameter relating the variance of the regression coefficients to the residual variance, while v_0 and σ_0^2 act as pseudo data points - v_0 pseudo data points of variance σ_0^2. Putting these together, the marginal probability for a substring of the data, $Y_{i:j}$, is:

$f(Y_{i:j}|X) = \iint f(Y_{i:j}|\beta, \sigma^2, X) f(\beta\ |\sigma^2, X)\ f(\sigma^2|X) d\beta\ d\sigma^2$. Let n be the number of data points in a substring, $v_n = v_0 + n$, $\beta^* = (X^T X + k_0 I)^{-1} X^T Y_{i:j}$, $s_n = (Y_{i:j} - X\beta^*)^T (Y_{i:j} - X\beta^*) + k_0 \beta^{*T} \beta^* + v_0 \sigma_0^2$), and let m be the number of model components. Integration yields:

$$f(Y_{i:j}) = f(Y_{i:j}|X) = \frac{(v_0\sigma_0^2/2)^{v_0/2}\ \Gamma(v_n/2)(k_0)^{m/2}}{\Gamma(v_0/2)(s_n/2)^{v_n/2}(2\pi)^{n/2}\ |X^T X + k_0 I|^{1/2}}$$

This quantity is calculated and then stored in memory for all possible substrings of the data, $Y_{i:j}$, with $1 \le i < j \le N$.

2. **Forward Recursion:**

Instead of finding the minimum as in the Least Squares Change Point algorithm (Section 2.2), we take a sum over all possible placements of the final change point. Let $F_k(Y_{1:j})$ be the density of the data [$Y_1 \ldots Y_j$] with k change points. Define $F_1(Y_{1:j}) = \sum_{v<j} f(Y_{1:v}) f(Y_{v+1:j})$ and $F_k(Y_{1:j}) = \sum_{v<j} F_{k-1}(Y_{1:v}) f(Y_{v+1:j})$

3. **Stochastic Backtrace**:

Here, we use Bayes Rule to draw samples from the posterior distribution created by the Forward Recursion step rather than choosing the optimal parameter values as was done for the Least Squares Change Point algorithm (Section 2.2). Each of the posterior distributions described below has an exact representation that is straightforward to sample from. Sampling a set of solutions allows us to address questions related to uncertainty in the number of change points, their locations, and the parameters of the regression model.

In order to have a completely defined partition function (or normalization constant), $f(Y_{1:N})$, two additional quantities need to be specified: 1) a prior distribution on the number of change points, $f(K=k)$; and 2) a prior distribution on the locations of the change points, $f(c_1, c_2, \dots c_k \mid K=k)$. i.e.

$$f(Y_{1:N}) = \sum_{k=0}^{k_{max}} \sum_{c_1 \dots c_k} f(Y_{1:N} \mid K = k, c_1 \dots c_k) * f(c_1 \dots c_k \mid K = k) * f(K = k)$$

The inner summation is exactly the quantity calculated on the Forward Recursion step. As for $f(K=k, c_1, \dots, c_k)$, a priori, we place half the probability mass on zero change points [$f(K=0) = 0.5$] and assume a uniform prior on a positive number of change points, [$f(K=k) = 0.5/(k_{max}+1)$, where k_{max} is the maximal number of allowed change points]. Additionally, we assume that all change point solutions with exactly k change points are equally likely, i.e. $f(c_1, \dots, c_k \mid K = k) = {}^{1}\Big/{\binom{N}{k}}$. This last assumption is a combinatorial prior that directly accounts for the greater number of solutions as the number of change points increases. Together, $f(K=0) = 0.5$ and for $k>0$, $f(K = k, c_1, \dots, c_k) = \frac{0.5}{(k_{max}+1)\binom{N}{k}}$. Since this prior distribution is uniform, it does not have to be included in the Forward Recursion, but instead

can be factored out and multiplied in at this point. However, use of a non-uniform prior must be built in to the Forward Recursion step.

3.1. **Sample a Number of Change Points**:

The Forward Recursion calculates the density of the entire data set, $Y_{1:N}$, given k change points, $F_k(Y_{1:N}) = f(Y_{1:N} | K = k)$. Using Bayes Rule, a posterior distribution on the number of change points given the data can be derived:

$f(K = k|Y_{1:N}) = \frac{F_k(Y_{1:N}) f(c_1,\ldots,c_k|K=k) f(K=k)}{f(Y_{1:N})}$, with $f(Y_{1:N})$ defined as above. This step replaces using the 'break in the curve' to estimate the number of change points (Section 2.2.1).

3.2. **Sample the Locations of the Change Points:**

Additionally, Bayes Rule can be used to assess the uncertainty related to the exact timing of a change. Let $c_{K+1} = t_N$, the last data point. Then for $k=K, K$-1, …, 1,

$c_k \sim \frac{F_{k-1}(Y_{1:v}) f(Y_{v+1:c_{k+1}})}{\sum_{v\in[k-1,c_{k+1})} F_{k-1}(Y_{1:v}) f(Y_{v+1:c_{k+1}})}$. This step replaces the optimal with a probabilistic placement of the change points (Section 2.2).

3.3. **Sample the Regression Parameters for the Interval Between Adjacent Change Points:**

Let $n = c_{k+1} - c_k + 1$, the number of data points in the sub-interval and let $\beta^* = (X^T X + k_0 I)^{-1} X^T Y_{c_k:c_{k+1}}$, $s_n = \left(Y_{c_k:c_{k+1}} - X\beta^*\right)^T \left(Y_{c_k:c_{k+1}} - X\beta^*\right) + k_0 \beta^{*T} \beta^* + v_0 \sigma_0^2$), and $v_n = v_0 + n$. Using Bayes Rule one more time, we find that $f(\beta|\sigma^2) \sim N(\beta^*, (X^T X + k_0 I)^{-1} \sigma^2)$ and $f(\sigma^2) \sim Scaled - Inverse\ \chi^2(v_n\ {}^{s_n}/_{v_n})$. This step replaces Maximum Likelihood estimation of the parameters of the regression model.

Calculating the Probability of the Data, is $O(N^2)$, while the Forward Recursion and Stochastic Backtrace are both $O(kN)$. Therefore, the algorithm has a total time complexity of $O(N^2)$.

3.3: Application: Finding a Change in Mean - Copy Number Variation

Recall the Copy Number Variation simulation from Section 2.3, whose data is reproduced in Figure 3.1. In the least squares setting, the 'break in the curve' was used to determine the proper number of change points for analysis. Here, we seek a Bayesian solution to this problem, which removes the subjective decision regarding the number of change points and allows us to assess the uncertainty surrounding the optimal solution.

Before beginning the analysis, there are three model parameters that need to be chosen: k_0, v_0, and σ_0^2. We assume that the variance of the regression coefficients (σ^2/k_0) can be larger than the error variance, σ^2, so that k_0 is chosen small, for instance 0.1. This choice of

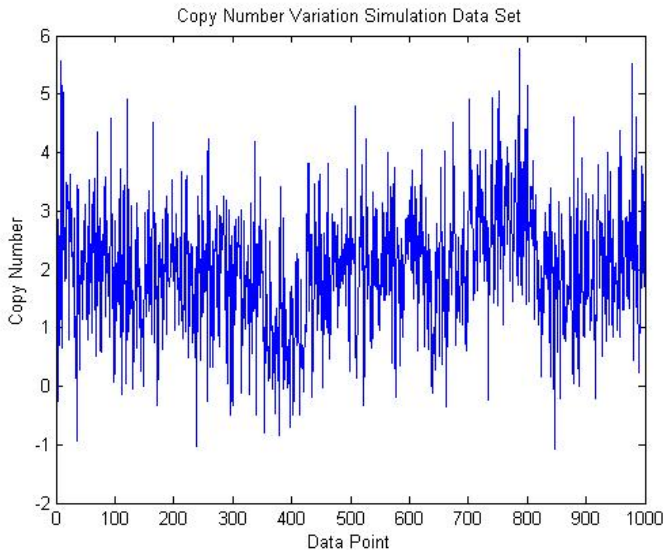

Figure 3.1 Copy Number Variation Simulation. Reproduced from Section 2.3

parameter gives support to a set of plausible regression coefficients. In an attempt to represent prior ignorance about the error variance, we choose the parameters v_0, and σ_0^2 of the *Inverse* χ^2 distribution to be 2 and 1 respectively, representing two pseudo data points each with variance one. Since there are 1000 observations in our data set and each interval generally contains

more than 100 data points, the prior parameters should have little impact on the final solution. Again, we seek to model this data by a constant, $f(x) = C$.

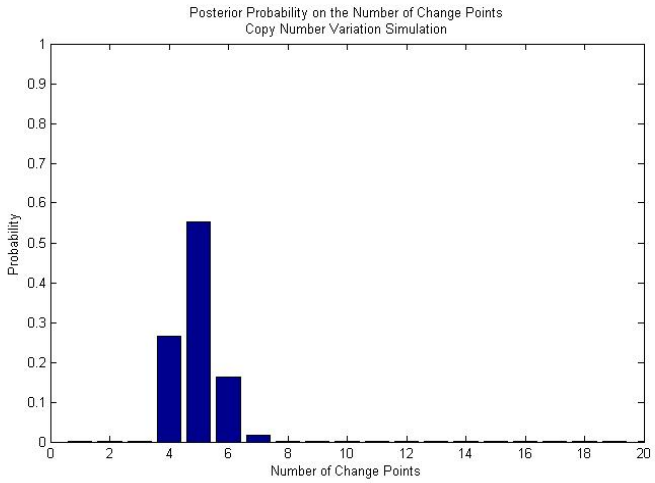

Figure 3.2: Posterior Distribution on the Number of Change Points for the Copy Number Variation Simulation.

After running the Bayesian Change Point algorithm, we find that the posterior distribution on the number of change points is concentrated between 4 and 6, with ~55% of the posterior mass on 5 change points [Figure 3.2]. Figure 3.3 shows that that Bayesian Change Point algorithm identifies all four of the true change points, along with a fifth change point near the beginning of the data set that corresponds to consecutive data points where the added noise was large and positive. Had these data points been located in the middle rather than at one end of the data set, it is likely that the algorithm would have identified these points as noise rather than an additional short segment. Not only was there uncertainty in the number of change points, but in their locations as well. The posterior distribution on the location of the change points is shown at the bottom of Figure 3.3, indicating a small amount of uncertainty (±5 data points) in both directions at each location.

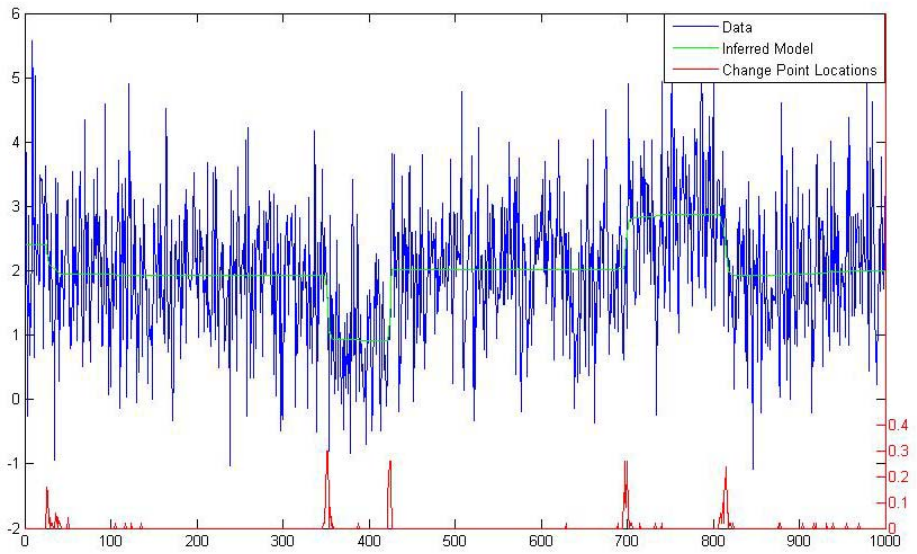

Figure 3.3: Inferred Model and Change Point Locations for the Copy Number Variation Simulation. The original data is colored blue and the inferred model green. The height of the red spikes indicates the posterior probability of a change point at a given location while the width of the red spikes gives an indication of the uncertainty in its location.

3.4: Application to Discrete Data: The Dishonest Casino

Consider a simple game of chance where you place a bet on the outcome of a coin flip. Because each coin flip has only two possible outcomes, 'Heads' or 'Tails', the regression model used in the previous section is not appropriate to model this type of game (i.e. a discrete rather than a continuous process). Instead, the likelihood of any sequence of coin flips (ex. 'HHTTHT') can be modeled as a series of Bernoulli Trials with probability of Heads, p_H, and probability of Tails, 1-p_H = p_T. i.e.

$$\mathrm{L}(\text{'HHTTHT'}) = \mathrm{P}(\mathrm{Y} = \text{'HHTTHT'}|\mathrm{p_H}) = \mathrm{p_H^{n_H} p_T^{n_T}}$$

where n_H is the number of 'Heads' and n_T is the number of 'Tails'. For a fair coin, p_H=p_T=0.5 and all sequences of coin flips are equally likely. Suppose that you record a series of coin flips. Change Point can attempt to answer whether or not the Casino was (always) using a fair coin throughout the day.

3.4.1:Derivation of the Bayesian Change Point Algorithm for Discrete Data

1. **Calculating the Probability Density of the Data *f($Y_{i:j}$ / X)*:**

   Step 1 of the Bayesian Change Point algorithm calculates the probability of every possible substring of the data. For the coin example, since $f(Y_{i:j}) = \int f(Y_{i:j}|p_H) f(p_H) dp_H$, a prior distribution on the nuisance parameters, p_H and p_T, must be chosen that will allow us to integrate these parameters out. Thus, we choose a conjugate prior for Bernoulli trials, the Beta Distribution, which has density:

$$f(p_H;\ \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} (\mathrm{p_H})^{\alpha-1}(1-\mathrm{p_H})^{\beta-1}$$

   where α and β are parameters to be specified. Therefore, the probability of any substring of the data is:

$$f(Y_{i:j}) = \int (\mathrm{p_H})^{\mathrm{n_H}}(1-\mathrm{p_H})^{\mathrm{n_T}} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} (\mathrm{p_H})^{\alpha-1}(1-\mathrm{p_H})^{\beta-1}\ dp_H$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int (\mathrm{p_H})^{\mathrm{n_H}+\alpha-1}(1-\mathrm{p_H})^{\mathrm{n_T}+\beta-1}\ dp_H$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\mathrm{n_H}+\alpha)\Gamma(\mathrm{n_T}+\beta)}{\Gamma(\mathrm{n}+\alpha+\beta)}$$

   where $n=n_H+n_T$ is the total number of coin flips in the substring being considered. If we choose a uniform prior distribution, α = β = 1, then the expression simplifies further to:

$$f(Y_{i:j}) = \frac{\Gamma(\mathrm{n_H}+1)\Gamma(\mathrm{n_T}+1)}{\Gamma(\mathrm{n}+2)} = \frac{n_H!\, n_T!}{(n+1)!}$$

2. **Forward Recursion:**

   The Forward Recursion is identical to that of Section 3.2:

   $F_1(Y_{1:j}) = \sum_{v<j} f(Y_{1:v}) f(Y_{v+1:j})$ and $F_k(Y_{1:j}) = \sum_{v<j} F_{k-1}(Y_{1:v}) f(Y_{v+1:j})$

3. **Stochastic Backtrace:**

We keep the prior distribution on the number of change points the same as in Section 3.2, *f(K=0)* = 0.5 and for *k*>0, $f(K = k, c_1, \dots, c_k) = \frac{0.5}{(k_{max}+1)\binom{N}{k}}$. Of the three sub-steps, only 3.3: Sampling the Parameter Values for the Model, will change from Section 3.2.

3.1. **Sample a Number of Change Points**:

$$f(K = k|Y_{1:N}) = \frac{F_k(Y_{1:N})f(c_1, \dots, c_k|K = k)f(K = k)}{f(Y_{1:N})}$$

3.2. **Sample the Locations of the Change Points**:

Let $c_{K+1} = t_N$, the last data point. Then for *k=K, K*-1, …, 1,

$$c_k \sim \frac{F_{k-1}(Y_{1:v})f(Y_{v+1:c_{k+1}})}{\sum_{v\in[k-1,c_{k+1})} F_{k-1}(Y_{1:v})f(Y_{v+1:c_{k+1}})}.$$

3.3. **Sample the model parameter for the interval between adjacent change points.**

Since we are no longer using a regression, the posterior distribution on the parameter p_H now takes the form of a Beta distribution (with updated parameters) rather than Normal Inverse-χ^2. Here,

$$P(p_H|Y_{i:j}) = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(\mathrm{n_H}+\alpha)\Gamma(\mathrm{n_T}+\beta)}(p_H)^{n_H+\alpha-1}(1-p_H)^{n_T+\beta-1}$$

$$\sim Beta(p_H;\, n_H+\alpha, n_T+\beta)$$

Again, if we consider the uniform prior, α = β = 1,

$$P(p_H|Y) \sim Beta(p_H;\, n_H+1, n_T+1)$$

This type of discrete problem can easily be generalized to more 'interesting' games of chance, for instance: flipping two coins, rolling a pair of dice, or spinning a roulette wheel. In each case, there are more than two possible outcomes for each round of the game. Therefore, the Bernoulli model should be replaced by a product multinomial, and the Beta prior distribution

should be replaced by a Dirichlet prior distribution. A description of the generalized model can be found in [88].

3.4.2: Simulated Games of Chance – A Biased Coin

This example contains a single switch from a fair to a biased coin (i.e. one change point) part way through a sequence of 50 simulated coin flips. A priori, we assume that all change point solutions with exactly *k* change points are equally likely. Furthermore, we assume the uniform Beta prior $\alpha=\beta = 1$ for each of the coins The location of the single change point, *k*, is selected according to a Binomial(50, 0.5) distribution. The first *k* coin flips are made using a fair coin, while the remaining 50-*k* coin flips are made with a biased coin whose probability of heads is 0.8. Let '1' represent 'Heads' and '0' represent 'Tails'. The 50 coin flips are listed below, with the change point occurring at position *k* =29:

Y='01010101011110111011010000100101111111011111111111'

After viewing the sequence of coin flips, one can ask: Is there evidence that more than one coin was used during this game? If so, when did the change occur? And what was the probability of 'Heads' for each coin? Figure 3.4 helps to illustrate the answers to these questions.

The marginal likelihood for the one coin model is $P(Y \mid \textit{one coin}) = 3.98 \times 10^{-15}$, while the marginal likelihood for the two coins model is $P(Y \mid \textit{two coins}) = 1.11 \times 10^{-14}$. Assuming, a priori, that the two models are equally likely, Bayes Rule tells us that the probability that the 50 simulated coin flips were generated by two different coins is $P(\textit{two coins} \mid Y) = 0.736$. Given two coins, Figure 3.4 shows that the posterior distribution of the change point is centered on coin flip number 31. Finally, if we sample values for the probability of 'Heads' from the posterior Beta distribution in each of the resulting sub-intervals and then average the results, we can make an inference about the probability of 'Heads' for each of the two coins. (Alternatively, you

could use the posterior mean of the Beta distributions to make this inference). The coin flips that occurred before the change point appear to have come from a virtually fair coin, while the coin flips that occurred after the change point likely came from a biased coin whose probability of 'Heads' is nearly 0.85. This agrees well with the true probability of 'Heads' for each of the two coins.

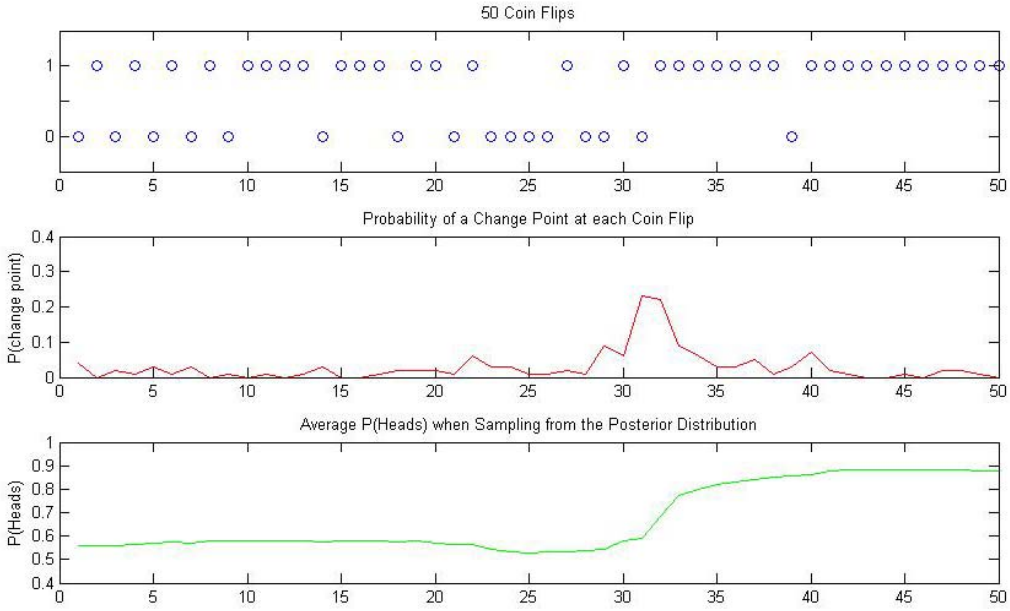

Figure 3.4: 50 Coin Flips. The top pane of the figure shows each of the 50 coin flips, with '1' representing 'Heads' and '0' representing 'Tails'. The middle pane shows the probability of a change in coins at each of the 50 coin flips. The bottom pane displays the average probability of 'Heads' as sampled from the posterior Beta distribution for each coin flip. The increase seen around coin flip 31 coincides with the high probability of a change as indicated in the middle pane.

Chapter 4

Efficient Calculations for Bayesian Variable Selection

4.1 Introduction: A Bayesian Approach to Variable Selection:

Recall the linear regression model introduced in Chapter 3:

$$Y = \sum_{j=1}^{m} \beta_j X_j + \varepsilon$$

where Y is the unknown dependent variable, $X_1, \ldots, X_m$ are m known predictor variables, β_j is the j^{th} regression coefficient and ε is a random error term. When a large number of predictors are available, interest often focuses on the selection of a subset of these variables.

The number of possible sub-models grows exponentially with the number of predictors, thus traditional methods of variable selection quickly become intractable for high dimensional data. Hence, existing methods use a two step approach:

1) A Screening procedure aimed at dimensional reduction through the removal of variables from further consideration.
2) Statistical evaluation of the posterior probability of the remaining sub-models.

Existing algorithms deal with these two aspects of efficiency in various ways.

Several approaches have been developed for variable screening, which include: greedy forward and backward eliminations ([100] and references therein); Forward Regression [150]; Least Angle Regression [40], which is a less greedy version of traditional forward selection methods; the Leaps and Bounds technique [51]; and Sure Independent Screening [SIS] [46], which uses the correlation of predictors to the dependent variable to screen. Typically these screening procedures are employed to reduce the number of predictors to a point that more

rigorous variable selection techniques become feasible, especially for the case when the number of predictors exceeds the number of observations.

Numerous methods also exist for the evaluation step. Ridge Regression [62, 63] has been studied extensively and has been reviewed in [36, 37, 136]. A number of penalized regression approaches have been developed including the nonnegative garrote [17, 100] and Lasso techniques [40, 111, 140]. Additionally, Zou and Hastie [156] developed the elastic net, which is a combination of Ridge Regression and the Lasso technique.

Bayesian approaches focus on finding the posterior distribution across the ensemble of candidate sub-models. Gibbs Sampling, or Stochastic Search Variable Selection (SSVS) [35, 52] as well as Markov Chain Monte Carlo (MCMC) approaches [48, 49, 64, 93, 118] have been introduced to explore these high-dimensional spaces, but provide only an approximate representation of the posterior space and leave open the difficult problem of assessing convergence. Mitchell and Beauchamp [103] found maximum a posteriori (MAP) estimates through an exhaustive search because of the small number of regression terms under consideration; Branch and Bound methods [81, 102] were suggested for larger sets of regression terms. As Fernandez et al [49] points out, the 'largest computational burden lies in the evaluation of the marginal likelihood of each model'. Therefore, the ability to make these calculations efficiently is of the utmost importance.

Bayesian Model Averaging (BMA) [117] uses a branch and bound technique known as Leaps and Bounds [51] to initially screen the solution space. To expedite this process, Leaps and Bounds utilizes an efficient 'matrix sweep' procedure and stores partial results from this step for later use in BMA's evaluation step. The posterior probability of each sub-model that survives the screening process is approximated by and then ranked using the Bayesian Information Criteria (BIC). The space is further truncated by removing any sub-model that does not exceed

an arbitrary bound, typically $1/20^{th}$ of the maximal probability for a sub-model. Iterative BMA (IBMA) [155] was developed in the context of problems where the number of predictors greatly exceeds the number of observations, specifically microarray data. IBMA first ranks variables in order of their individual predictive ability and then successively applies the BMA algorithm to groups of the ordered variables in an attempt to screen out unimportant sets of predictors.

Two classes of exact Bayesian approaches have been described. In the 'spike and slab' model of Mitchell and Beauchamp [103], regression coefficients (β) are modeled by a finite mixture; a point mass at zero was considered the 'spike' while a uniform alternative distribution was considered the 'slab'. In this model, variables assigned to the 'spike' are removed. Alternatively, George and McCulloch [52] use a mixture of normal distributions as their prior distribution on β. Both of the normal distributions are centered at zero, but each has a different variance [$\beta \sim N(0, \sigma^2/k_i)$ *or* $\beta \sim N(0, \sigma^2/k_e)$]. One of the variances is chosen small so that β's drawn from this distribution can 'safely' be replaced by zero with little loss; the other variance parameter is chosen large enough to give support to the set of predictor variables whose coefficients may differ greatly from zero.

These two approaches highlight the distinction between different schools of thought on variable selection: subset selection and 'shrinkage' procedures. In subset selection, regressors are either retained or dropped completely from the model. The results are easily interpretable, but as discrete processes, they can be highly variable as small changes in the data can cause very different models to be selected [140]. Included in the class of subset selection are BMA [117], IBMA [155], MCMC coupled with a Bayesian graphical model [48, 64, 93, 118], 'spike and slab' [103], and stepwise regression techniques [100]. The 'shrinkage' procedures are continuous processes that shrink the regression coefficients of 'uninteresting' regressors towards zero but do not set them to exactly zero. Thus, the results are not as easily interpretable. Included in the

'shrinkage' procedures are Ridge Regression [62, 63], nonnegative garrote [17, 100], and Stochastic Search Variable Selection [SSVS] [52]. The Lasso technique [40, 111, 140] combines elements of both schools of thought as this procedure shrinks some of the coefficients all the way to zero. Both Ridge Regression and the Lasso technique are equivalent to MAP estimators for Bayesian variable selection procedures with Normal and Double Exponential prior distribution on the regression coefficients, respectively [140].

In this Chapter, we describe Exact Bayesian Model Averaging (EBMA), a procedure that uses indicator variables for the selection of one of two classes of variables. Like SSVS, we employ a mixture of normal priors for the regression coefficients, but more efficiently compute the quantities necessary to calculate the posterior probability of sub-models for an exact Bayesian variable selection procedure. Thus, EBMA falls into the category of 'shrinkage' procedures. We also show how EBMA employs a binary tree that is identical to the tree used in the 'sweep' procedure of BMA. As a result, EBMA provides a means to replace approximations to the posterior probability (BIC) with the exact posterior probabilities using an algorithm whose complexity is equivalent to methods that have already been shown to handle problems with a very large number of predictors, specifically, the BMA and IBMA algorithms.

The rest of this chapter is organized as follows. Section 4.2 describes our variable selection model that employs normal priors. In Section 4.3, we describe the EBMA algorithm and prove that its time complexity is $O(m^2)$. Section 4.4 uses simulations as well as the Crime and Punishment data set [41, 42, 145] to compare the speed and accuracy of EBMA to brute force variable selection and MCMC approaches. In order to assess the potential limitations of the BIC approximation used in BMA, we also use the Crime and Punishment data set to compare the results of EBMA to BMA. Section 4.5 consists of discussion and conclusions.

4.2 Calculating the Probability Density of the Data *f(Y)*:

Our regression model assumes that the error terms, ε, are independent, mean zero, normally distributed random variables. Therefore, the likelihood function, $f(Y \mid \beta, \sigma^2, A_m) \sim N(X\beta, \sigma^2 I)$, where X is the matrix of regressors, I is the identity matrix, and A_m is a vector that indicates the included and excluded variables of the model being considered [the dependence on X, the underlying set of possible regressors, will be suppressed in all distributions]. Conjugate priors were chosen for the prior distributions on the vector of amplitudes, β, and the variance, σ^2. Specifically, β is a mixture of multivariate normal distributions [$\beta \sim N(0, \sigma^2/k_i)$ or $\beta \sim N(0, \sigma^2/k_e)$, depending on whether a specific model component is included or excluded from the final model, respectively] and $\sigma^2 \sim Scaled - Inverse\ \chi^2(v_0, \sigma_0^2)$. The marginal probability of the data given a specific sub-model, A_m, is then:

$$f(Y|\ A_m) = \iint f(Y|\beta, \sigma^2, A_m) f(\beta\ |\sigma^2, A_m)\ f(\sigma^2) d\beta\ d\sigma^2\ .$$

Denote N as the number of data points and m as the total number of possible predictors. Let m_i be the number of variables included in sub-model A_m and let m_e be the number of excluded variables [$m_i + m_e = m$]. Associated with the included variables is a 'wide' prior variance parameter k_i and associated with the excluded variables is a 'narrow' prior variance parameter k_e. Let I_{Am} be a diagonal matrix with either k_i or k_e on the diagonal corresponding to whether or not a specific variable is included. Furthermore, define $v_N = v_0 + N$, $\beta^* = \left(X^T X + I_{A_m}\right)^{-1} X^T Y$, and $s_N = (Y - X\beta^*)^T (Y - X\beta^*) + \beta^{*T} I_{A_m} \beta^* + v_0 \sigma_0^2$. Integration yields:

$$f(Y|A_m) = \frac{\left(v_0\sigma_0^2/2\right)^{v_0/2}\ \Gamma\left(v_N/2\right)(k_i)^{m_i/2}(k_e)^{m_e/2}}{\Gamma\left(v_0/2\right)\left(S_N/2\right)^{v_n/2}(2\pi)^{N/2}\ |X^T X +\ I_{A_m}|^{1/2}}$$

The marginal probability of the data, independent of the specific model (Model Averaging), can be obtained by summing over all possible vectors A_m:

$$f(Y) = \sum_{all\ A_m} \iint f(Y\ |\beta, \sigma^2, A_m) f(\beta\ |\sigma^2, A_m)\ f(\sigma^2)\ d\beta d\sigma^2\ f(A_m).$$

Let p_i be the probability of including a predictor ('slab') and let p_e be the probability of excluding a predictor ('spike') [$p_i + p_e = 1$]. If we define the prior probability of a sub-model as a product Bernoulli, $f(A_m) = p_i^{m_i} p_e^{m_e}$, integration yields:

$$f(Y) = \frac{\left(\nu_0 \sigma_0^2 / 2\right)^{\nu_0/2}\ \Gamma\left(\nu_N / 2\right)}{\Gamma\left(\nu_0 / 2\right)(2\pi)^{N/2}} \sum_{All\ A_m} \frac{\left(p_i^2 k_i\right)^{m_i/2} \left(p_e^2 k_e\right)^{m_e/2}}{\left(S_N / 2\right)^{\nu_n/2}\ |X^T X + I_{A_m}|^{1/2}}$$

Only a matrix determinant, matrix inverse (involved in the calculation of s_N), and a count of the number of included (or excluded) model components needs to be completed for each sub-model A_m; the rest of the terms in the density function are independent of A_m.

Each of the matrix operations is naively $O(m^3)$, but can be proven to be on the order of matrix multiplication [19, 32, 148]. To date, matrix multiplication is at best $O(m^{2.376})$ by the Coppersmith-Winograd algorithm [31]. If we pre-compute the quantities Y^TY, X^TX, and X^TY, then the complexity of s_N depends only on the matrix inverse and therefore has a time complexity $O(m^{2.376})$. Together with the determinant and the count on the number of included variables, the overall time complexity for an exact representation of the posterior space is $O(L\ m^{2.376})$, where L is the total number of sub-models to be examined. Space complexity is upper bounded by the space required for the calculations on an individual sub-model [matrix storage – $O(m^2)$, creation of vector β^* – $O(m)$] and any overhead associated with the data set [pre-computed values such as X^TY – $O(m)$, X^TX – $O(m^2)$, etc.], and is therefore $O(m^2)$ per sub-model. The high time complexity of this exhaustive calculation points to the need for efficient matrix calculations if the regression is to be feasible for a large number of predictors.

4.3 Efficient Calculations

Miller [101] provided an efficient way to calculate the inverse of a sum of matrices. Suppose that *G* and *H* are arbitrary nonsingular square matrices of the same dimension and we seek the inverse of the matrix *G +H*, should it too be nonsingular. Miller [101] develops a recursive algorithm to calculate this inverse based upon a fundamental lemma stating that if H is a matrix of rank one, then $(G+H)^{-1} = G^{-1} - \frac{1}{1+g}G^{-1}HG^{-1}$, where g = tr($HG^{-1}$). If the matrix *H* is of rank *r*>1, then we can write $H = E_1 + E_2 + \dots + E_r$, and therefore $G + H = G + E_1 + E_2 + \dots + E_r$ where each E_k, $1 \le k \le r$, has rank one [56], and iteratively apply the lemma [This decomposition is not unique]. Finally, Miller [101] shows that there does exist a decomposition such that each of the 'partial sums' $C_{k+1} = G + E_1 + E_2 + \dots + E_k$ is nonsingular for $k = 1,\dots,r$, ensuring that the lemma can be applied iteratively. His theorem is reproduced below:

> THEOREM: *Let G and G +H be nonsingular matrices and let H have positive rank r. Let* $H = E_1 + E_2 + \dots + E_r$ *where each* E_k *has rank one and* $C_{k+1} = G + E_1 + E_2 + \dots + E_k$ *is nonsingular for k = 1,…,r. Then if* $C_1 = G$,
>
> $$C_{k+1}^{-1} = C_k^{-1} - v_k C_k^{-1} E_k C_k^{-1}, \qquad k = 1, \dots, r$$
>
> *Where*
>
> $$v_k = \frac{1}{1 + tr C_k^{-1} E_k}$$
>
> *In particular*
>
> $$(G+H)^{-1} = C_r^{-1} - v_r C_r^{-1} E_r C_r^{-1}$$

Because of the special structure of our regression problem, our '*H*' matrix is a diagonal matrix whose entries depend on whether a specific variable will be included or excluded in the final model. Therefore, the rank one matrix that we are adding is exactly the matrix associated

with adding one more variable to the current model (or subtracting one variable, depending on whether you start with the null or a full model). To visualize, think of a full binary tree where

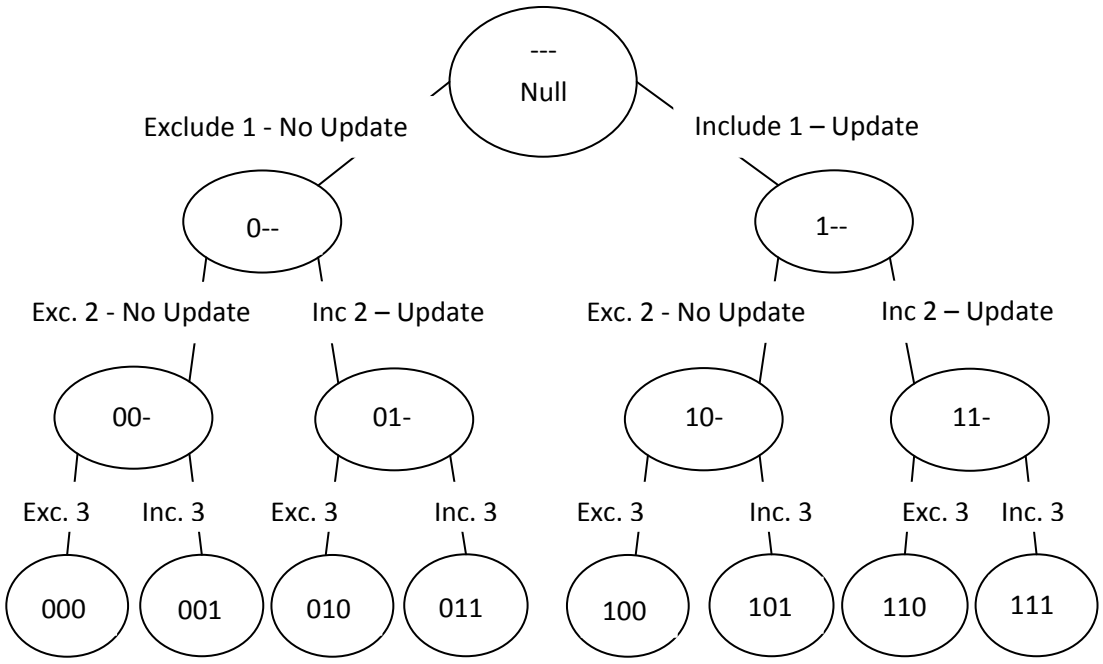

Figure 4.1 A Visualization of the Recursive EBMA Procedure. Given three possible predictors, the binary tree shows how to use the initial matrix, represented by '--- Null', to iteratively generate each of the eight possible sub-models. A '0' represents a specific predictor being excluded, while a '1' represents inclusion by variable selection. Each level of the tree corresponds to a specific predictor and calculations are performed only when the 'Include' branch is traversed. Since each branch of the tree is independent of its sibling, EBMA is easily parallelizable.

one branch signifies 'inclusion' or '1' while the other branch signifies 'exclusion' or '0' and whose depth equals the number of possible predictors. Figure 4.1 provides a visual aid with three possible predictors where the root of the tree consists of the null model. Miller's algorithm allows us to work through the tree, performing a calculation only when the 'inclusion' branch is traversed (or the 'exclusion' branch if you start with the full rather than the null model). Each leaf on the tree holds the probability of one sub-model. We now have only a single matrix inverse and matrix determinant to calculate (rather than one for each possible sub-model) with the iterative procedure providing all additional matrix inversions and determinants necessary.

This binary tree is identical to the one utilized by the Leaps and Bounds procedure [51], differing only in the matrix operations ('Sweep' versus EBMA) performed at each node.

***Theorem**: Let L be the number of sub-models under consideration, N be the length of the data set, and m be the total number of possible predictors. Miller's algorithm in the context of variable selection (EBMA) has time complexity $O(m^2 L)$ and space complexity $O(m^3)$*

In the most general setting, $L=2^m$. However, restrictions on allowable sub-models, such as a cap on the number of included components, can reduce this number.

***Proof**:*

Time Complexity: For examples of the complexities of recursion trees, see [32]. The algorithm can be broken into two parts: Initialization and Recursion. The Recursion can be further partitioned into three stages: Divide [into a number of sub-problems], Conquer [each of the sub-problems], and Combine [solutions to the sub-problems]

i. Initialization: This step includes all calculations that are independent of sub-model choice [i.e. $(2\pi)^{N/2}$ or $\Gamma(\nu_N/2)$ – O(N)] or which need to be done only once [i.e. X^TY used in the calculation of $\beta^* – O(Nm)$, $(X^TX)^{-1}$ and $|X^TX|^{1/2} – O(m^{2.376})$]. All operations are of polynomial order and thus will be dominated by the recursion step.

ii. Recursion

   a. Divide: Each internal node of our recursion tree with two children divides the set of sub-models in half based on whether or not the current predictor is included or excluded. If excluded, no computation needs to be done and the values are simply passed to the next level of the tree – *O(1)*. If included, we need to update our matrix determinant and matrix inverse. In our case,

the iterative calculations of Miller [101] can be further simplified because each of the rank one matrices, E_k, has only one nonzero entry (denoted e_k = the k^{th} position on the diagonal). Therefore, $v_k = \frac{1}{1+C_k^{-1}(k,k)e_k}$, where C_k^{-1} *(k,k)* is the specified position in the matrix C_k^{-1}; $C_k^{-1}E_kC_k^{-1}$ reduces to vector multiplication [e_k * k^{th} column of C_k^{-1} *k^{th} row of C_k^{-1}]. Thus, after the initial matrix inversion and matrix determinant calculation at the root of the tree, each subsequent matrix inverse is an $O(m^2)$ multiplication and an $O(m^2)$ matrix addition, while the matrix determinant is simply an *O(1)* multiplication of constants [$\det C_{r+1} = \frac{1}{v_1 v_2 \ldots v_r} detC_1$]. In total, the divide step is $O(m^2)$

b. Conquer: The set of sub-models under consideration is continually divided in half until we reach the leaves of the recursion tree. Here, we need to compute $f(Y|A_m) \propto \frac{(p_i^2 k_i)^{m_i/2}(p_e^2 k_e)^{m_e/2}}{(S_N/2)^{v_n/2}|X^TX+I_{A_m}|^{1/2}}$. The numerator is *O(m)*, the determinant has already been calculated, and so it remains only to compute $\beta^* = \left(X^TX + I_{A_m}\right)^{-1}X^TY$ and then $s_N = (Y - X\beta^*)^T(Y - X\beta^*) + \beta^{*T}I_{A_m}\beta^* + v_0\sigma_0^2$. The matrix inverse and X^TY have already been computed so β^* is $O(m^2)$, while s_N is $O(m^2) + O(m) + O(1) = O(m^2)$. In total, the conquer step is $O(m^2)$

c. Combine: The cost of combining the sub-problems is *O(1)* if once calculated, the probabilities of each-sub-model are stored in memory.

To find the overall time complexity of the Recursion step, we can add the time spent at each level of the tree. Each internal node is both a Divide and a Combine step. For a tree with *L* leaves, there are *L*-1 internal nodes with two children, a total of $O(m^2L)$. Each leaf on the tree is

a Conquer step. Thus, computation of all leaves costs $O(m^2L)$. In total, the recursion costs $O(m^2L) + O(m^2L) = O(m^2L)$. Since the Initialization is of polynomial order, the entire algorithm has time complexity $O(m^2L)$. Alternatively, if $L=2^m$, define $T(n)$ to be the time complexity of a problem of size n and define the recurrence:

$$T(2^n) = \begin{cases} 2T(2^{n-1}) + O(m^2) & n > 0 \\ O(m^2) & n = 0, i.e. T(1) \end{cases}$$

Solving the recurrence for $T(2^m)$ gives the desired result.

Space Complexity: At Initialization, we create several $O(1)$ variables, the vector X^TY [$O(m)$], as well as the initial matrix [$O(m^2)$] whose inverse [$O(m^2)$] and determinant [$O(1)$] are calculated. In total, $O(m^2)$. Performing a depth-first traversal of the tree requires the storage of one matrix $O(m^2)$, the value of the determinant $O(1)$, and a vector recording the inclusion/exclusion status of the variables $O(m)$ at each level. Temporary variables involved in the updating of the matrix inverse and determinant are maximally $O(m)$. Since the depth of the tree is m, the space requirement of 'divide' is $O(m^3) + O(m) + O(m^2) + O(m) = O(m^3)$. At each leaf, calculation of $f(Y|A_m)$ requires the temporary creation of the $O(m)$ vector β^*. By writing to disk $f(Y|A_m)$ once computed for each sub-model, the total space complexity for 'conquer' is $O(m)$. 'Combine' requires no space. Thus, the overall space complexity for initialization and recursion is $O(m^2)+O(m^3) +O(m)= O(m^3)$. ■

4.4 Simulation Study and Comparison to Existing Methods

To illustrate our algorithm, we have created a simulation that displays the improvement in speed of the recursion versus a brute force calculation. In addition, we use the well-studied Crime and Punishment data set [41, 42, 145] to characterize its compatibility with other models.

In all cases, EBMA was run on Matlab 2008b on a laptop with an Intel® Core™2 Duo CPU 2.26GHz processor with 2 GB of RAM. For this first example, we choose the parameters of the variable selection algorithm as: $p_i = p_e$ [all sub-models equally likely] and k_i =0.01, k_e=100 which is equivalent to the parameter setting $\left(\frac{\sigma_{\beta_i}}{\tau_i}, c_i\right) = (10,100)$ in [52].

4.4.1 Improvements in Speed over Brute Force Regression – A Simulation:

To evaluate the efficiency of the recursion to brute force methods, we compared the amount of time required to enumerate and calculate the probability of all possible sub-models using EBMA and one without the matrix speed-up. The simulation involves a progression from m=12 to 26 potential predictors using N=200 observations, yielding $L=2^{12}$ to 2^{26} possible sub-models. Predictors are obtained as independent standard normal vectors $X_1,\ldots, X_{26}$ iid ~$N_{200}(0,1)$ so that they are essentially uncorrelated. In all cases, the dependent variable, Y, was generated according to the model

$$Y = 10X_1 - 12X_2 - 7X_3 + 5X_4 + 2X_5 - X_6 + \varepsilon$$

where $\varepsilon \sim N_{200}(0, \sigma^2 I)$ with σ = 2. As shown in Table 4.1, the advantage of EBMA grows with

# Variables	Brute Force	EBMA	Fold Reduction
12	0.64	0.31	2.06
14	3.11	1.22	2.55
16	14.54	5.00	2.91
18	70.44	20.60	3.42
20	325.94	84.99	3.84
22	1531.05	347.85	4.40
24	6918.71	1419.45	4.87
26	34260.74	5921.75	5.79

Table 4.1 Improvement in Speed of EBMA over Brute Force Enumeration. Time (in seconds) required to enumerate and calculate the probability of all possible sub-models by brute force (Column 1), or via the recursive algorithm (Column 2). The fold reduction in time of Column 3 compares the timing EBMA to brute force regression. The posterior distributions are identical for the two methods.

the number of variables, resulting in an almost 6-fold reduction in time when 26 variables are considered. Both types of calculations yield an identical posterior distribution on the sub-models.

Suppose that you had a very large number of predictors (>100), but suspected that a smaller number of the predictors were significant. Taking advantage of the recursion, we can eliminate from consideration any sub-model whose number of included variables exceeds some threshold (k_{max}). In doing so, entire branches of the binary tree are pruned, reducing *L* from exponential to $\sum_{k=1}^{k_{max}} \binom{m}{k} \sim O(m^{k_{max}})$. To simulate this type of analysis, we will assume that we have *N*=250 observations and a progression of *m*= 50 to 250 predictors. We 'expect' that 3 or fewer predictors are significant. Again, predictors are represented as independent standard normal vectors $X_1,\ldots, X_{250}$ iid ~$N_{250}(0,1)$ so that they are essentially uncorrelated. The dependant variable, *Y*, is generated according to the model

$$Y = 5X_{17} - 6X_{29} + 3X_{41} + \varepsilon$$

Where $\varepsilon \sim N_{250}(0, \sigma^2 I)$ with $\sigma = 2$. As Figure 4.2 shows, the advantage of EBMA increases rapidly with the number of candidate variables, *m*, and shows an almost 9-fold reduction in time with 250 candidate variables over a brute force approach.

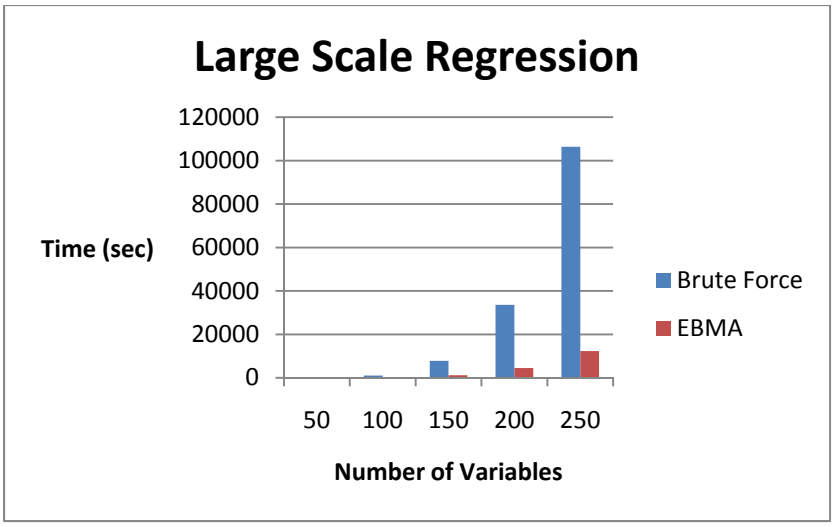

Figure 4.2: Simulation of a Large-Scale Regression. Given a large number of predictors, the simulation assumes that only a small number of predictors are significant. Here, we consider all possible interactions of three or fewer variables for a given number of predictors (*m*). Shown is the amount of time (in seconds) required to compute the probability of all possible sub-models that have three or fewer interacting predictors both by brute force and through the recursion given a data set of size *N* = 250.

4.4.2 Comparison to MCMC Approaches – The Crime and Punishment Data Set

Traditionally, crime was characterized by deviant behavior that was linked to the offender's presumed unique motivation, albeit psychological, social, or family circumstances (For an overview, see [139]). In the late 1960s and early 1970s, this paradigm changed and investigators instead sought to examine the relationship between crime and various measureable quantities such as juvenile delinquency, variations in income and unemployment conditions [50] and the probability and severity of punishment ([41] and references therein). Becker [6] and Stegler [138] argued for an 'economics of crime' - the decision to engage in criminal activity was a rational choice determined by the costs and benefits relative to other (legitimate) opportunities. If criminal activity was the outcome of a rational economic decision, then the probability of punishment should act as a deterrent. Ehrlich [41] developed a theory of the participation in

illegitimate activities, specified it mathematically, and tested it against empirical data obtained in 1960 from 47 states (excluding Hawaii, Alaska, and New Jersey). Errors in Ehrlich's [41] empirical analysis were corrected by Vandaele [145] whose data is used here. No attempt is made to consider the merits of Ehrlich's theory. This data set is used only to compare the results of our variable selection algorithm to that of Raftery et al. [118] and Fernandez et al. [48] who also used this data for illustrative purposes.

Predictor Number	Predictor	k_i=0.01, k_e=100	k_i=0.09, k_e=100	R97	F01	Ehrilch's Models	
1	Percentage of males age 14-24	42	75	79	86		*
2	Indicator variable for southern state	4	12	17	22		
3	Mean years of schooling	71	95	98	99		
4	Police Expenditure in 1960	66	71	72	67		
5	Police Expenditure in 1959	42	49	50	42		
6	Labor Force Participation Rate	3	7	6	15		*
7	Number of Males per 1000 Females	4	8	7	15		
8	State Population	7	19	23	33		
9	Number of Nonwhites per 1000 People	21	60	62	69	*	*
10	Unemployment rate Urban Males, age 14-24	3	10	11	20		*
11	Unemployment rate Urban Males, age 35-39	11	36	45	60		
12	Wealth	13	30	30	31	*	*
13	Income Inequality	99	100	100	100	*	*
14	Probability of Imprisonment	36	75	83	91	*	*
15	Average Time Served in State Prisons	4	18	22	33	*	*

Table 4.2: The Crime and Punishment Data Set. The marginal probability (x100) of including each of the 15 possible regressors in the prediction of the 1960 crime rate across 47 states using two different parameter settings for EBMA, k_i=0.01, k_e=100 and k_i=0.09, k_e=100. Results are compared to the analysis of Raftery et al [118] [denoted R97], Fernandez et al [48] [denoted F01], and the models hypothesized by Ehrlich [41].

The model that we consider is the basic regression model given above, with the dependent variable, *Y*, representing the per capita crime rate. The fifteen possible predictors are shown in Table 4.2. As in the original analysis, all data were log transformed except for the indicator variable for southern states. Given the 15 predictors, there are a total of 2^{15} = 32,768 possible sub-models to explore.

Rank:	Posterior Model Probability (%), k_i=0.01, k_e=100															
1	6.89			3	4									13		
2	6.87	1		3	4									13		
3	5.82				4									13		
4	3.91			3		5								13		
5	3.83					5								13		
6	3.05	1		3	4									13	14	
7	2.84			3	4									13	14	
8	2.38			3		5								13	14	
9	2.38	1		3		5								13		
10	2.30			3	4					9				13	14	
R97, F01	0.40	1		3	4					9		11		13	14	
E1	2.50E-08	1					6			9	10		12	13	14	15
E2	5.40E-06									9			12	13	14	15
Rank:	Posterior Model Probability (%) k_i=0.09, k_e=100															
1	2.95	1		3	4					9				13	14	
2	2.20	1		3	4					9		11		13	14	
3	1.89	1		3	4									13	14	
4	1.57			3	4					9				13	14	
5	1.54	1		3	4							11		13	14	
6	1.53	1		3		5				9				13	14	
7	1.48	1		3	4									13		
8	1.41	1		3	4					9				13	14	15
9	1.24	1		3		5				9		11		13	14	
10	1.23	1		3	4					9			12	13	14	
R97, F01	2.20	1		3	4					9		11		13	14	
E1	4.81E-07	1					6			9	10		12	13	14	15
E2	1.20E-05									9			12	13	14	15

Table 4.3: MAP estimates of the Crime and Punishment Data Set. The top 10 sub-models in terms of their posterior probability (x100) as determined by EBMA under two different parameter settings. For comparison purposes, the top sub-model of Raftery et al [118] [denoted R97] and Fernandez et al [48] [denoted F01], as well as the models hypothesized by Ehrlich [41] [denoted E1, E2], have also been included.

The results of the Crime Data analysis using EBMA are shown in Tables 4.2 and 4.3. Table 4.2 displays the marginal probability of a given regressor being selected for inclusion while Table 4.3 shows the top performing sub-models. As George and McCulloch [52] demonstrate, altering the parameters of the prior distribution associated with inclusion and exclusion places more or less focus on the variables most strongly associated with the response. When

comparing these results to previous analyses on the Crime Data set [48, 118], we find that setting the parameters for prior distribution on β to values used by George and McCulloch [52] [k_i =0.01, k_e=100 equivalent to $\left(\frac{\sigma_{\beta_i}}{\tau_i}, c_i\right) = (10,100)$] push the algorithm to be conservative in its selection for inclusion. Changing these parameters to, for example, k_i =0.09, k_e=100 [equivalent to $\left(\frac{\sigma_{\beta_i}}{\tau_i}, c_i\right) = (10,100/3)$] makes the algorithm less conservative in its selections and provides results more akin to those obtained by both Raftery et al [118] and Fernandez et al [48] [Table 4.2]. In the latter case, the top sub-model of Raftery et al [118] and Fernandez et al (2001a) has the second largest posterior probability while both of Ehrlich's [41] models fail to garner much support [Table 4.3]. Perhaps the most interesting feature is that the entire variable selection algorithm takes a mere ~1.4 seconds to compute all 2^{15} possible sub-models, far outperforming the MCMC approach of Fernandez et al [48] [80 seconds], even after consideration for advances made in computing over the last several years. However, one must keep in mind that the time required for MCMC approaches depends on the threshold of tolerance set for the algorithm which will affect the length of the chain.

4.4.3: Comparison to BMA - The Crime and Punishment Data Set

The Leaps and Bounds procedure [51] employed by BMA uses a tree structure that is identical to the binary tree described in Section 4.3, differing only by the specific ordering of the variables by Leaps and Bounds, which alters the traversal order. The Leaps and Bounds algorithm creates two trees which are traversed in tandem. The root of one tree is the full model while the root of the other tree is the null model. In an attempt to cluster 'good' models near the roots of the trees and find the optimal sub-models of each size as quickly as possible, the variables are ordered by their individual fit to the data. The 'good' models provide sharper bounds for the

algorithm and the quicker that they are found, the more of the sample space that can be truncated by the branch and bound technique.

The move from one sub-model to the next in Leaps and Bounds technique is accomplished via matrix 'sweep' operation akin to Gaussian elimination [51]. The 'sweep' operation, like EBMA, uses one sub-model to quickly derive another and like EBMA, is $O(m^2)$. Additionally, since the 'sweep' operation is completed once for each possible sub-model, it will be done an equivalent number of times as the EBMA posterior probability calculation if the same truncation rules are applied to both procedures. 'Sweeping' can be used to produce a matrix inverse $(X^T X)^{-1}$, as well as the quantities β* and $(Y - X\beta^*)^T(Y - X\beta^*)$ needed for *least squares* regression [quantities utilized by the BMA procedure [117]]. Replacing the 'sweep' operation with the $O(m^2)$ EBMA algorithm described above provides all of the necessary components of the *probabilistic* calculation, including the matrix inverse $\left(X^T X + I_{A_m}\right)^{-1}$ and the matrix determinant $|X^T X + I_{A_m}|^{1/2}$ without increasing the theoretical time complexity.

Leaps and Bounds returns the sum of squared error for each sub-model that it analyzes. BMA converts this squared error to a BIC statistic which approximates the probability of the remaining sub-models given the data. Given the squared error, calculation of the BIC statistic requires a constant number of operations [two logarithms and three $O(1)$ additions and multiplications]. On the other hand, once the EBMA matrix operations are complete, two $O(m)$ vector multiplications and a constant number of other operations [two logarithms, and fewer than a dozen $O(1)$ additions and multiplications, depending on how parameter values are stored in memory] need to be computed to obtain the exact posterior probability. Thus, when same elimination rules are employed, only a small difference in the constants exists between the two procedures.

For example, in a direct comparison on the Crime and Punishment data set, the ratio of the compute time for EBMA to BMA was ~1.5:1, implying that the cost of an exact answer is minimal. Furthermore, these calculations were less than 25% of the total run time of Leaps and Bounds. These specific results are of course dependent on the computer and programming language used, and the efficiency of the written code. Since our exact probabilistic calculation can be done in nearly the same amount of time as the BIC approximation, an approximation of the posterior probability is no longer necessary for a method that has been shown to work for very large problems [43, 155].

Given the set of sub-models produced by the Leaps and Bounds algorithm on the Crime and Punishment data set [117], a further truncation of the sample space is done by BMA based on the BIC statistic. *Symmetric Occam's Window* eliminates all sub-models that are much less likely than the most likely sub-model, by default 20 times less likely. *Strict Occam's Window* further reduces the set of models by eliminating all sub-models with more likely sub-models nested within them. The *Symmetric Occam's Window* leaves 51 possible sub-models covering ~31% of the posterior space, while the *Strict Occam's Window* leaves 14 possible sub-models covering ~19% of the posterior space for the Crime and Punishment data set [117]. Table 4.4 compares the BIC approximation of the probability of each sub-model left by *Strict Occam's Window* to their true posterior probabilities, normalized by the sum of these 14 sub-models. Table 4.4 also provides the probabilistic ranking of each of the sub-models according to the BMA procedure and the posterior probabilities from EBMA. As shown in the table, the model ranks based on BMA do not correspond well with ranks based on exact posterior probabilities as only two have the same rank. Furthermore, the model with the highest posterior probability has BMA rank #9, while the model ranked highest by BMA is ranked #16 based on posterior probability.

BMA Rank	Models														BIC Score	BIC Prob	Exact Prob.	True Rank
1	1		3	4					9		11		13	14	15	-55.9	24.0	5.2	**16**
2	1		3	4					9		11		13	14		-55.4	18.3	11.7	**2**
3	1		3		5				9		11		13	14		-54.4	11.3	6.6	**9**
4	1		3	4					9				13	14	15	-53.8	8.3	7.5	**8**
5	1		3	4							11		13	14		-53.6	7.7	8.2	**5**
6	1		3	4				8	9				13	14		-53.1	5.8	4.5	**20**
7	1		3		5						11		13	14		-52.7	4.9	4.9	**17**
8			3	4				8	9				13	14		-52.4	4.2	5.7	**14**
9	1		3	4					9				13	14		-52.4	4.2	15.6	**1**
10	1		3		5				9				13	14	15	-51.5	2.7	2.5	**43**
11			3		5			8	9				13	14		-51.3	2.4	3.1	**31**
12	1		3		5				9				13	14		-51.2	2.3	8.1	**6**
13	1		3	4							11		13			-50.9	2.0	6.4	**11**
14	1		3	4									13	14		-50.9	1.9	10.0	**3**

Variables															
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	
93	0	100	76	24	0	0	12	84	0	68	0	100	98	35	BIC PMP
91	0	100	75	25	0	0	13	71	0	43	0	100	94	15	Exact PMP
75	12	95	71	49	7	8	19	60	10	36	30	100	83	22	True PMP

Table 4.4: Comparison to BMA. Each row of the table represents one of the 14 (corrected) sub-models selected by *Strict Occam's Window* with its corresponding BIC score and BIC Probability, normalized to only the 14 models shown above [117]. The Exact Probability column gives the probability of the data using EBMA, again normalized to only the 14 models shown above [which account for only ~19% of the posterior probability space]. The True Rank column shows the overall probabilistic rank of the given sub-model from the true distribution when the exact probabilities are calculated on the entire sample space. This column can be compared to the first, which gives the rank of the sub-models according to the BMA procedure. The last three rows of the table give the marginal probability of each variable being selected for inclusion [PMP = Posterior Model Probability]. The 'True PMP' is taken from the analysis done for Table 4.3. All probabilities have been multiplied by 100.

4.5 Discussion and Conclusions:

EBMA provides an exact representation of the posterior space of all possible sub-models via an efficient exact calculation when the number of models is not too large. The algorithm was tested not only on the publicly available Crime and Punishment data set [145], but using simulations as well. The simulation studies showed a significant and increasing reduction in time as the number of predictors grew [Table 4.1, Figure 4.2]. While no claim can be made about the 'correct' sub-model for the Crime and Punishment data, our algorithm was able to

duplicate the results found in Raftery et al [118] and Fernandez et al [48]. Although not discussed in this context, this method could also make a MCMC approach more efficient by traversing the probability space through models that are 'neighbors' as defined in [118].

There are three circumstances in which EBMA may be used to advantage. 1) Any application in which BMA or IBMA has been or could be used to advantage; 2) Problems involving a large number of regression analyses. For example, in change point analysis [3, 88, 129], variable selection is required on every possible substring of the data. In this case, the reduction in compute time can be realized for each of the $N(N+1)/2$ substring calculations that need to be performed on a set of data of length N; 3) Problems involving very large numbers of predictors, in which there is interest in examining posterior distributions for all pairs or all triples, such as in genome studies where epitasis plays a role in phenotype or disease risk. To address the practicality of this application, we simulated a moderate sized problem and project this to the very large size problem that could be run on current parallel computing systems.

Since, any internal node of the binary tree [Figure 4.1] can have its independent 'include' and 'exclude' branches computed separately, this algorithm is readily amenable to parallelization. Thus, when parallel systems are available, the number of models that can be included in exact posterior probability calculations increases nearly linearly with the number of available processors. Setting a hard limit on the number of variables selected for inclusion in an exhaustive search can reduce the time complexity from exponential to polynomial, and thus posterior distributions can be obtained for all permitted models. With current large scale parallel computing systems, EBMA could easily be employed to find the posterior distribution over all combinations of 3 predictors selected from several hundred to a thousand candidate predictors [Figure 4.2]. Computing all pairs of candidate predictors can be completed even faster or for larger problems (~10K candidate variables) since there are fewer to compute. We

encountered no numerical difficulties as our results consistently matched brute force methods. However, it is possible that numerical difficulties specific to this approach may arise in other applications, but we expect that these are unlikely. While we believe that our projection from moderate sized problems to very large problem is reasonable, practical problems could arise in the implementation of this approach on parallel computing systems.

This procedure, like all others, becomes overwhelmed by the exponentially increasing number of sub-models when all combinations of candidate predictors are considered. Thus, a preliminary screening procedure is required. Since EBMA employs the same tree configuration as the 'sweep' procedure of Leaps and Bounds and because it has the same time complexity, $O(m^2)$, exact posteriors can be employed in a Leaps and Bounds based screening procedure with little addition to the computational load. In this setting, our procedure provides a way to efficiently calculate the posterior probabilities of the models selected by the screening step up to the unknown normalizing constant and thus a means to rank the selected models according to their posterior probabilities. Additionally, since BMA (specifically, IBMA which repeatedly uses BMA) has been shown to work for very large problems, EBMA's improvement to this algorithm will therefore also applicable to problems of this size without a substantial difference in time. Of course, since the posterior space is truncated, EBMA will give posterior probabilities only for the remaining portions of the model space.

In Table 4.4, we report differences between BMA and EBMA for the 14 models left after the BMA screening procedure. As these results show, BIC ranks these 14 models in a very different order than the posterior probabilities obtained by EBMA. One possible explanation is that the BIC approximation [117] is based upon a convergence result - an approximation that improves as the number of observations increases relative to the number of predictors.

Alternatively, as with all Bayesian procedures, the posterior distribution is dependent on the prior models and their parameter settings. Thus, the differences in the BMA and EBMA probabilistic rankings may stem from the differences in their prior models. The prior model assumptions of BIC are implicit, and as pointed out by Chen and Chen [23] they can be quite unreasonable. Specifically, the constant prior behind BIC amounts to assigning probabilities for classes of sub-models that are proportional to their sizes. Accordingly, the prior increases almost exponentially in class size. "This is obviously unreasonable, being strongly against the principle of parsimony [23]."

As indicated by George and McCulloch [52], altering the parameters for the 'spike' and the 'slab' can change the set of sub-models with high posterior probability, a result echoed for EBMA by the Crime and Punishment example. Specifically, EBMA is an algorithm in the shrinkage family of variable selection models and in each of these models, the investigator can tune the degree of shrinkage. Thus, the use of this approach requires an investigator to select a range of parameter values near zero that are too small to be on "interest", just as an investigator is required to set the level of type I error that merits further study. For EBMA, the larger the difference between the values of k_e and k_i, the larger the 'penalty' will be for including an additional variable in the model. Further study is required to fully understand the impact of these settings.

One important argument for the 'shrinkage' approach is based on the reasoning of Efron [39] who argues for the use of the terms 'interesting' and 'uninteresting' rather than describing variables in terms of their statistical significance. An investigator may be uninterested either because the discovery of such small differences would be of no practical value or because such small differences could too easily be the result for artifacts such as unobserved confounding or unexpected correlation [39]. Variables that fall in the 'uninteresting' category have their

coefficients close to zero, so that even if they are statistically significant, their effect is too small to be of interest.

We found that Miller's [101] algorithm is well suited to increase the efficiency of variable selection procedures since addition or deletion of a variable is equivalent to adding a rank one matrix, circumventing the most computationally intensive calculations – matrix inversions and determinants. Using the EBMA algorithm, the number of models amenable to exact posterior calculations can now be extended.

Chapter 5

Combining the Bayesian Change Point Algorithm with Variable Selection

5.1: Introduction

The Bayesian Change Point algorithm addresses many of the shortcomings associated with the Least Squares Change Point algorithm. Specifically, a method to address the uncertainty inherent to the inference is now available as the ensemble of sampled solutions can give an indication as to the shape of the posterior space. In the least squares setting, choosing the optimal number of change points is a known hard problem. This difficulty has been replaced by a posterior distribution on the number of change points. Additionally, sampling from the posterior distribution cures the lack of knowledge surrounding the uncertainty in the optimal placement of the change points.

Combining variable selection with the Bayesian Change Point Algorithm further increases the flexibility in statistical modeling. By allowing variables to be added or deleted as suggested by the data, a more realistic model of the underlying phenomenon can be obtained, especially at the boundaries of two regimes. The key to a reasonable run time for this approach lies in an economical variable selection procedure. The recursive nature of the EBMA procedure from Chapter 4 significantly reduces the compute time of the algorithm by efficiently computing the parameters of one sub-model from another.

While this chapter focuses on a regression model that uses sinusoidal functions, the examples from the preceding two chapters show that neither the Bayesian Change Point

algorithm nor EBMA are restricted to these functions. The combined approach can also be adapted to more complex model functions so long as the density function for the residual error can be calculated in every sub-interval for each of the allowable sub-models. Thus, a wide range of possible applications exist.

5.2: The Bayesian Change Point and Variable Selection Algorithm

Recall from Chapter 3 that the initial step of the Bayesian Change Point algorithm is to calculate the probability density of the data $f(Y_{i:j} \mid X)$ for a fixed subset of the regressors, X, and for each possible sub-interval of the time series, $Y_{i:j}$, with $1 \leq i < j \leq N$, where N is the total number of observations. Our regression model assumed that the error terms, ε, are independent, mean zero, normally distributed random variables. Therefore, the likelihood function is $f(Y \mid \beta, \sigma^2, X) \sim N(X\beta, \sigma^2 I)$, where I is the identity matrix. Conjugate priors were chosen for the prior distributions on the vector of amplitudes, β, and the error variance, σ^2. Specifically, β is multivariate normal [$\beta \sim N(0, \sigma^2/k_0)$], and $\sigma^2 \sim Scaled - Inverse\ \chi^2(v_0, \sigma_0^2)$, where k_0 is a scale parameter relating the variance of the regression coefficients to the residual variance, while v_0 and σ_0^2 act as pseudo data points. Thus, the marginal probability of the data was:

$f(Y_{i:j}|X) = \iint f(Y_{i:j}|\beta, \sigma^2, X) f(\beta\ |\sigma^2, X)\ f(\sigma^2|X) d\beta\ d\sigma^2$. Let n be the number of data points in a substring, $v_n = v_0 + n$, $\beta^* = (X^T X + k_0 I)^{-1} X^T Y_{i:j}$, $s_n = \left(Y_{i:j} - X\beta^*\right)^T \left(Y_{i:j} - X\beta^*\right) + \beta^{*T}\beta^* + v_0\sigma_0^2$), and let m be the number of model components. Integration yielded:

$$f(Y_{i:j}) = f(Y_{i:j}|X) = \frac{\left(v_0\sigma_0^2/2\right)^{v_0/2}\ \Gamma\left(v_n/2\right)(k_0)^{m/2}}{\Gamma\left(v_0/2\right)\left(s_n/2\right)^{v_n/2}(2\pi)^{n/2}\ |X^T X + k_0 I|^{1/2}}$$

The goal was to be able to make inferences from the joint distribution $f(Y, \beta, \sigma^2, \{C\}|X)$, where {C} is a set of change points whose locations are $c_0 = 1, c_1, \ldots, c_k, c_{k+1} = N$.

To relax the assumption of a fixed set of regressors, we carry out a variable selection procedure for each possible sub-interval. Define A_m to be a vector that indicates the set of included and excluded variables from the model being considered. Inferences can now be made from a more general joint distribution

$$f(Y, \beta, \sigma^2, \{C\}, A_m)$$

$$= \left[\prod_{i=0}^{k} f\left(Y_{c_i:c_{i+1}} \middle| \beta, \sigma^2, c_i c_{i+1}, A_m\right) f(\beta | \sigma^2, c_i c_{i+1}, A_m) f(\sigma^2 | c_i c_{i+1}, A_m) f(A_m)\right]$$

$$* f(\{C\} = c_0, c_1 \ldots, c_k, c_{k+1})$$

This distribution has the ability to choose a different subset of regressors for each sub-interval. The large number of sub-intervals that exist for any time series [$N(N+1)/2$] require an efficient variable selection procedure in order for the algorithm to be practical. Two distinct approaches exist: 1) Choose the 'best' sub-model for each sub-interval; or 2) Model Averaging (i.e. EBMA). The latter fits better with a Bayesian approach.

The efficient calculation of the exact posterior distribution on the set of predictor variables provided by EBMA [Chapter 4] makes the combined Bayesian Change Point and Variable Selection algorithm feasible. Once the density (probability) of the data has been calculated for each possible sub-model, A_m, the choice of sub-model can be marginalized out: $f(Y) = \sum_{all\ A_m} f(Y\ |A_m)\ * f(A_m)$, where $f(A_m)$ is a product Bernoulli on the number of regressors included in a sub-model, as defined below.

By combining EBMA with the Bayesian Change Point algorithm, the density (probability) of our data now takes the following form: Let m_i be the number of variables included in sub-model A_m and let m_e be the number of excluded variables [$m_i + m_e = m$, total number of variables]. Associated with the included variables is a 'wide' prior variance parameter k_i and associated with the excluded variables is a narrow prior variance parameter k_e. Let I_{Am} be a diagonal matrix with either k_i or k_e on the diagonal corresponding to whether or not a specific

variable is included in the sub-model being considered. Furthermore, define $v_N = v_0 + N$, $\beta^* = \left(X^T X + I_{A_m}\right)^{-1} X^T Y_{i:j}$, and $s_N = \left(Y_{i:j} - X\beta^*\right)^T \left(Y_{i:j} - X\beta^*\right) + \beta^{*T} I_{A_m} \beta^* + v_0 \sigma_0^2$. Finally, let p_i be the probability of including a variable ('slab') and let p_e be the probability of excluding a variable ('spike') [$p_i + p_e = 1$]. The prior distribution on a sub-model, A_m is defined as $f(A_m) = p_i^{m_i} p_e^{m_e}$. Then

$$f\left(Y_{i:j}\right) = \sum_{all\ A_m} \iint f(Y \mid \beta, \sigma^2, A_m) f(\beta \mid \sigma^2, A_m)\, f(\sigma^2)\, d\beta d\sigma^2\, f(A_m).$$

Or after integrating:

$$f\left(Y_{i:j}\right) = \frac{\left({}^{v_0\sigma_0^2}/_2\right)^{{}^{v_0}/_2}\, \Gamma\left({}^{v_n}/_2\right)}{\Gamma\left({}^{v_0}/_2\right)(2\pi)^{{}^{n}/_2}} \sum_{All\ A_m} \frac{\left(p_i^2 k_i\right)^{{}^{m_i}/_2} \left(p_e^2 k_e\right)^{{}^{m_e}/_2}}{\left({}^{S_n}/_2\right)^{{}^{v_n}/_2}\, |X^T X + I_{A_m}|^{{}^{1}/_2}}$$

Again, notice how only a determinant, matrix inverse (involved in the calculation of s_n), and a count of the number of included (or excluded) model components needs to be completed for each sub-model A_m, as the rest of the terms in the density function are independent of A_m.

This calculation replaces Step 1 of the Bayesian Change Point algorithm: Calculating the Probability Density of the Data, which is done for each possible substring of the data. The updated algorithm is outlined below with altered steps indicated by (*):

1. **(*)Calculating the Probability Density of the Data $f(Y_{i,j})$:**

   EBMA is used to perform model averaging for each possible sub-string of the data:

$$f\left(Y_{i:j}\right) = \frac{\left({}^{v_0\sigma_0^2}/_2\right)^{{}^{v_0}/_2}\, \Gamma\left({}^{v_n}/_2\right)}{\Gamma\left({}^{v_0}/_2\right)(2\pi)^{{}^{n}/_2}} \sum_{All\ A_m} \frac{\left(p_i^2 k_i\right)^{{}^{m_i}/_2} \left(p_e^2 k_e\right)^{{}^{m_e}/_2}}{\left({}^{S_n}/_2\right)^{{}^{v_n}/_2}\, |X^T X + I_{A_m}|^{{}^{1}/_2}}$$

2. **Forward Recursion:** As before, let $F_k(Y_{1:j})$ be the density of the data $[Y_1 \ldots Y_j]$ with k change points. Define $F_1\left(Y_{1:j}\right) = \sum_{v<j} f(Y_{1:v}) f(Y_{v+1:j})$ and $F_k\left(Y_{1:j}\right) = \sum_{v<j} F_{k-1}(Y_{1:v}) f(Y_{v+1:j})$

3. **Stochastic Backtrace**:

Not only can we draw samples to evaluate the uncertainty in the number of change points, their locations, and the parameters of the regression model, but we can also now sample to assess the uncertainty of including a given variable in each sub-interval of the data.

To have a completely defined partition function (or normalization constant), $f(Y_{1:N})$, two additional quantities will need to be specified as previously done: 1) a prior distribution on the number of change points, $f(K=k)$; and 2) a prior distribution on the locations of the change points, $f(c_1, c_2, \dots c_k \mid K=k)$. i.e.

$$f(Y_{1:N}) = \sum_{k=0}^{k_{max}} \sum_{c_1 \dots c_k} f(Y_{1:N} \mid K = k, c_1 \dots c_k) * f(K = k, c_1 \dots c_k)$$

The inner summation is exactly the quantity calculated on the forward step of the recursion. As for $f(K=k, c_1, \dots, c_k)$, a priori, we place half the probability mass on zero change points [$f(K=0) = 0.5$] and assume a uniform prior on a positive number of change points, [$f(K=k) = 0.5/(k_{max}+1)$, where k_{max} is the maximal number of allowed change points]. Additionally, we assume that all change point solutions with exactly k change points are equally likely, i.e. $f(c_1, \dots, c_k \mid K = k) = {}^{1}\!/\!_{\binom{N}{k}}$. This last assumption is a combinatorial prior that directly accounts for the greater number of solutions as the number of change points increases. Together, $f(K=0) = 0.5$ and for $k>0$, $f(K = k, c_1, \dots, c_k) = \frac{0.5}{(k_{max}+1)\binom{n}{k}}$. Since this prior distribution is uniform, it does not have to be included in the Forward Recursion, but can instead be factored out and multiplied in at this point. However, use of a non-uniform prior must be built in to the Forward Recursion step

3.1. **Sample a Number of Change Points:**

The Forward Recursion calculates the density of the entire data set, $Y_{1:N}$, given k change points, $F_k(Y_{1:N}) = f(Y_{1:N} \mid K = k)$. Using Bayes Rule, a posterior distribution on the

number of change points given the data can be derived:

$f(K = k|Y_{1:N}) = \frac{F_k(Y_{1:N})f(c_1,\ldots,c_k|K=k)f(K=k)}{f(Y_{1:N})}$, with $f(Y_{1:N})$ defined as above.

3.2. **Sample the Locations of the Change Points:**

Additionally, Bayes Rule can be used to assess the uncertainty related to the exact timing of a change. Let $c_{K+1} = t_N$, the last data point. Then for $k=K, K-1, \ldots, 1$,

$c_k \sim \frac{F_{k-1}(Y_{1:v})f(Y_{v+1:c_{k+1}})}{\sum_{v\in[k-1,c_{k+1})} F_{k-1}(Y_{1:v})f(Y_{v+1:c_{k+1}})}$

3.3. **(*)Sample a Sub-Model for the Interval Between adjacent Change Points:**

Samples are drawn from the posterior distribution on the set of possible sub-models given the data

$$f(A_m|\, Y_{c_k:c_{k+1}}) = \frac{\iint f(Y_{c_k:c_{k+1}}\,|\beta,\sigma^2,A_m)f(\beta\,|\sigma^2,A_m)\,f(\sigma^2)d\beta d\sigma^2\, f(A_m)}{f(Y_{c_k:c_{k+1}})}$$

3.4. **Sample the Regression Parameters for the Interval Between Adjacent Change Points:**

Let $n = c_{k+1} - c_k + 1$, the number of data points in the sub-interval and let $\beta^* = (X^TX + I_{A_m})^{-1}X^TY_{c_k:c_{k+1}}$, $s_N = (Y_{c_k:c_{k+1}} - X\beta^*)^T(Y_{c_k:c_{k+1}} - X\beta^*) + \beta^{*T}I_{A_m}\beta^* + v_0\sigma_0^2$, and $v_n = v_0 + n$. Using Bayes Rule one final time, we find that $f(\beta|\sigma^2) \sim N(\beta^*, (X^TX + I_{A_m})^{-1}\sigma^2)$ and $f(\sigma^2) \sim Scaled - Inverse\ \chi^2(v_n, {}^{s_n}/_{v_n})$

5.3: Proof of Concept: A Simulation Study

Suppose that there are 10 possible predictors $X_1,\ldots,X_{10}$, corresponding to sinusoids of periods 20, 30, 40, 50, 60, 70, 80, 100, 150, and 200, respectively. The dependent variable, *Y*, will consist of 1000 observations randomly generated in the following manner:

1) Select 3-8 change points ~U(3,8)
2) Choose the locations of these change points ~U(1,1000)

3) Select 2-5 predictors to include in each sub-interval ~U(2,5)

4) Generate the regression coefficients for each of the included predictors ~ N(0,2)

5) Add White Noise at three different levels ~N(0,1), N(0, 1.5), and N(0,2)

Using Matlab R2008b's built-in random number generator, the following data set was built:

$$Y_{1:49} = 0.4076X_2 - 2.4401X_3$$
$$Y_{50:362} = -0.062X_5 + 1.3588X_6 - 2.0968X_7$$
$$Y_{363:489} = 0.4495X_3 - 0.8996X_5 - 0.7212X_6 - 1.8691X_7$$
$$Y_{490:614} = 1.6361X_1 + 1.504X_8 + 0.9344X_{10}$$
$$Y_{615:1000} = -0.2275X_1 - 0.5789X_2 - 1.2659X_4 + 0.5789X_{10}$$

The Bayesian Change Point and Variable Selection algorithm was run with the following parameter settings: p_i = 0.5, p_e=0.5, all sub-models equally likely; v_0 = 4, s_0 =0.25, and (k_i, k_e) = (0.01, 100) which is equivalent to the parameter setting $\left(\frac{\sigma_{\beta_i}}{\tau_i}, c_i\right) = (10,100)$ in [52].

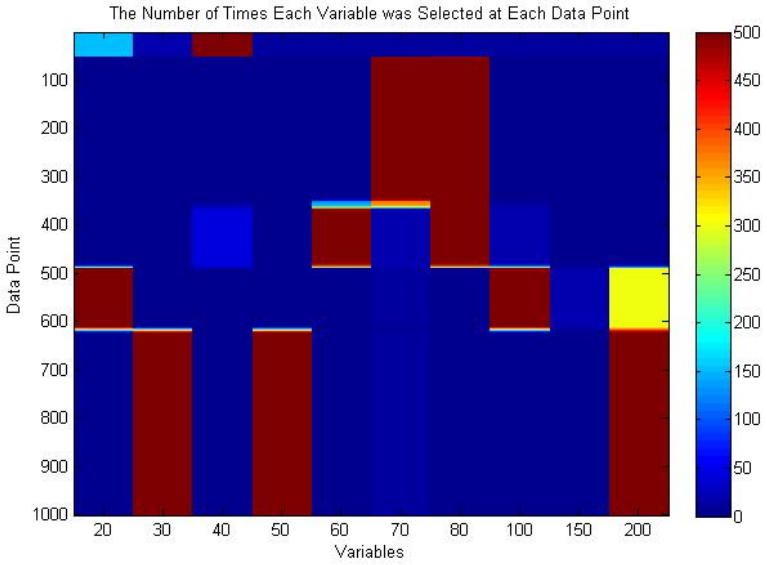

Figure 5.1: Heat Map of Simulation. A heat map indicating the number of times out of the 500 sampled solutions that each variable was selected for inclusion at each data point of the simulation.

Five Hundred samples were drawn from the posterior distribution. The results for a white noise level of 1.0 are shown in Figures 5.1-5.3. The posterior probability of selecting the proper number of change points, 4, was over 99% for white noise levels 1.0 and 1.5, and over 95% for a white noise level of 2.0. Figure 5.1 is a heat map displaying the number of times that

each variable was selected at each data point. As shown in Figure 5.1, the algorithm is conservative in its selection of variables. Variables with small regression coefficients (i.e. X_5 in $Y_{50:362}$) were overwhelmed by the noise in the system and thus not selected by the algorithm. As discussed in Chapter 4, changing the parameters k_i and k_e can alter how conservative the algorithm is in its selection. Figure 5.2 shows the average sampled coefficient for each variable

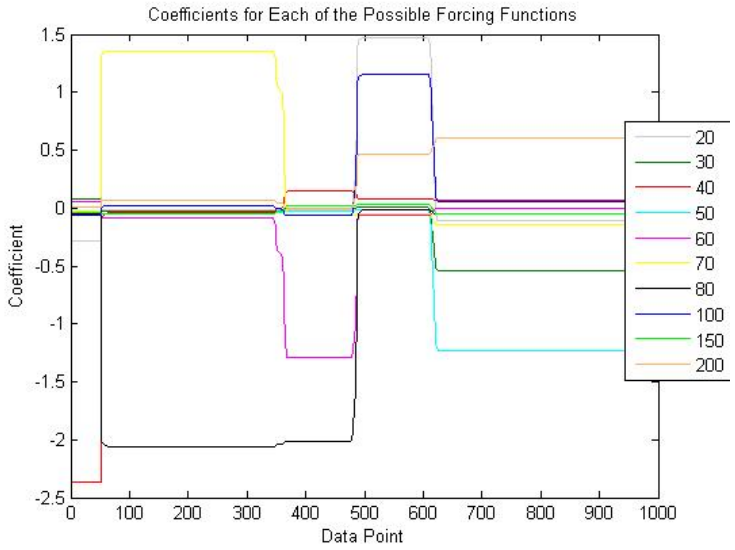

Figure 5.2: Average Coefficients for the Simulation. The average sampled coefficients for each of the 10 variables at each data point in the simulation.

at each data point. These coefficients match well with their true values. Figure 5.3 shows the inferred model in conjunction with the simulated data. The location of the sampled change points is indicated in red at the bottom of the figure. The height of the red 'spikes' is indicative of the number of times a change point was selected (sampled) at that exact location and hence, the probability of a change point at that location. Tall spikes indicate relative certainty (i.e. high probability) in the location of a change point whereas wider 'spikes' indicate a corresponding amount of uncertainty in the location of the change point. Because of the large amount of noise in this simulated data relative to the coefficients of the generating model, the overall R^2 for a white noise level of 1.0 is 0.7188. As the level of noise increases, the change point locations remain accurate, but their distribution becomes more spread out around the true timing of the

change. Similarly, the number of variables selected in each sub-interval decreases with increasing levels of noise as variables with small regression coefficients have the uncertainty surrounding their inclusion grow.

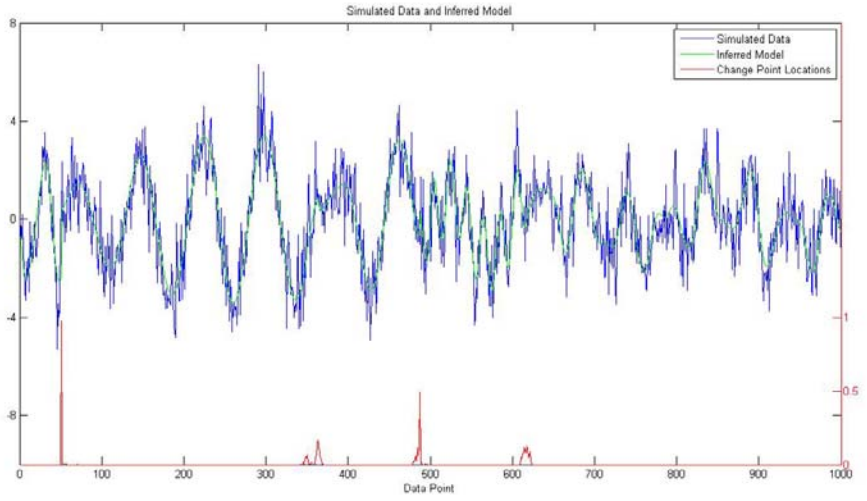

Figure 5.3: The Inferred Model. The simulated data set (blue) is plotted together with the model inferred by the Bayesian Change Point and Variable Selection algorithm (green). The red spikes at the bottom of the figure indicate the proportion of samples that chose a specific data point as a change point.

5.4: Application: $\delta^{18}O$ record of the Plio-Pleistocene

We aim to test several theories of ice sheet formation and destruction through the Bayesian Change Point and Variable Selection Algorithm. Specifically, we seek to compare the following theories using sinusoidal approximations to each of the orbital components:

1. **A Linear Response to Milankovitch Forcing**: Milankovitch forcing is comprised of precession, obliquity, and eccentricity. Precession is represented by a 23 kyr sinusoid, Obliquity by sinusoids at 41 and 53 kyr, and eccentricity by a triplet of sinusoids at 95, 124, and 404 kyr. Each of these frequencies represents the strongest spectral components for each of the orbital forcing functions [9, 10, 69].
2. **Harmonics of Obliquity**: The Harmonics (or Bundles) of Obliquity hypothesis [66, 68, 89] claims that ice sheets terminate with every second or third obliquity cycle after the Mid-

Pleistocene Transition, whereas ice sheets terminated with every obliquity cycle prior to the Mid-Pleistocene Transition. Therefore, the second and third obliquity cycles will be represented by 82 and 123 kyr sinusoids, respectively.

3. **Harmonics of Precession**: The Harmonics (or Bundles) of Precession hypothesis [124, 126] claims that ice sheets grow when summer insolation is unusually low for a full precession cycle and once established, do not last beyond the next increase in summer insolation. Thus, ice sheets will terminate at the fourth or fifth precession cycle. The Harmonics of Precession will be represented by sinusoids at 69, 92, and 115 kyr, corresponding to the third, fourth, and fifth precession cycles, respectively.
4. **Orbital Inclination**: Muller and MacDonald [105, 106, 107] claim that the narrow spectral peak of orbital inclination at 100 kyr is a good match to the narrow spectral peak at 100 kyr in the $\delta^{18}O$ data. Therefore, we represent the Orbital Inclination hypothesis as a single 100 kyr sinusoid.

Because a regression model is being used for this analysis, each of the models will attempt to describe the full pattern of variation in the $\delta^{18}O$ record, rather than describing solely the timing or frequency of deglaciations as in their original formulations.

Based on these four models, the sinusoids used as input to the algorithm are: 23, 41, 53, 69, 82, 92, 95, 100, 115, 123, 124, and 404 kyr. The 123 and 124 kyr sinusoids have frequencies equivalent to three decimal places. As a result, they will be combined into a single sinusoid for this analysis. Furthermore, because each of these hypotheses is exclusive of the others, any sub-models containing sinusoids from two exclusive hypotheses are eliminated from consideration. For example, an allowed sub-model cannot contain both the second harmonic of

obliquity and the fourth harmonic of precession or the 100 kyr inclination sinusoid. Consequently, our 2^{11} possible sub-models are reduced to 144 for each possible sub-interval.

The Bayesian Change Point and Variable Selection algorithm was run with the same parameter settings as the previous example: k_i =0.01 and k_e = 100 (a setting suggested by [52]); v_0 = 4 and σ_0^2 = 0.25 (described below); and p_i = 0.5, p_e = 0.5, indicating all sub-models equally likely. Variations in the values of k_i, k_e, p_i, and p_e affect the probability of a sub-model in a given sub-interval, but otherwise have little impact on the overall number of change points. However, the analysis of the $\delta^{18}O$ proxy record by the Bayesian Change Point and Variable Selection algorithm is highly sensitive to the choice of parameters for the prior distribution of the error variance, v_0 and σ_0^2. These parameters must be chosen carefully to avoid an unbounded likelihood function. Given a data set with a small variance and a model with a large number of free parameters (here, 23 coefficients, one for each of the eleven sine and cosine waves and one constant), the model may be able to fit small segments of the data nearly perfectly under certain parameter settings. If this happens, the density of the data, $f(Y_{i,j})$, can become unbounded and many spurious changes points in close proximity to one another will be observed in the final output. To counteract this phenomenon, penalized likelihood techniques are often implemented [25, 135]. Since the prior distribution on the error variance is akin to adding pseudo-data points of a given residual variance, we have chosen to add a small number of 'large' variance points to neutralize spurious change points of this nature. The key to this choice is to make sure that the product $v_0\sigma_0^2 \geq 1$. This ensures that $(s_n)^{v_n/2}$, where $s_n = \left(Y_{c_k:c_{k+1}} - X\beta^*\right)^T\left(Y_{c_k:c_{k+1}} - X\beta^*\right) + \beta^{*T} I_{A_m}\beta^* + v_0\sigma_0^2$ acts to shrink, rather than enlarge the density function $f(Y_{i,j})$, even when the residual variance (which is related to the variance of substring $Y_{i,j}$) is small. We note that if MAP estimates were being employed by the Bayesian Change Point and Variable Selection algorithm, the procedure outlined above is equivalent to

penalized regression [25, 135]. Additionally, any two change points are required to be the minimum of 50 kyr or 50 data points apart from each other in order to ensure that enough data is available to estimate the parameters of the model accurately.

The results of the Bayesian Change Point and Variable Selection analysis on the $\delta^{18}O$ proxy record are shown in Figures 5.4-5.6. Five hundred samples were drawn from the posterior distribution to give an indication of the shape of posterior space. Figure 5.4 displays the

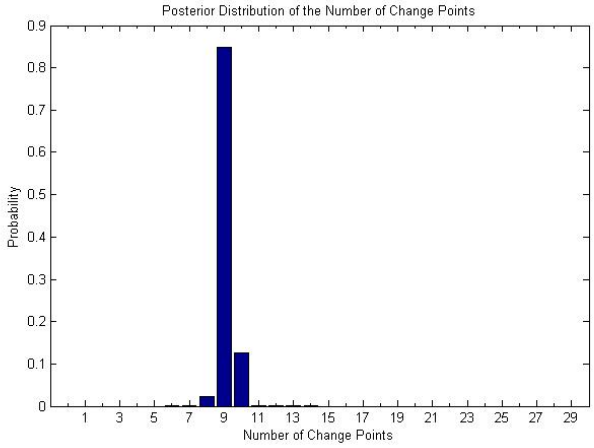

Figure 5.4: Posterior distribution on the number of change points for the $\delta^{18}O$ proxy record

posterior distribution on the number of change points. The posterior probability of nine change points is nearly 0.85, while ten change points are selected 12% of the time. Of the 500 sampled solutions, Figure 5.5 indicates the number of times that each variable was selected at each data point. From this figure, a few conclusions can be drawn:

1) The 41 kyr sinusoid representing obliquity is selected at almost every time point in nearly all of the sampled solutions. In the short interval of time where it is not (135-345ka), the 53 kyr sinusoid, also representing obliquity, is chosen instead roughly two-thirds of the time.
2) Before the onset of permanent glaciers in the Northern Hemisphere (~2.7Ma), only obliquity (41kyr sinusoid) is inferred by any of the 500 samples. Additionally, with the exception of

the 92 kyr sinusoid during 2.0 -2.7Ma, the 41 kyr sinusoid was the only frequency regularly chosen for the entire period prior to the Mid-Pleistocene Transition.

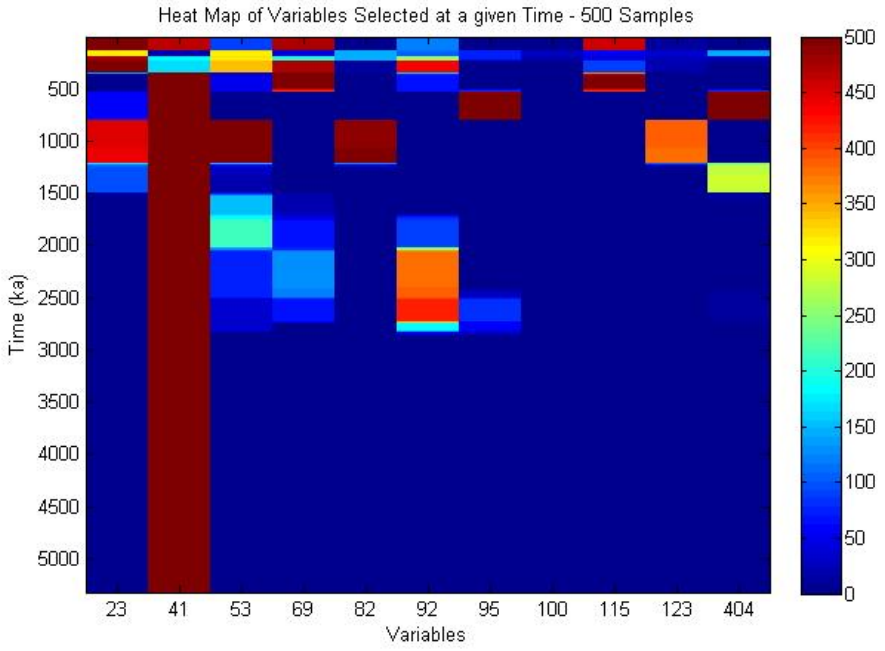

Figure 5.5: Heat map for the $\delta^{18}O$ Proxy Record. A heat map indicating the number of times out of 500 sampled solutions that each variable was selected for inclusion at each data point.

3) The algorithm does not provide a definitive answer as to which of the models being tested is superior to the others after the Mid-Pleistocene Transition (0.8-1.2Ma). Obliquity Bundles are chosen during the transition, while Eccentricity is chosen in one time interval and Precession Bundles in three others after the Mid-Pleistocene Transition. However, the Orbital Inclination model (100 kyr) was not chosen in more than 10% of the samples at any point in time and can therefore be eliminated as a potential ice sheet mechanism.

Figure 5.6 shows the fit of the inferred model (averaged over the 500 samples) to the $\delta^{18}O$ proxy record with the change points locations indicated as red spikes at the bottom of the figure. The height of these spikes indicates the proportion of samples in which the data point was chosen as a change point. The width of these spikes is indicative of the uncertainty in the

timing of a change. Overall, the inferred model was able to account for 85.4% of the variance in the $\delta^{18}O$ data.

5.5 Conclusions:

Simulation results (Figures 5.1-5.3) show a highly accurate placement of change points, but a more conservative approach to variable selection. Of the variables that were chosen, their inferred coefficients were quite similar to their true values. On the other hand, the contributions of some variables were overwhelmed by the added noise and therefore the variables were not selected for inclusion even though they were present, albeit minor, in the data. In general, a longer the interval between change points (i.e. the more data you have) will yield more accurate inferences.

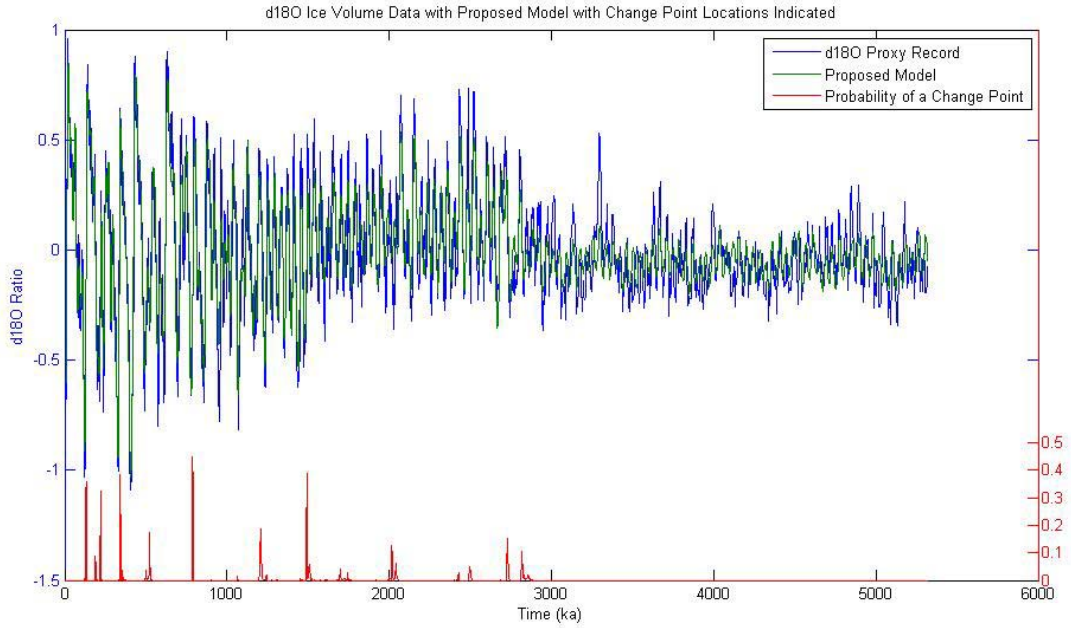

Figure 5.6: The Inferred Model for the $\delta^{18}O$ Proxy Record. The $\delta^{18}O$ proxy record (blue) plotted in conjunction with the model inferred by the Bayesian Change Point and Variable Selection algorithm (green). The red spikes at the bottom of the figure indicate the proportion of the 500 sampled solutions where a given data point was chosen as the location of a change point.

The analysis of the $\delta^{18}O$ proxy record shows the omnipresence of Obliquity, but indecisiveness as to a cause of the recent glacial cycles. However, the algorithm gives strong evidence against the Orbital Inclination theory. Although no one model is superior to the others, on average, nine change points fit over 85% of the variance in the data. Further study is needed to find the true source of change for the Earth's ice sheets.

Chapter 6

Summary of Results and Future Directions

6.1: Summary of Results

This dissertation has explored the waxing and waning of the Earth's ice sheets using highly flexible statistical models – Least Squares Change Point, Bayesian Change Point, and Bayesian Variable Selection (EBMA) – each of which uses dynamic programming to navigate through a very high dimensional solution space.

The first of these models, the Least Squares Change Point algorithm, finds the optimal placement of a given number of changes points, but lacks an objective way to choose the appropriate number. In analyzing the $\delta^{18}O$ proxy record for the single dominant Milankovitch frequency, we find that the change from 41 kyr to 100 kyr glaciations at the Mid-Pleistocene Transition occurs ~798ka. An examination of linear combinations of Milankovitch frequencies reveals that 41 kyr glacial cycles do not disappear at the Mid-Pleistocene Transition, but become buried under a more dominant 100 kyr glacial signal. This analysis also shows that alternate hypotheses to Milankovitch Theory, specifically the 'Harmonics of Obliquity' theory are able to fit the data as well if not better than traditional Milankovitch Theory. A more in depth study of Milankovitch Theory reveals that the Mid-Pleistocene Transition was a time of change from a forced to a paced glacial system. Moreover, during the period of direct orbital control of the ice sheets, solar insolation at high latitudes in the winter match the $\delta^{18}O$ proxy record better than Milankovitch's high latitude summer insolation. Several explanations of this phenomena are proposed, all of which focus on the lack of a precession signal in the proxy record, the main difference between summer and winter insolaiton.

The Bayesian Change Point algorithm eliminates the difficult decision of selecting an appropriate number of change points to analyze faced by the Least Squares Change Point algorithm. In addition, the algorithm now provides a way to assess uncertainty in the number of change points, their locations, and the parameters of the regression model. Conjugate priors are utilized to facilitate the summations and integrations required to calculate the probability of the data given the model. Using the simplest of models, $f(x) = C$, we showed how the Bayesian Change Point algorithm can accurately detect changes in the mean of a continuous random variable. The algorithm takes a slightly different form for discrete variables, allowing us to test for a 'Dishonest Casino' or other, more general types of discrete applications.

EBMA combines a binary tree with an efficient method for calculating the inverse of a sum of matrices in order to reduce the computational complexity [from $O(m^{2.376})$ to $O(m^2)$] of calculating the probability of a sub-model given the data. The computational complexity of EBMA matches that of a dominant approximation technique, the BIC statistic, but can provide an exact representation of the posterior space. Simulations show a significant reduction in compute time over brute force methods, while an analysis of the Crime and Punishment data set illustrates the limitations of the BIC approximation technique. By combining EBMA with a variable screening procedure, problems with a very large number of predictors can be solved.

Combining the Bayesian Change Point algorithm with Variable Selection is, to the best of our knowledge, a novel technique. Brute force attempts to solve either of the problems are impractical. Together, these procedures would be even further out of reach for a brute force approach. The dynamic programming techniques described here bring this method into the realm of practicality, allowing us to define a probabilistic model on an enormous solution space. An analysis of the $\delta^{18}O$ proxy record indicates that before the Mid-Pleistocene Transision, obliquity stands alone as a forcing function for the Earth's glacial system. After the Mid-

Pleistocene Transition, the story is not as clear, as no one theory of ice sheet dynamics dominates the rest.

6.2 Future Directions

The Change Point and Variable Selection algorithms can be extended in several ways to deal with various types of models and data sets. We will briefly mention three of these areas.

6.2.1 New Applications of Change Point

The Change Point algorithm is applicable to any set of sequential data in which you can quantify how well a proposed model matches the data. Here, we have discussed squared error and the probability of the data as two measures of fitness. Other measures of distance or fitness, such as Hamming Distance, are also possible. Thus, a large number of data sets can be analyzed using the change point model. In the past, the Change Point algorithm has been applied to DNA sequences to look for differences in nucleotide proportions across segments of the genome. Currently, the Bayesian Change Point algorithm is being used to study actual Copy Number Variation data sets. A future application of the Change Point algorithm is to study changes in sea surface temperature through time. Unlike the $\delta^{18}O$ proxy record, changes in sea surface temperature are not uniform across the globe. Here, not only will the timing of change points be sought, but whether or not they are concurrent at different locations on the globe.

6.2.2 More Complex Regression Models

The ability to sum and integrate out all of the parameters of a model is essential to being able to calculate the partition function efficiently. Any model in which you cannot integrate out all of

the nuisance parameters, such as one that does not use conjugate priors, will affect your ability to find an exact representation of the posterior space. Additionally, the Change Point algorithm does not have a continuity constraint between adjacent segments. Enforcing continuity disrupts the local nature of the regression parameters being used to fit the data in a given interval. To circumvent these problems, approximation techniques such as MCMC or Gibbs Sampling approaches can be introduced to traverse the posterior space, thus expanding the types of models allowed in a change point analysis.

6.2.3 Expanding Variable Selection

EBMA is limited by the exponentially increasing number of sub-models that comes with an increasing number of predictors. In this situation, any exhaustive algorithm, no matter how efficient, will eventually become intractable. Thus, a variable screening procedure must be developed to be used in conjunction with EBMA for large problems. The Leaps and Bounds procedure is the most obvious choice, as the tree structure utilized by this branch and bound procedure matches the tree structure of EBMA. However, the Leaps and Bounds algorithm truncates the regression tree by eliminating branches that cannot produce the optimal sub-models of a given size. Truncating a tree optimally, while economical, is not the best way to approach Bayesian Variable Selection. Alternatively, the decision to truncate a branch of the regression tree can be made based on the probability of the sub-models that make up that branch. One possible approach is to sample sub-models from a branch in order to obtain a quick estimate of the probability mass associated with that branch before making a truncation decision. While not fully Bayesian, this would at least attempt to account for the entropic effect of large branches and could prevent the truncation of branches which contain sub-models that are near-optimal [i.e. sub-models that have nearly equal probabilities to the optimal models]. If

the algorithm can be scaled to handle a large enough number of predictors, applications in Genome Wide Association Studies are possible.

6.3 Final Thoughts

I have been asked several times how this study impacts the current debate on global warming. The $\delta^{18}O$ proxy record indicates that ice sheets have been melting and reforming for millions of years. I like to tell people that while man is not the sole cause of the Earth's melting glaciers, we may be speeding up the process.

Along the same lines, I have also been asked about this model's ability to make predictions about the future. As indicated by our study, the Earth's glacial system is sufficiently complex that we do not yet have an appropriate way to model its activity; no one theory of ice sheet variability was shown to be superior to the others. Further study is needed for a better understanding of ice sheet dynamics. After reminding an audience that data points are spaced 1000 years apart, I tell them that of course the model can make a prediction, but that it is useless for the current global warming debate, which focuses on a much shorter time scale. On the other hand, the model's prediction for the future would not be proven right or wrong until long after we are gone.

References

[1] H. Aksoy et al. Fast segmentation algorithms for long hydrometeorological time series. *Hydrological Processes*, **22**(23): 4600-4608, 2008.

[2] Y. Ashkenazy and E. Tziperman. Are the 41 ka glacial oscillations a linear response to Milankovitch forcing? *Quat. Sci. Rev*. **23**: 1879-1890, 2004.

[3] I.E. Auger and C.E. Lawrence. Algorithms for the Optimal Identification of Segment Neighborhoods. *Bull. Math. Bio.,* **51:** 39-54, 1989.

[4] J. Bai and P. Perron. Computation and Analysis of Multiple Structural Change Models. *J. Appl. Econ.***18:** 1-22, 2003. doi:10.1002/jae.569

[5] D. Barry and J.A. Hartigan. A Bayesian Analysis for Change Point Problems. *J. Amer. Stat. Assoc*., **88**(421): 309-319, 1993.

[6] G.S. Becker. Crime and Punishment: An Economic Approach. *Journal of Political Economy*, **76:**169-217, 1968.

[7] R. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, New Jersey, 1957.

[8] R.B. Bendel and A.A. Afifi. Comparison of Stopping Rules in Forward 'Stepwise' Regression. *J. Am. Stat. Assoc.* **72**: 46-53, 1977.

[9] A.L. Berger. Long-Term Variations of Daily Insolation and Quaternary Climatic Changes. *J Atoms Sci.* **35:** 2362-2367, 1978.

[10] A.L. Berger and M. F. Loutre. Insolation values for the climate of the last 10 million years. *Quat. Sci. Rev.*, **10**: 297-317, 1991. doi:10.1016/0277-3791(91)90033-Q

[11] W.H. Berger, T. Bickert, H. Schmidt, and G. Wefer. Quaternary oxygen isotope record of pelagic foraminifers: Site 806, Ontong Java Plateau, In *Proc. ODP Sci. Results 130*, Berger, W.H., L.W. Kroenke, and L.A. Mayers (eds.), 381-395, 1993.

[12] R. Bintanja and R.S.W. van de Wal. North American ice-sheet dynamics and the onset of 100,000-year glacial cycles. Nature **454,** 2008. doi:10.1038/nature07158

[13] G.E. Birchfield and M. Ghil. Climate evolution in the Pliocene and Pleistocene from marine-sediment records and simulations: internal variability versus orbital forcing. *J. Geophys. Res.* ***98*****:** 10385-10399, 1993.

[14] J. Bloemendal and P. deMenocal. Evidence for a change in the periodicity of tropical climate cycles at 2.4 Ma from whole-core magnetic susceptibility measurements. *Nature* **342**: 897-900, 1989.

[15] E.W. Bolton, K.A. Maasch, and J.M. Lilly. A wavelet analysis of Plio-Pleistocene climate indicators: A new view of periodicity evolution. *Geophys. Res. Lett*., **22:** 2753-2756, 1995.

[16] L. Bretthorst. *Bayesian Spectrum Analysis and Parameter Estimation*. Springer-Verlag, New York, 1988.

[17] L. Brieman. Better Subset Regression Using the Nonnegative Garrote. *Technometrics*, **37:** 373-384, 1995.

[18] K.R. Briffa, P.D. Jones, F.H. Scheingruber, S.G. Shiyatov, and E.R.Cook. Unusual twentieth-century summer warmth in a 1,000-year temperature record from Siberia. *Nature* **376**: 156-159, 1995.

[19] J.R. Bunch and J.E. Hopcroft. Triangular Factorization and Inversion by Fast Matrix Multiplication. *Mathematics of Computation*, **28**: 231-236, 1974.

[20] B.P. Carlin, A.E. Gelfand, and A.F.M. Smith. Hierarchical Bayesian analysis of changepoint problems. *Appl*. *Statist*. **41**: 389-405, 1992.

[21] H. Caussinus and O. Mestre. Detection and Correction of Artificial Shifts in Climate Series. *J. Roy. Stat. Soc. Series C,* **53**(3): 405-425, 2004.

[22] J. Chappellaz, T. Blunier, D. Raynaud, J.M. Barnola, J. Schwander, and B. Stauffer. Synchronous changes in atmospheric CH4 and Greenland climate between 40 and 8 ka bp. *Nature* **366**: 443-445, 1993.

[23] J. Chen and Z. Chen. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, **95**(3): 759-771, 2008. doi:10.1093/biomet/asn034

[24] S. Chib. Estimation and comparison of multiple change-point models. *J. Econometrics*, **86**(2): 221-241, 1998. doi:10.1016/S0304-4076(97)00115-2

[25] G. Ciuperca, A. Ridolfi, and J Idier. Penalized Maximum Likelihood Estimator for Normal Mixtures. *Scand J. Statist.* **30:** 45-59, 2003.

[26] P.U. Clark and D. Pollard. Origin of the Middle Pleistocene Transition by Ice Sheet Erosion of Regolith. *Paleoceanography,* **13**(1): 1-9, 1998.

[27] S. Clemens. An astronomical tuning strategy for Pliocene sections: implications for global-scale correlation and phase relationships. *Phil. Trans. R. Soc. Lond. A*, **357:** 1949-1973, 1999.

[28] S. Clemens and R. Tiedemann. Eccentricity forcing of Plio-Early Pleistocene climate revealed in a marine oxygen-isotope record. *Nature*, **385:** 801-803, 1997.

[29] K.M. Cobb, C.D. Charles, H. Cheng, and R.L. Edwards. El Nino/Southern Oscillation and tropical Pacific climate during the last millennium. *Nature* **424**: 271-276, 2003.

[30] J.E. Cole, R.G. Fairbanks, and G.T. Shen. The spectrum of recent variability in the Southern Oscillation: results from a Tarawa Atoll coral. *Science,* **260**: 1790-1793, 1993.

[31] D. Coppersmith and S. Winograd. Matrix Multiplication via Arithmetic Progressions. *Journal of Symbolic Computation,* **9:** 251-280, 1990.

[32] T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein. *Introduction to Algorithms (2nd ed)*. MIT Press, New York, 2001.

[33] W. Dansgaard, S.J. Johnsen, H.B. Clausen, D. Dahl-Jensen, N.S. Gundestrup, C.U. Hammer, C.S. Hvidberg, J.P. Steffensen, A.E. Sveinbjornsdottir, J. Jouzel, and G. Bond. Evidence for general instability of past climate from a 250-ka ice-core record. *Nature* **364**: 218-220, 1993.

[34] R.A. Davis, T.C.M. Lee, and G.A. Rodriguez-Yam. Structural Break Estimation for Nonstationary Time Series Models. *J. Am. Stat. Soc.* **101**(473): 223-239, 2006. doi 10.1198/016214505000000745

[35] J. Diebolt and C.P. Robert. Estimation of Finite Mixture Distributiuons through Bayesian Sampling. *J. Roy. Statist. Soc.* **B, 56**: 363-375, 1994.

[36] N.R. Draperand H. Smith. *Applied Regression Analysis* (3rd ed) Wiley, New York, 1998.

[37] N.R. Draper and R.C. van Nostrand. Ridge Regression and James-Stein Estimation: Review and Comments. *Technometrics*, **21:** 451-466, 1979.

[38] R. Durbin, S.R. Eddy, A. Krogh, and G. Mitchinson. *Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids.* Cambridge Univ Press, Cambridge, UK, 1998.

[39] B. Efron. Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis. *J. Am. Stat. Assoc.* **99**(465): 96-104, 2004. doi: 10.1198/016214504000000089

[40] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least Angle Regression. *The Annals of Statistics*, **32:** 407-499, 2004.

[41] I. Ehrlich. Participation in Illegitimate Activities: A Theoretical and Empirical Investigation. *Journal of Political Economy*, **81**(3)**:** 521-565, 1973.

[42] I. Ehrlich. The Deterrent Effect of Capital Punishment: A Question of Life and Death. *AER,* **65**, 395-417, 1975.

[43] T.S. Eicher, C. Papageorgiou, and O. Roehn. Unraveling the fortunes of the fortunate: An Iterative Bayesian Model Averaging (IBMA) approach. *Journal of Macroeconomics*, **29:** 494-514, 2007.

[44] J. Esper, E.R.Cook, and F.H. Schweingruber. Low-frequency signals in long tree-ring chronologies for reconstructing past temperature variability. *Science* **295**: 2250-2253, 2002.

[45] X. Estivill and L. Armengol. Copy Number Variants and Common Disorders: Filling the Gaps and Exploring Complexity in Genome-Wide Association Studies. *PLoS Genet* **3**(10): e190, 2007.. doi:10.1371/journal.pgen.0030190

[46] J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *J.R. Statist. Soc. B*, **70:** 849-911, 2008.

[47] P. Fearnhead. Exact and Efficient Bayesian Inference for Multiple Changepoint problems. *Stat. Comput.* **16:** 203-213, 2006. doi 10.1007/s11222-006-8450-8.

[48] C. Fernandez, E. Ley, and M.F.J. Steel. Benchmark Priors for Bayesian Model Averaging. *J. Econometrics*, **100**: 381-427, 2001a.

[49] C. Fernandez, E. Ley, and M.F.J. Steel. Model Uncertainty in Cross-Country Growth Regression. *J. Appl. Econometrics,* **16**: 563-576, 2001b.

[50] B.M. Fleisher. *The Economics of Delinquincy.* Quadrangle, Chicago, 1966.

[51] G.M. Furnival and R.W. Wilson. Regression by leaps and bounds. *Technometrics*, **16**:499-511, 1974.

[52] E.I. George and R.E. McCulloch. Variable Selection via Gibbs Sampling. *J. Amer. Stat. Assoc.*, **88:** 881-890, 1993.

[53] M. Ghil, M.R. Allen, M.D. Dettinger, K. Ide, D. Kondrashov, M.E. Mann, A.W. Robertson, A. Saunders, Y. Tian, F. Varadi, P. Yiou. Advanced spectral methods for climatic time series. *Rev. Geophys.* **40** (1): Art. No. 1003 AUG-SEP 2002. doi 10.1029/2000RG000092

[54] H. Gildor and E. Tzipernam. Sea ice as the glacial cycles' climate switch: Role of seasonal and orbital forcing. *Paleoceanography*, **15:** 605-615, 2000.

[55] P.J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**(4):711-732, 1995; doi:10.1093/biomet/82.4.711

[56] P.R. Halmos. *Finite Dimensional Vector Spaces.* VanNostrand, Princeton, 1958.

[57] D.M. Hawkins. Point Estimation of the Parameters of Piecewise Regression Models. *Appl. Statist.* **25** (1): 51-57, 1976.

[58] D.M. Hawkins. Fitting multiple change-point models to data. *Comp. Stat. Data Analysis*, **37**(3): 323-341, 2001. doi: 10.1016/S0167-9473(00)00068-2.

[59] J.D. Hays, J. Imbrie, and N. J. Shackleton. Variations in the Earth's orbit: Pacemakers of the ice ages. *Science*, **194:**1121– 1132, 1976.

[60] T.D. Herbert and L. Mayer. Long climatic time series from DSDP/ODP physical property measurements. *Journal of Sedimentary Petrology*, **61**: 1089-1108, 1991.

[61] F.S. Hillier and G.J. Lieberman. *Introduction to Operations Research*, 8th Edition. McGraw-Hill Professional, Boston, 2005.

[62] A.E. Hoerl and R.W. Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics,* **12:** 55-67, 1970a.

[63] A.E. Hoerl and R.W. Kennard. Ridge Regression: Applications to Nonorthogonal Problems. *Technometrics*, **12:** 69-82, 1970b.

[64] J.A. Hoeting, D. Madigan, A.E. Raftery, and C.T. Volinsky. Bayesian Model Averaging: A Tutorial. *Statist. Sci.* **14:** 382-417, 1997. (Corrected version from http://www.stat.washington.edu/www/research/online/hoeting1999.pdf).

[65] P. Hubert. The segmentation procedure as a tool for discrete modeling of hydrometeorological regimes. *Stochastic Environmental Research and Risk Assessment,* **14**: 297-304, 2000. doi:10.1007/PL00013450

[66] P. Huybers. Glacial variability over the last two million years: an extended depth-derived agemodel, continuous obliquity pacing, and the Pleistocene progression. *Quat. Sci. Rev.* **26:** 37-55, 2007.

[67] P. Huybers and C. Wunsch. A depth derived Pleistocene age model: Uncertainty estimates, sedimentation variability, and nonlinear climate change. *Paleoceanography*, **19:** PA1028, 2004. doi:10.1029/2002PA000857.

[68] P. Huybers and C. Wunsch. Obliquity pacing of the late Pleistocene glacial terminations. *Nature*, **434:** 491-494, 2005.

[69] J. Imbrie, E.A. Boyle, S.C. Clemens, A. Duffy, W.R. Howard, G. Kukla, J. Kutzbach, D.G. Martinson, A. McIntyre, A.C. Mix, B. Molfino, J.J. Morley, L.C. Peterson, N.G. Pisias, W.L. Prell, M.E. Raymo, N.J. Shackleton, and J.R. Toggweiler. On the structure of major glaciation cycles 1Linear responses to Milankovitch forcing, *Paleoceanogr*. **7**: 701-738, 1992.

[70] J. Imbrie, J.D. Hays, D.G. Martinson, A. McIntyre, A.C. Mix, J.J. Morley, N.G. Pisias, and N.J. Shackleton. The orbital theory of Pleistocene climate: support from a revised chronology of the marine ^{18}O record, In: A. Berger, J. Imbrie, J. Hays, G. Kukla, and B. Saltzman (eds.) Milankovitch and Climate, pp. 269-305. Hingham, MA: Reidel, 1984.

[71] J. Imbrie, et al. On the Structure and Origin of Major Glaction Cycles 2, The 100,000-Year Cycle. *Paleoceanography,* **8**(6): 699-735, 1993.

[72] J. Imbrie and J. Z. Imbrie. Modeling the climatic response to orbital variations. *Science*, **207:**943–953, 1980.

[73] J. Imbrie, A. McIntyre, and A. Mix. Oceanic response to orbital forcing in the late Quaternary: Observational and experimental strategies, in A. Berger et al. (eds.) Climate and Geo-Sciences, Kluwer, 121-164, 1989.

[74] F. Jin, J. D. Neelin, and M. Ghil. El Nino on the devil's staircase: annual subharmonic steps to chaos. *Science* **264**: 70-72, 1994.

[75] N. Jones and P. Pevzner. *An Introduction to Bioinformatics Algorithms*. MIT Press, Cambridge, MA, 2004.

[76] P.D. Jones and M.E. Mann. Climate over past millennia. *Rev. Geophys.* **42** Art. No. RG2002, 2004.

[77] J.E. Joyce, L.R.C. Tjalsma, and J.M. Prutzman. High resolution planktonic stable isotopic record and spectral analysis for the past 5.35 My: Ocean Drilling Program Site 625 Northeast Gulf of Mexico. *Paleoceanogr*. ***5***: 507-529, 1990.

[78] A. Kehagias. A hidden Markov model segmentation procedure for hydrological and environmental time series. *Stochastic Environmental Research and Risk Assessment,* **18**(2): 117-130, 2004. doi:10.1007/s00477-003-0145-5.

[79] T. King. Quantifying nonlinearity and geometry in time series of climate. *Quat. Sci. Rev*., **15**: 247-266, 1996.

[80] A.K. Konopka and M.J.C. Crabbe. *Compact Handbook of Computational Biology*. Marcel Dekker, New York, 2004.

[81] A.H. Land and A.G. Doig. An Automatic Method of Solving Discrete Programming Problems. *Econometrica* **28**(3): 497-520, 1960.

[82] K.M. Lau and H. Weng. Climate Signal Detection using Wavelet Transform: How to make a Time Series Sing. *Bull. Amer. Meteor. Soc.*, **76**: 2391-2402, 1995.

[83] M. Lavielle and E. Lebarbier. An application of MCMC methods for the multiple change-points problem. *Signal Processing*, **81**(1): 39-53, 2001. doi:10.1016/S0165-1684(00)00189-4

[84] T.C.M. Lee. On Algorithms for Ordinary Least Squares Regression Spline Fitting: A Comparative Study. *J. Statist. Comput. Simul.*, **72**, 647-663, 2002.

[85] A. Lew and H. Mauch. *Dynamic Programming: A Computational Tool*. Springer, New York, 2007.

[86] L.E. Lisiecki and M. E. Raymo. A Pliocene-Pleistocene stack of 57 globally distributed benthic $\delta^{18}O$ records. *Paleoceanography*, **20**, PA1003, 2005. doi:10.1029/2004PA001071.

[87] H.S. Liu, and B.F. Chau. Wavelet Spectral Analysis of the Earth's Orbital Variations and Paleoclimatic Cycles. *J. Atmos. Sci*., **55:** 227-236, 1998.

[88] J.S. Liu and C.E. Lawrence. Bayesian inference on Biopolymer models. *Bioinformatics,* **15**:38-52, 1999.

[89] Z. Liu, L. C. Cleaveland, and T.D. Herbert. Early onset and origin of 100-kyr cycles in Pleistocene tropical SST records. *Earth and Planetary Science Letters **265***: 703–715, 2008.

[90] M.F. Loutre, D. Paillard, F. Vimeux, and E. Cortijo. Does mean annual insolation have the potential to change the climate? *Earth and Planetary Science Letters,* **221:** 1-14, 2004.

[91] Q. Lu, R. Lund, and T.C.M. Lee. An MDL Approach to the Climate Segmentation Problem, *Ann. Appl. Stat.*, 2010.

[92] K.A. Maasch. Statistical detection of the mid-Pleistocene transition. *Climate Dynamics* **2:** 133-143, 1998.

[93] D. Madigan and J. York. Bayesian Graphical Models for Discrete Data. *International Statistical Review*, **63:** 215-232, 1995.

[94] M.E. Mann. On smoothing potentially non-stationary climate time series. *Geophys. Res. Lett.* **31** (7): Art. No. L07214 APR 15 2004.

[95] M.E. Mann, R.S. Badley, and M.K. Hughes. Global-scale temperature patterns and climate forcing over the past six centuries. *Nature* **392**: 779-783, 1998.

[96] L.C. Marsh and D.R. Cormier. *Spline Regression Models*. Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-137. Thousand Oaks, CA: Sage, 2001.

[97] M.A. Maslin and A. Ridgwell. Mid Pleistocene Revolution and the 'eccentricity myth'. Special Publications of the *Geological Society of London* **247**: 19-34, 2005.

[98] P.A. Mayewski, L. D. Meeker, M. C. Morrison, M. S. Twickler, S. Whitlow, K. K. Ferland, D. A. Meese, M. R. Legrand, and J. P. Steffensen. Greenland ice core "signal" characteristics: An expanded view of climate change. *J. Geophys. Res.* **98**: 12839-12847, 1993.

[99] M. Milankovitch. Canon of Insolation and the Ice-Age Problem. *Israel Program for Scientific Translations*. Jerusalem (1969), 1941.

[100] A.J. Miller. *Subset Selection in Regression* (2nd ed), Chapman and Hall, New York, 2002.

[101] K.S. Miller. On the Inverse of the Sum of Matrices. *Mathematics Magazine,* **54**(2): 67-72, 1981.

[102] T.J. Mitchell and J.J. Beauchamp. Algorithms for Bayesian Variable Selection in Regression. In: T.J. Boardman (Ed.), *Computer Science and Statistics: Proceedings of the 18th Symposium on the Interface*, American Statistical Association, Washington, D.C., 181-182, 1986.

[103] T.J. Mitchell and J.J. Beauchamp. Bayesian Variable Selection in Linear Regression. (with discussion) *Journal of the American Statistical Association*, **83**: 1023-1036, 1988.

[104] M. Mudelsee and M. Schulz. The Mid-Pleistocene climate transition: onset of 100 ka cycle lags ice volume build-up by 280 ka. *Earth and Planetary Science Letters* **151:** 117-123, 1997.

[105] R. Muller and G. MacDonald. Glacial Cycles and Orbital Inclination. *Nature* **377:** 107-108, 1995. doi:10.1038/377107b0

[106] R. Muller and G. MacDonald. Glacial Cycles and Astronomical Forcing. *Science* **277:** 215-218, 1997a. doi: 10.1126/science.277.5323.215

[107] R. Muller and G. MacDonald. Simultaneous presence of orbital inclination and eccentricity in proxy climate records from Ocean Drilling Program Site 806. *Geology* **25:** 3-6, 1997b.

[108] D. Paillard. The timing of Pleistocene glaciations from a simple multiple-state climate model. *Nature* **391**: 378 – 381, 1998.

[109] J. Park and T.D. Herbert. Hunting for paleoclimatic periodicities in a sedimentary series with uncertain time scale. *Journal of Geophysical Research, **92B***: 14,027-14,040, 1987.

[110] J. Park and K. Maasch. Plio-Pleistocene time evolution of the 100- ka cycle in marine paleoclimate records. *J. Geophys. Res*., **98:** 447- 461, 1993.

[111] T. Park and G. Casella. The Bayesian Lasso. *J. Amer. Stat. Assoc*., **103**(482): 681-686, 2008. doi:10.1198/016214508000000337.

[112] L. Perreault, et al. Retrospective multivariate Bayesian change-point analysis: A simultaneous single change in the mean of several hydrological sequences. *Stochastic Environmental Research and Risk Assessment* **14:** 243-261, 2000. doi: 10.1007/s004770000051

[113] J.R. Petit, J.R., J. Jouzel , D. Raynaud, N.I. Barkov, J.-M. Barnola, I. Basile, M. Bender, J. Chappellaz, M. Davis, G. Delaygue, M. Delmotte, V.M. Kotlyakov, M. Legrand, V.Y. Lipenkov, C. Lorius, L. Pepin, C. Ritz, E. Saltzman, and M. Stievenard. Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature* **399**: 429-436, 1999.

[114] D.B. Phillips and A.F.M. Smith. Bayesian model comparison via jump diffusions. *Markov Chain Monte Carlo in Practice*, Ed. W. R. Gilks, S. T. Richardson and D. J. Spiegelhalter, Ch. 13. London: Chapman and Hall, 1995.

[115] N.G. Pisias, A.C. Mix, and R. Zahn. Nonlinear response in the global climate system: evidence from benthic oxygen isotope record in core RC13-110. *Paleoceanography*, **5**(2): 147-159, 1990.

[116] T.M. Quinn, T.M. Crowley, F.W. Taylor, C. Henin, P. Joannot, and Y. Join. A multicentury stable isotope record from a New Caledonia coral: interannual and decadal SST variability in the southwest Pacific since 1657. *Paleoceanography,* **13**: 412-426, 1998.

[117] A.E. Raftery. Bayesian Model Selection in Social Research. *Sociological Methodology*, **25**:111-163, 1995.

[118] A.E. Raftery, D. Madigan, and J.A. Hoeting. Bayesian Model Averaging for Linear Regression Models. *J. Amer. Stat. Assoc.*, **92:** 179-191, 1997.

[119] A.C. Ravelo, D.H. Andreasen, M. Lyle, A. Olivarez, and M.W. Wara. Regional climate shifts caused by gradual global cooling in the Pliocene epoch. *Nature* **429** (6989): 263-267, 2004.

[120] M.E. Raymo. The initiation of Northern Hemisphere glaciations. *Ann. Rev. Earth Planet. Sc*i., **22**: 353-383, 1994.

[121] M.E. Raymo. The timing of major terminations. *Paleoceanography*, **12:** 577– 585, 1997.

[122] M.E. Raymo, L.E. Lisiecki, and K.H. Nisancioglu. Plio-Pleistocene Ice Volume, Antarctic Climate, and the Global $\delta^{18}O$ Record. *Science*, **313**: 492-495, 2006.

[123] M.E. Raymo and K. Nisancioglu. The 41k World: Milankovitch's Other Unsolved Mystery. *Paleoceanography,* **18**(1): 11, 2003. doi:10.1029/2002PA000791

[124] M.E. Raymo, D. W. Oppo, and W. Curry. The mid-Pleistocene climate transition: a deep-sea carbon perspective. *Paleoceanography*, **12:** 546**–**559, 1997.

[125] R. Redon, et al. Global variation in copy number in the human genome. *Nature,* **444**(7118): 444-454, 2006. doi:10.1038/nature05329

[126] A.J. Ridgwell, A. J. Watson, and M. E. Raymo. Is the spectral signature of the 100 kyr glacial cycle consistent with a Milankovitch origin? *Paleoceanography*, **14:** 437–440, 1999.

[127] W.F. Ruddiman. Orbital insolation, ice volume, and greenhouse gases. *Quaternary Science Reviews*, **22**: 1597-1629, 2003.

[128] W.F. Ruddiman, M. Raymo, and A. McIntyre. Matuyama 41000 year cycles: North Atlantic ocean and northern hemisphere ice sheets. *Earth & Planet. Sci. Letters,* **80***:* 117-129, 1986.

[129] E. Ruggieri, T. Herbert, K.T. Lawrence, and C.E. Lawrence. Change point method for detecting regime shifts in paleoclimatic time series: Application to d18O time series of the Plio-Pleistocene. *Paleoceanography*, **24:** PA1204, 2009. doi:10.1029/2007PA001568.

[130] S. Rutherford and S. D'Hondt. Early onset and tropical forcing of 100,000-year Pleistocene glacial cycles. *Nature* **408**: 72 – 75, 2000.

[131] S. Rutherford, M.E. Mann, T.L. Delworth, and R.J. Stouffer. Climate field reconstruction under stationary and nonstationary forcing. *J. Climate* **16** (3): 462-479, 2003.

[132] G.A.F. Seber and A.J. Lee. *Linear Regression Analysis.* John Wiley & Sons, Hoboken, NJ, 2003.

[133] D.J. Seidel and J.R. Lanzante. An assessment of three alternatives to linear trends for characterizing global atmospheric temperature changes. *J. Geophys. Res.* **109:** D14108, 2004. doi:10.1029/2003JD004414

[134] N.J. Shackleton, J. Backman, H. Zimmerman, D.V. Kent, M.A. Hall, D.G. Roberts, D. Schnitker, J.G. Baldauf, A. Desprairies, R. Homrighausen, P. Huddlestun, J.B. Keene, A.J. Kaltenback, K.A.O. Krumsiek, A.C. Morton, J.W. Murray, and J. Westberg-Smith. Oxygen isotope calibration of the onset of ice-rafting and history of glaciation in the North Atlantic region. *Nature*, **307**: 620-623, 1984.

[135] B.W. Silverman. Penalized Maximum Likelihood Estimation. *Encyclopedia of Statistical Sciences*, **6**: 664-667, 1985.

[136] G. Smith and F. Campbell. A Critique of some Ridge Regression Methods. *J. Amer. Stat. Assoc.*, **75:** 74-103, 1980 (incl. discussion by Thisted, Marquardt, van Nostrand, Lindley, Overchain, Peele and Ryan, Vinod, and Gunst, and authors' rejoinder).

[137] D.A. Stephens. Bayesian Retrospective Multiple-changepoint Identification. *Appl. Statist*, **43**(1): 159-178, 1994.

[138] G.J. Stigler. The Optimum Enforcement of Laws. *J.P.E.* **78**: 526-536, 1970.

[139] D.R. Taft and R.W. England Jr. *Criminology (4th ed)* MacMillan, New York, 1964.

[140] R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *J. Roy. Statist. Soc,* **B, 58:** 267-288, 1996.

[141] A.R. Tome and P. M. A. Miranda. Piecewise linear fitting and trend changing points of climate parameters. *Geophys. Res. Lett.,* **31:** L02207, 2004. doi:10.1029/2003GL019100.

[142] C. Torrence and G.C. Compo. A Practical Guide to Wavelet Analysis. *Bull. Am. Meteor. Soc.* **79**: 61-78, 1998.

[143] E. Tziperman and H. Gildor. On the mid-Pleistocene transition to 100-kyr glacial cycles and the asymmetry between glaciation and deglaciation times. *Paleoceanography*, **18**(1): 1001, 2003. doi:10.1029/2001PA000627.

[144] E. Tziperman, M. E. Raymo, P. Huybers, and C. Wunsch. Consequences of pacing the Pleistocene 100 kyr ice ages by nonlinear phase locking to Milankovitch forcing. *Paleoceanography*, **21:** PA4206, 2006. doi:10.1029/2005PA001241.

[145] W. Vandaele. Participation in Illegitimate Activities: Ehrlich Revisited. In: A Blumstein, J. Cohen, and D. Nagin *(Eds.), Deterrence and Incapacitation,* National Academy of Sciences Press, Washington, D.C., 270-335, 1978.

[146] R. Vautard and M. Ghil. Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series. *Phys. D,* **35:** 395–424, 1989.

[147] R. Vautard, P. Yiou, and M. Ghil. Singular spectrum analysis: A toolkit for short noisy chaotic signals. *Phys. D*, **58**:95–126, 1992.

[148] G. Villard. Computation of the Inverse and Determinant of a Matrix. In: F. Chyzak (Ed.) *Algorithms Seminar,* INRIA Rocquencourt, France, 29-32, 2003.

[149] H.M. Wagner. *Principles of Operations Research*. Prentice Hall, Inc., Englewood Cliffs, NJ, 1975.

[150] H. Wang. Forward Regression for Ultra-High Dimensional Variable Screening. *J. Amer. Stat. Assoc.,* **104**(488): 1512-1524, 2009. doi 10.1198/jasa.2008.tm08516

[151] B. Western and M. Kleykamp. A Bayesian Change Point Model for Historical Time Series Analysis. *Political Analysis* **12**(4):354-374, 2004; doi:10.1093/pan/mph023

[152] W. L. Winston. Operations *Research: Applications and Algorithms*, 4th Edition. Duxbury Press, 2003.

[153] C. Wunsch. The interpretation of short climate records with comments on the North Atlantic and Southern Oscillations. *Bull. Am. Met. Soc.,* **80**: 245-255, 1999.

[154] P. Yiou, J. Genthon, J. Jouzel, M. Ghil, H. Le Treut, J.M. Barnola, C. Lorius, and Y.N. Korotkevich.

High-frequency paleovariability in climate and in CO2 levels from Vostok ice-core records. *J. Geophys. Res.* ***96****(B12)***:** 20365-20738, 1991.

[155] K.A. Yueng, R.E. Bumgarner, and A.E. Raftery. Bayesian Model Averaging: Development of an improved multi-class, gene selection and classification tool for microarray data. *Bioinformatics,* **21**:2394-2402, 2005.

[156] H. Zou and T. Hastie. Regularization and Variable Selection via the Elastic Net. *J.Roy. Statist. Soc.,* **B, 67:** 301-320, 2005.