

## Statement of Research Interests – Eric Ruggieri

Climate change remains one of today's hot-button issues. Over the last several decades, geologists have analyzed oxygen isotope ratios ( $\delta^{18}\text{O}$ ) from benthic foraminifera that have been shown to accurately record the amount of glacial cover on the Earth<sup>1</sup>. Given that the quantity of ice on the Earth is believed to be controlled by the amount of solar insolation received at high latitudes in the summer, three important questions emerge:

- 1) Which insolation curve, in terms of latitude and season, best matches the  $\delta^{18}\text{O}$  record?
- 2) Is the climate system being paced or forced by solar insolation?
- 3) The Mid-Pleistocene Transition, a major change in the record, occurs around 1 Ma (Million years ago). Is this change gradual or abrupt? Additionally, given a set of hypotheses, can the cause of this change be determined?

Because of the large number of data points in the time series, together with the multitude of insolation curves and alternative hypotheses for ice sheet dynamics, an answer to these questions can only be found through inference in high dimensional space.

By combining elements of variable selection with change point regression, my work has begun to provide an answer to these questions. In the least squares setting, optimization over all possible locations of change points and regression coefficients is done in a computationally efficient way to find the globally optimal solution; from a probabilistic perspective, after marginalizing over the locations of the change points and the regression coefficients, Bayes Rule can be used to obtain estimates of the uncertainty surrounding the solution by drawing samples from the posterior distribution. Analyzing the data by these methods, I have found that the insolation curves that best match the structure of the  $\delta^{18}\text{O}$  record is that of high latitude winters rather than the widely perceived high latitude summer insolation curves. Additionally, the Mid-Pleistocene Transition was determined to be a gradual change from a forced to a paced dynamic system, whose exact mechanism remains uncertain.

This  $\delta^{18}\text{O}$  record<sup>2</sup> indicates that during the last 5 million years, the Earth's glaciers have melted and reformed at regular intervals with the last glacial minima occurring 120ka (thousand years ago), a time at which the ocean was 6m higher than today's levels. Perhaps the most interesting part of the 5 Myr glacial record is the two distinct changes that have occurred:

- 1) 2.7 Ma: Prior to 3.3Ma, the absence of ice sheets in the Northern Hemisphere caused glacial influence on climate to be dominated by Southern Hemisphere phenomena. Between 3.3 Ma and 2.7Ma, this balance began to shift as small glaciers began to form in the Northern Hemisphere. By 2.7Ma, permanent Northern Hemisphere ice sheets had formed and became the dominant glacial influence on climate, an effect which remains true today. The  $\delta^{18}\text{O}$  glacial record gives evidence of this transition by a large amplitude change - glacial formation and destruction retained its 41 kyr (thousand years) periodic structure, but the magnitude of glaciations increased.
- 2) 0.8-1.2Ma: Different researchers have concluded that the change around 1 Ma, known as the Mid-Pleistocene Transition, was both gradual and abrupt, and occurred at 0.6 Ma, 1.5 Ma, or somewhere in between (for a list of references, see [3]). The  $\delta^{18}\text{O}$  record again showed an increase in amplitude at this time, but more significantly, the frequency of glaciations changed from 41 kyr to 100 kyr.

The dominant theory of ice volume, Milankovitch Theory<sup>4</sup>, states that the quantity of ice on the Earth is controlled by the amount of solar insolation received at the top of the Earth's atmosphere at high latitudes during the summer. 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). Eccentricity

controls the absolute amount of insolation, whereas obliquity and precession control the distribution of insolation on the Earth's surface. As early as 1980, Imbrie and Imbrie<sup>5</sup> noted that "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".

My background in Statistics and Computational Biology gave me a unique perspective from which to view this problem. A change point algorithm<sup>6</sup> initially developed for DNA sequence problems seemed ideally suited to study the  $\delta^{18}\text{O}$  record. Given an unknown dependent variable,  $y$ , and  $m$  known predictor variables  $x_1, \dots, x_m$ , linear regression methods are based upon the statistical model

$$y = \sum_{i=1}^m \beta_i x_i + \varepsilon$$

Where  $\beta_i$  is the  $i^{\text{th}}$  regression coefficient and  $\varepsilon$  is a random error term. The change point problem is concerned with the identification of a point in a time series at which some change occurs, here the parameters of the model. The time series as a whole is heterogeneous, but between any two change points the time series is assumed to be homogeneous. If we are interested in the optimal placement of  $k$  change points among the  $N$  observations of a time series, then the change point algorithm uses dynamic programming to reduce the  $O(N^k)$  calculation of all possible solutions – a combinatorial nightmare – to a more manageable  $O(kN^2)$ . I modified this recursive algorithm to perform least squares linear regression with periodic functions. Specifically, if  $f_k(y_{1:j})$  represents the minimum squared error obtained by fitting data points  $y_1 \dots y_j$  with  $k$  change points, then the forward step of the dynamic programming algorithm calculates

$$f_k(y_{1:j}) = \min_{1 \leq v \leq j} \{ f_{k-1}(y_{1:v}) + f_0(y_{v+1:j}) \}$$

where  $f_0(y_{v+1:j})$  is the least squares solution for regression on data points  $y_{v+1} \dots y_j$ . On the backward step, the optimal solution can be obtained by finding the argument which minimizes this function.

Working together with collaborators from the Department of Geological Sciences at Brown University, the analysis of the  $\delta^{18}\text{O}$  record by the change point algorithm was published in *Paleoceanography*<sup>3</sup>. Both of the transitions noted above, along with several other novel changes were identified by the least squares change point procedure. More importantly, we found that the Mid-Pleistocene Transition was not a time when 100 kyr glaciations replaced 41 kyr glaciations, but a time when the dominant 100 kyr glaciations arose in addition to the existing 41 kyr glaciations. The algorithm has also been published as open source software together with a Graphical User Interface (GUI) after receiving inquiries from other geologists about its implementation and use in their own research.

Several unanswered questions remained. In particular, a characterization of the uncertainty surrounding the globally optimal least squares change point solution was needed. The Bayesian version of the change point algorithm<sup>7</sup> uses a probabilistic setup together with dynamic programming to navigate through each possible solution. Assuming that the error terms are IID (Independent and Identically Distributed) Normal and by using a Normal-Inverse $\chi^2$  conjugate prior, one can integrate over the parameter space to find the marginal probability of any segment of the time series  $y_i \dots y_j$ ,  $1 \leq i < j \leq N$  given the regression model,

$$f(y_{i:j}) = \iint f(y_{i:j} | \beta, \sigma^2) f(\beta | \sigma^2) f(\sigma^2) d\beta d\sigma^2$$

Let  $f_k(y_{1:j})$  be the probability of the data points  $y_1 \dots y_j$  with  $k$  change points. Instead of optimizing as in the least squares algorithm, the forward step for the Bayesian version of the change point algorithm will marginalize over the placement of the change points:

$$f_k(y_{1:j}) = \sum_{1 \leq v \leq j} f_{k-1}(y_{1:v}) * f_0(y_{v+1:j})$$

where  $f_0(y_{v+1:j}) = f(y_{v+1:j})$  given above. The benefit of the Bayesian algorithm is that one can now answer questions about the uncertainty in the number of change points, their locations, and the regression coefficients for the proposed regression model by sampling, on the backward step, from the posterior distribution on these parameters.

However, many competing hypotheses for ice sheet dynamics exist, from the forced orbital system implied by Milankovitch Theory, to a paced system, where the timing, rather than the structure of glacial events are set by the orbital parameters. Hence, my current research focuses on Bayesian variable selection in the context of the change point algorithm with an application to the  $\delta^{18}\text{O}$  record. The time required to calculate all possible sub-models for each possible time interval would be prohibitive by brute force methods. Quick, greedy algorithms for variable selection exist<sup>8</sup>, but they fail to give a probabilistic perspective to the problem. Approximation techniques, such as MCMC algorithms<sup>9</sup>, can sample their way through the space, but the number of samples needed to achieve convergence can be large. Because the time complexity of brute force probabilistic computations is limited by the calculation of a matrix determinant and a matrix inverse, I developed a recursive procedure to reduce the complexity of these calculations which is related to finding the inverse of a sum of matrices<sup>10</sup>.

The variable selection algorithm can be visualized as a complete binary tree. Each level of the tree corresponds to one variable with internal nodes dividing the set of sub-models in half; one branch signifies exclusion, while the other branch signifies inclusion of the variable at that level with calculations performed only when the inclusion branch is traversed. Each leaf on the regression tree holds the probability of one possible sub-model. The algorithm quickly and efficiently calculates the probability of one sub-model from another, moving left to right (or right to left) through the leaves of the regression tree. Instead of calculating a matrix inverse and a matrix determinant for each possible sub-model, one need only perform vector multiplication and matrix addition to calculate the matrix inverse and multiplication by a constant to calculate the matrix determinant. The results of an analysis on a well-studied 'Crime and Punishment' data set<sup>11</sup> show the algorithm's exact representation of the posterior space matches the results of approximating MCMC based approaches, but in a fraction of the time.

Further reductions in the complexity can be made if, for example, there is a restriction on the number of variables that can be included in a given model or if variables are orthogonal to each other. For the former, entire branches are pruned from the regression tree, whereas for the latter, a single deep regression tree can be replaced by several more shallow regression trees. In both cases, a reduction in the number of internal nodes on the binary tree reduces the number of calculations that need to be performed. In what could be viewed as the Bayesian equivalent to the Leaps and Bounds algorithm<sup>12</sup>, an additional way to reduce the number of internal nodes on the tree is to combine this technique with a branch and bound algorithm. This variable selection technique can be used by itself, but its true utility for answering questions on ice volume is shown when it is used in conjunction with the change point algorithm.

Combining variable selection with the change point algorithm is a novel contribution to the field of statistics which enables highly flexible modeling. Now, not only can the parameters of a model change through time, but the model components themselves can change. After marginalizing out the choice of a sub-model,  $A_m$ , the marginal probability of any segment of the time series  $y_i \dots y_j$   $1 \leq i < j \leq N$  becomes:

$$f_0(y_{i:j}) = \sum_{\text{all } A_m} \iint f(y_{i:j} | \beta, \sigma^2, A_m) f(\beta | \sigma^2, A_m) f(\sigma^2) d\beta d\sigma^2 f(A_m)$$

This result can now be used for the forward step of the Bayesian change point described previously.

Analysis of the  $\delta^{18}\text{O}$  glacial record using the Bayesian change point and variable selection algorithm has identified, on average, nine changes within the  $\delta^{18}\text{O}$  record, including the two major changes noted by geologists, along with uncertainty estimates as to their timing. While this analysis has not been able to declare one hypothesis for ice sheets dynamics superior to the rest, it can categorically rule out some of the proposed mechanisms based upon their relative inability to fit the  $\delta^{18}\text{O}$  record. More importantly, the Mid-Pleistocene Transition can be viewed as a gradual change that was not the result of one mechanism replacing another, but the emergence of a new mechanism (such as feedbacks due to the ever increasing magnitude of ice sheets or the emergence of marine-based ice sheets) in addition to the existing forcing and response dynamics (terrestrial-based ice sheets), confirming the conclusions drawn from the least squares analysis.

Future research projects and opportunities for undergraduate research are abundant. This work has already spawned three summer projects for undergraduates and first year graduate students. Briefly, these projects are related to: 1) Clustering of the sampled solutions from the variable selection algorithm; 2) Exploring the correlation through time between a set of three proxy records related to the El Nino climate system using the change point algorithm and a Hidden Markov Model (HMM); and 3) Analyzing the differences in enrichment around transcription start sites in both young and old flies using change point together with a clustering algorithm. Another inviting project would be to use the change point algorithm to study sea surface temperature records in order to determine whether temperature changes through time are coherent across the globe or whether they differ based on location.

My research interests in Computational Biology lie in statistical modeling in genomics, specifically Genome Wide Association Studies (GWAS) in the context of phasing algorithms and variable selection. Phasing algorithms and measures of Linkage Disequilibrium (LD) are often used without discrimination in GWAS, even though no two will generate the same output<sup>13</sup>. A Computer Science group that I have worked with, led by Sorin Istrail, is looking to create a set of axiomatic principles that all phasing algorithms should abide by with the goal of trying to develop an algorithm that satisfies these axioms. In the context of this project, I have written a pair HMM phasing algorithm with similarities to the fastPhase algorithm<sup>14</sup>. This graphical model can be shown to be more accurate and efficient than Parsimony, Maximum Likelihood, or Expectation Maximization (EM) based algorithms.

Variable selection in the context of GWAS appears to be a promising, but computationally challenging issue to explore. Successful GWAS studies<sup>15</sup> have been able to find the one genomic marker that correlates in a statistically significant way with the disease in question. However, while simple Mendelian diseases may have a single genetic cause, the vast majority of diseases are complex and likely have multiple genes that are at least partially responsible. Efficient variable selection may help to lessen the combinatorial burden of searching for correlations between multiple markers and the disease.

Finally, a long-term project still in its beginning stages is to use an Information Theoretic approach to explore the parallels between the cell and a computer. Specifically, a computer built from unreliable (able to fail) parts can use error detection and correction techniques to store information and execute code without error. In a similar manner, a cell has built-in error detection and correction schemes. While parts of a cell may be unreliable, a cell's DNA must store the cell's vital information and then transcribe and translate that information without error. An error in either system can cause a computer program to crash or the cell to die. The ability to transmit information

either without error or with a small amount of correctible error is intimately tied to the Information Theoretic notions of Channel Capacity and Rate Distortion.

The use of dynamic programming and recursions in the context of graphical models can lessen the combinatorial burden in many problems. The change point algorithm answers questions regarding the optimal or probabilistic placement of regime boundaries; variable selection allows the model to change with time. Together, they provide an increased flexibility in statistical modeling. Moving forward, change point and variable selection can be applied to genome studies, temperature records, and can allow me to dive further into ice sheet dynamics. The true nature of ice sheet variability may remain a mystery, but a better understanding about the glacial changes that have occurred on the Earth over the last 5 Myr now exists.

## References and Notes:

- [1] For additional background information on ice sheet dynamics, refer to: Ruddiman, W.F. 2008. *Earth's Climate: Past and Future* (2<sup>nd</sup> Ed.), W.H. Freeman and Co., New York.
- [2] Lisiecki, L. E., and M. E. Raymo (2005), A Pliocene-Pleistocene stack of 57 globally distributed benthic  $\delta^{18}O$  records, *Paleoceanography*, 20, PA1003, doi:10.1029/2004PA001071.
- [3] Ruggieri, E., Herbert, T., Lawrence, K.T., and Lawrence, C.E., (2009). 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, doi:10.1029/2007PA001568.
- [4] Milankovitch, M. (1941). Canon of Insolation and the Ice-Age Problem. Israel Program for Scientific Translations. Jerusalem (1969).
- [5] Imbrie, J., and J. Z. Imbrie (1980), Modeling the climatic response to orbital variations, *Science*, **207**, 943– 953.
- [6] Auger, I.E., and Lawrence, C.E., (1989). Algorithms for the Optimal Identification of Segment Neighborhoods. *Bull. Math. Bio.*, **51**, 39-54.
- [7] Lui, J.S. and Lawrence, C.E., (1999). Bayesian Inference on Biopolymer Models. *Bioinformatics*, **15**(1): 38-52.
- [8] For an overview, see: Miller, A.J., 2002. *Subset Selection in Regression* (2<sup>nd</sup> ed), Chapman and Hall, New York.
- [9] For example: Fernandez, C., Ley, E., and Steel, M.F.J., (2001). Benchmark Priors for Bayesian Model Averaging. *J. Econometrics*, **100**, 381-427.
- [10] For a derivation, see: Miller, K.S., (1981). On the Inverse of the Sum of Matrices. *Mathematics Magazine*, **54**(2) 67-72.
- [11] Ehrlich, I., (1973). Participation in Illegitimate Activities: A Theoretical and Empirical Investigation. *Journal of Political Economy*, **81** 521-565.
- [12] Furnival, G.M. and Wilson, R.W. (1974). Regression by Leaps and Bounds, *Technometrics* **16** 499-511.
- [13] Devlin, B. and Risch, N. (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **96**, 523-536.
- [14] Scheet, P. and M. Stephens, (2006), A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase, *American Journal Human Genetics*, **78**(4), 629-644. Doi: 10.1086/502802
- [15] For example: Hafler D.A., et al., (2007). Risk Alleles for Multiple Sclerosis Identified by a Genomewide Study, *N. Engl. J. Med.* **357** 851–862. doi: 10.1056/NEJMoa073493.