Parametric Resampling Methods for Retrospective Changepoint Analysis

by

Jonathan W. Duggins

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Statistics

Eric P. Smith, Chairman
Dong-Yun Kim
William H. Woodall
Pang Du
Feng Guo

June 25, 2010
Blacksburg, Virginia

Keywords: Changepoint, bootstrapping, resampling, threshold detection, frequentist

Parametric Resampling Methods for Retrospective Changepoint Analysis

Jonathan W. Duggins
Virginia Polytechnic Institute and State University, 2010
Advisor: Eric P. Smith, Ph.D.

Abstract

Changepoint analysis is a useful tool in environmental statistics in that it provides a methodology for threshold detection and modeling processes subject to periodic changes in the underlying model due to anthropogenic effects or natural phenomena. Several applications of changepoint analysis are investigated here. The use of inappropriate changepoint detection methods is first discussed and the need for a simple, flexible, correct method is established and such a method is proposed for the mean-shift model. Data from the Everglades, Florida, USA is used to showcase the methodology in a real-world setting. An extension to the case of time-series data represented via transition matrices is presented as a result of joint work with Matt Williams (Department of Statistics, Virginia Tech) and rainfall data from Kenya, Africa is presented as a case-study. Finally the multivariate changepoint problem is addressed by a two-stage approach beginning with dimension reduction via principal component analysis (PCA). After the dimension reduction step the location of the changepoint in principal component space is estimated and assuming at most one change in a mean-shift setting, all possible submodels are investigated.

Dedication

To Helen, for making life's road worth traveling while we had a chance.

Acknowledgements

I would be remiss if I did not thank my advisor, Dr. Eric P. Smith, for his patience and support during not only the dissertation process, but during my entire tenure as a doctoral student. Similarly the feedback from my other committee members, Dr. Bill Woodall, Dr. Dong-Yun Kim, Dr. Pang Du and Dr. Feng Guo, was invaluable. Of course, the Virginia Tech community is much broader than just the department of statistics and I owe an enormous amount of gratitude to the community at large. In particular my time working with the GSA allowed me a chance to meet colleagues and make friends across all disciplines, an invaluable opportunity for any graduate student – but especially for a statistician. Last, but certainly not least I want to thank my friends and family. For the last decade you have helped me cope with the insanity of school, Tomi, Stef, Ian and Megan. I owe you more than you know. Mom, Dad and Flatt, there is no way to measure how much I have appreciated your counsel over the many, many, many years I have been in school. Thank you to all my friends and family for sticking it out with me.

Contents

List of Tables

List of Figures

CHAPTER 1

INTRODUCTION

The changepoint problem is certainly not a new area of interest in the field of statistics, whether applied or theoretical, and it can be thought of as a search for homogeneous subsets in a collection of data that is ordered with respect to some gradient. When a point is found such that the data is no longer considered to be homogeneous, using some form of discrimination rule, a changepoint is said to have occurred in the data. This heterogeneity may be due to a change in mean, variance, skewness, distributional family or any other characteristic of the data. In more complex settings more than one changepoint may exist in the collected data and the changes may be in different directions, e.g. an increase in mean followed by a decrease in mean, a change in different parameters such as a mean shift followed by a variance shift, or even a change from one distribution to another at each change, e.g. a mixture distribution problem.

More formally, let $\{y_i\}_{i=1}^n$ be a sequence of observed data with $y_i \sim F_i$. The question of interest is whether $F_i \equiv F$ or if there exist $1 \leq \tau_1 < \tau_2 < \cdots < \tau_k \leq n$ such that

$$F_i = \begin{cases} F_1 & 1 \leq i \leq \tau_1 \\ F_2 & \tau_1 < i \leq \tau_2 \\ \vdots & \vdots \\ F_{k+1} & \tau_k < i \leq n \end{cases},$$

thus dividing the data into $k+1$ homogeneous groups. These groups are commonly referred to as regimes or segments and the gradient i typically represents time or some other physical gradient, especially in ecological and environmental settings. The $\{\tau_j\}_{j=1}^k$ are the changepoints and are sometimes referred to as *joinpoints* or *split-points*. Some of the earliest work on changepoint problems differed from this approach in that, rather

than considering the data to be already collected, it attempted to detect the changepoint as each new data point was collected. In this way Shewhart [1939] introduced the very first control charts that are still currently in use in quality control applications. Both the prospective approach (or real-time or on-line setting), popular in quality control and health surveillance, and the retrospective approach, popular in environmental statistics, are still generating great interest among researchers, both inside and outside, the field of statistics.

The changepoint problem has received attention from both frequentists and Bayesians, developing an extensive history in both parametric and nonparametric literature. Bhattacharya [1994] gives a relatively thorough overview of changepoint analysis as it had developed into the mid-1990s. Recent advances in computing resources have allowed for the development and implementation of more computationally complex approaches including expansive simulation studies to support theoretical results, computation of approximate $p-values$ in significance tests and allowing for the employment of bootstrap, randomization and other resampling based tests. As such there has been growing attention in evaluating the fit of changepoint models and providing accurate significance and confidence levels for inference in changepoint problems. It is this aspect of the changepoint problem that is the main thrust of the work discussed here.

While some research has been carried out without assuming independent observations, a large majority of the changepoint literature is devoted to the detection of a changepoint under the assumption of independence. In particular research on weakly dependent data and data exhibiting spatial autocorrelation is becoming more common. The spatial changepoint problem is of great interest in environmental statistics, where most data sets have some spatial component. The development of accurate and efficient inferential techniques in this area is not the goal here, but comments on the applicability of the methods detailed here to spatial changepoint problems will be made when appropriate.

CHAPTER 2

LITERATURE REVIEW

The field of changepoint analysis is an ever burgeoning one, leading to a vast literature discussing many aspects of such analysis. The wide variety of applications, from industrial to biological, gives proof to the need for such analysis but also gives rise to multiple approaches to most changepoint problems. In industrial and health care applications, for example, the analysis usually proceeds sequentially, typically using control charts or stopping rules to perform real-time monitoring. Lai [2001] gives a review of problems in sequential analysis, including a discussion about sequential changepoint detection in quality control. In biological and ecological settings the interest is more often in retrospective analysis, i.e. detection of a changepoint given the entire set of collected data. Examples include modeling dose-response curves, ground water monitoring, climate change and many others.

Regardless of whether the interest is in sequential or retrospective changepoint analysis (SCPA or RCPA), one typically first decides on whether a non-, semi- or fully-parametric approach is of interest. Semi- and nonparametric methods have gained popularity as modern computing has allowed for complicated and computationally intensive algorithms to be routinely applied. The framework of changepoint analysis typically requires answers to the following three questions.

- What is the true model?
- How many changepoints are truly present?
- Where are the changepoints truly located?

Changepoint problems vary based on the approaches taken to answer these three

questions. As in any modeling problem, choosing a model requires decisions to be made about both the deterministic and stochastic portions of the model. Changepoint problems often have the deterministic portion of the model specified, with the label of parametric or nonparametric referring to the stochastic portion of the model. In RCPA common choices for the deterministic portion of the model in each segment tend to belong to the class of piecewise linear functions and which may or may not be constrained to be continuous, in particular a step-function or hockey-stick model are often used [Barrowman and Myers, 2000]. The nonparametric treatment of such models simply avoids placing a distributional assumption on the collected data. For simplicity suppose that at most one changepoint exists in the data, then the classical nonparametric, RCPA problem is given $x_1, x_2, \ldots, x_n$ to determine whether or not a τ exists such that for $1 \leq \tau \leq n$ $x_i \equiv F_1$ for $i \leq \tau$ and $x_i \equiv F_2$ for $i > \tau$. If such a τ exists the estimation of the location of this changepoint is of obvious interest.

2.1 Quality Control

Statistical process control (SPC) is primarily concerned with establishing the stochastic setting that represents the "in-control" state for a particular system then detecting any deviation from this in-control state as soon as possible after it occurs. This "out-of-control" state is then analyzed and the nature of the change as well as the time at which it occurred are estimated and steps are theoretically taken to ensure that the in control state is recovered, assuming that the in-control state is preferred – as is usually the case. This on-line or sequential analysis of the data is an incarnation of the changepoint problem where the data is updated with each new observation collected and subsequently reanalyzed. It lends itself particularly to the problem in which at most one change exists since in SPC changes are almost always unwanted and so the process is stopped immediately when a change is detected and corrective action is typically applied. The changepoint problem in SPC has received more attention of late since the traditional charting methods, Shewhart, CUSUM, and EWMA in particular, rely on having *a priori* knowledge as to the parameters that define the in control situation. Typical quality

control settings where these charting techniques are to be applied rely on an initial collection period, termed Phase I, in which data is collected to establish estimates of the parameters. In the monitoring period, called Phase II, those estimates are taken as the true values and the on-line analysis of the data collected begins. Unfortunately the consequences of ignoring the variability inherent in these estimates being treated as true values can affect the average run length (ARL) of the methods, which is clearly undesirable [Hawkins et al., 2003]. Jensen et al. [2006] give a review of the literature discussing the effects such oversight can have.

According to Hawkins et al. [2003] one of the primary benefits to the changepoint formulation of the on-line detection problem is that no longer requires the distinction between Phase I and Phase II data collection, since all parameters can be treated as unknowns to be estimated and analysis can be started sooner than in traditional methods. Hawkins et al. propose a control chart based on the goal of detecting a change in normally distributed data. Specifically they take

$$X_i \sim N(\mu_1, \sigma_1^2) \quad \text{for} \quad i = 1, 2, \ldots, \tau$$

$$X_i \sim N(\mu_2, \sigma_2^2) \quad \text{for} \quad i = \tau + 1, \ldots, n$$

where τ is the unknown location of the changepoint, if it exists, and n is not fixed. To preserve the desirable property that the conditional probability of a false alarm at time n given that no false alarm has occurred up to $n-1$ is constant they consider a set of control limits indexed by n but which depend on α, and which they denote $h_{n,\alpha}$. Here α is the false alarm rate desired. The lack of analytically obtainable values for the control limits led the authors to provide a suite of simulations allowing for implementation of their methodology, for which efficient computing schemes were suggested. Performance evaluation of the changepoint formulation shows that it is not superior to the closest comparable method, but that it was sometimes-superior while never being far from the best choice. The gains from not requiring a Phase I study and incorporation of the estimates inherent variability are suggested to more than compensate, thus making this analytic approach an the superior one under consideration.

Hawkins and Zamba [Hawkins and Zamba, 2005] continue this approach by provid-

ing an additional component that, in the case of a changepoint being detected, look at the traditional two-sample t- and F-tests for mean and variance respectively to determine the most likely cause for the signal. The authors are careful to point out that the significance of these tests is not to be trusted since they were split based on a changepoint estimated from the data, but that they are still useful for determining whether the shift was in the mean, the variance, or both. Again, simulations provide the $h_{n,\alpha}$ for easy implementation, as well as a function for approximating the h for other α and n not included in their work. Simulations in the mean shift only, variance shift only, and mean and variance shift cases show the method to be very useful for detecting changepoints. Furthermore, it is shown that in the case where a changepoint is present their method may signal even when the individual t and F components would not have signaled, which the authors consider a success since it implies a shorter out-of-control ARL. Zamba and Hawkins [Zamba and Hawkins, 2006] further extend this work to the multivariate case following a very similar formulation.

Perry and Pignatiello Jr. [2008] investigate the changepoint problem for the on-line problem for data following any distribution belonging to the exponential family. However, they are only interested in a change in the location parameter and based on this interest use a max-log-likelihood approach to estimate the changepoint. In addition to the traditional estimation problem the authors were concerned with demonstrating that their method could detect changepoints even in data for which the in-control portion included natural variability, or seasonality. Trigonometric functions were incorporated to show that even in the case of a sinusoidal steady-state model, accurate changepoint estimates could still be provided. Another application is provided by Mahmoud et al. who focus on a fundamentally different problem of detecting changepoints in profile data [Mahmoud et al., 2007]. They assume the profiles are linear and that no changepoints occur within each profile and then develop a method for detecting whether or not changes occur between profiles. Specifically they take the model to be $y_{ij} = A_i + B_i x_{ij}$ in profile i and test for changes in A_i and B_i between profiles. Performance evaluations show their likelihood ratio test to be superior in many cases while being competitive in many others.

2.2 Nonparametric RCPA

Even focusing on retrospective changepoint analysis (RCPA) still leaves a breadth of problems to address, in particular how exactly to obtain the best answers to the questions posed above regarding the model between, number of, and location of, the true changepoints. Carlstein [1988] points out that most estimation methods depend on the distributions (1) belonging to a specific parametric family or (2) simply differing with respect to some measure of center. Carlstein then proposes three estimators for the changepoint without requiring either (1) or (2) to be met, and in fact the distributions can be discrete, continuous, or mixed and the supports may even be unknown. His three estimators are proposed and through simulation studies and examples one is found to be superior, according to Carlstein. This estimation approach is generalized by Dümbgen [1991] who additionally provides a bootstrap based technique for creating confidence sets for the true changepoint. In both Carlstein [1988] and Dümbgen [1991] the assumption that exactly one changepoint exists is made, and the focus is primarily on estimating the changepoint based on this presupposition.

The issue of testing hypotheses about changepoints using nonparametric methods is linked to the estimation problem since such inference necessitates point estimates of the changepoint be available. Early nonparametric tests for a simple mean shift were developed by Page [1955] and Bhattacharyya and Johnson [1968]. Page assumes a change in mean is of interest but does not make assumptions regarding the data's distribution. He then proposes a sign-test approach for the case where the data are analyzed assuming the initial mean is known, but does not address the case where the initial mean is unknown.

Bhattacharyya and Johnson cover both the known and unknown initial level (not necessarily the mean) when the alternative of interest is an increase (which they label Δ) in the level. For both cases they provide test statistics which are optimal with respect to the derivative of the average power at $\Delta = 0$, where the average is taken over m. Additionally their tests are shown to be invariant to any transformation, $h(\cdot)$, of the data where h is continuous, odd and strictly increasing for the case of a known initial level

and where h is continuous and strictly increasing for the unknown initial level.

Zou et al. [2007] consider a similar question, but rather than specify which of the distribution's moments are to change, they only specify that after a certain point, the distribution switches from F to G. Specifically they assume that for some data $\mathbf{y}$ it is the case the $\int \mathbf{y} d\mathbf{F} \neq \int \mathbf{y} d\mathbf{G}$, so that at least one moment of the distribution F changes at some point which they label k^*. Their test is based on testing whether or not a change occurs in $\int \mathbf{y}_k d\mathbf{F}_k$ for some $k = 1, \ldots, n$ and makes use of an empirical likelihood ratio test as an analog of the traditional likelihood ratio test available in the parametric setting discussed below. Results are provided regarding the asymptotic null distribution as well as the consistency of the associated estimator.

The nonparametric estimation of potential changepoints is also prevalent in a regression context where a changepoint, τ, may exist in the regression curve, f, or in $f^{(p)}$ for some $p \in \mathbb{Z}_+$. Müller provides a kernel-estimation based method for estimating τ and fitting a subsequent model given the estimate [Müller, 1992]. Loader [1996] provides improved results on the same problem by using different kernels, leading to improved convergence rates and, at least in the example provided in Loader [1996], less bias in the estimates of the amount of change. Horváth and Kokoszka use a similar approach, but by using polynomial fits rather than kernel-based fits they provide a test of continuity of $f^{(p)}$ and thus a test for the presence of a changepoint [Horváth and Kokoszka, 2002].

2.3 Parametric RCPA

While nonparametric research has focused on point estimation and hypothesis testing it has not been as extensively researched nor as widely applied as its parametric counterpart. This is in part due to the availability of maximum likelihood and Bayesian methods when working in a parametric setting. In the parametric community there is also a great deal of interest in the estimation of, and significance testing for, changepoints in regression settings. Confidence sets for changepoints are also of interest and are covered as well, though to a lesser extent. There are many changepoint models available in the literature, from simple abrupt mean shifts assuming equal variances to more

general piecewise linear regression models, both continuous and discontinuous. Early work on estimating the changepoint when at most one is assumed to exist and where the data follows a linear regression in each segment is provided by Quandt [1958] who postulates a likelihood ratio based statistic that does not assume constant variance but does not constrain the estimated model to be continuous. In subsequent work he provides numerical results that suggest that the negative doubled logarithm of the likelihood ratio does not follow a chi-square distribution, but without any theoretical justification [Quandt, 1960].

In the analysis of changepoints in linear regression models the bulk of the literature is focused on situations in which the regressor $\mathbf{x}$, which may be scalar or vector valued, is non-stochastic. Kim and Siegmund mention both an analysis for a stochastic and non-stochastic regressor, but primarily focus on the fixed x case [Kim and Siegmund, 1989]. In contrast, Koul and Qian [2002] specifically address the issue of changepoint analysis in a two-phase, i.e. one changepoint, linear regression model with a random regressor. Koul and Qian allow for changes in the intercept or slope to occur at the changepoint and provide maximum likelihood estimators for each of the four model parameters as well as the changepoint. Additionally the consistency of these estimators are provided and Monte Carlo simulations in the cases of normal, logistic and Student's-t errors are provided to support the theoretical results. The smallest sample size used is $n = 100$ which does lead to estimates which are very close to the true values and for which the coverage probabilities (computed via $\overline{x} \pm 2SE$ by Koul and Qian) are all above 90% and the other sample sizes investigated ($n = 200, 500$) of course show even higher probabilities. However, the results are not evaluated for sample sizes less than 100 which are more common in ecological applications since larger data sets tend to have spatial, temporal or spatio-temporal autocorrelation components that cannot be ignored.

As in other branches of statistics likelihood ratio based methods show continued popularity in the area of changepoint analysis, in part due to their asymptotic properties. James et al. [1987] expand the likelihood ratio concept of Quandt [1958] and provide a likelihood ratio test statistic for the case of a simple mean shift. In addition the authors give a modified likelihood ratio test statistic based on restricting the search

area for the changepoint by bounding it away from the endpoints of the data. Along with these two statistics the authors present three alternatives previously suggested by other authors and compare them based on an evaluation of their approximate power. In particular they compare their two likelihood ratio statistics to the recursive residual approach developed by Brown et al. [1975], the score based estimator of Pettitt [1980] and the Bayesian estimator of Chernoff and Zacks [1964]. The result of this comparison is that the optimal test, in terms of power, depended both on the location of the changepoint (whether τ falls early or late versus falling in the middle of the observed sequence) and on how small the smallest segment was allowed to be, e.g. setting a minimum group size of three versus five observations. As a result no optimal test is stated, but the modified likelihood ratio statistic, proposed by the authors, is suggested since it allows for easy estimation of the changepoint and the amount of mean shift that occurs, while performing moderately better at the boundaries where most changepoint estimation methods suffer [James et al., 1987].

Returning to the analysis of changepoints in regression problems, Kim and Siegmund [1989] consider the case of a linear regression model with at most one changepoint where the lone regressor is non-stochastic and the model is not constrained to be continuous. Specifically they consider $\{Y_i\}_{i=1}^n$ being i.i.d. normal with common variance σ^2 and where $\mathrm{E}[Y_i] = \alpha_0 + \beta_0 x_i$ for $i \leq j$ and $\mathrm{E}[Y_i] = \alpha_1 + \beta_1 x_i$ for $i > j$ without any restriction forcing continuity onto the model. In this setting the hypotheses of a change in α only or a change in both α and β simultaneously were of interest and likelihood ratio statistics are derived. Through the usage of Gaussian processes, approximate tail probabilities are given for their statistics and numerical results are presented, showing the approximations to be very good, even for sample sizes as small as $n = 10$ when testing for a change in α alone. When testing for a simultaneous change the results are again presented with numerical justification and again with very reasonable approximations, though the smallest sample size now considered is $n = 20$. However, these results are obtained assuming equally spaced data, specifically $x_i = i/n$, which is often not the case for environmental application and the consequences of which are only briefly mentioned. Furthermore the variance is assumed to be constant across the entire range of

the data which is unnecessarily restrictive and often violated in practical applications.

As an extension of this approach Gurevich and Vexler [2006] consider a very flexible class of two-phase, i.e. one changepoint, linear regression models. Their model is

$$y_{ij} = \left(\beta_{00} + \beta_{01} x_{1i} + e_{0ij}\right) I\{q_i < \gamma\} + \left(\beta_{10} + \beta_{11} x_{1i} + \beta_{12} x_{2i} + e_{1ij}\right) I\{q_i > \gamma\}$$

for $j = 1, \ldots, n_i$ and $i = 1, \ldots, n$. Here the y_{ij}, x_{1i}, x_{2i} and q_i are observed and the e_{1ij}, e_{2ij} are independent error terms with possibly different densities. The q_i is referred to as the threshold variable and serves as the "gradient" along which a search will proceed to find the estimate of the changepoint, γ. In this way they allow the model to be piecewise linear, without a continuity constraint, in such a way that several common models are special cases of this Gurevich-Vexler formulation. In particular, let $q_i = i$, $\beta_{12} \equiv 0$ and both error terms distributed $N(0, \sigma^2)$ with σ unknown to get the model

$$y_{ij} = \begin{cases} \beta_{00} + \beta_{01} x_{1i} & i \leq \gamma \\ \beta_{10} + \beta_{11} x_{1i} & i > \gamma \end{cases}$$

which is equivalent to the model of Kim and Siegmund above [Kim and Siegmund, 1989]. By setting $\beta_{01} = \beta_{11} \equiv 0$ and again taking $q_i = i$ the discontinuous hockey-stick model is recovered.

In Gurevich and Vexler [2006] tests are provided based on maximum likelihood methods whose power asymptotically approaches one under certain conditions on the error densities and sample size. Two tests, one Bayesian and one based on invariant statistics, are given and upper bounds on their significance levels are provided. In the case where the null hypothesis of no changepoint is rejected the authors provide a maximum likelihood estimate for the changepoint similar in concept to that of Koul and Qian [2002]. The tests and subsequent estimator are applied to an example and results from Monte Carlo simulations are given to shed light on the conservative nature of the upper bound placed on the significance levels as well as in an attempt to verify the asymptotic power of the tests does approach one. From the results it is clear that the provided methodology is quite conservative, having empirical type I error rates whose size tends to be approximately 80% of the nominal level, even for $n = 110$ which was the smaller of the

two sizes studied. For $n = 170$, the only other size under consideration, the rates were marginally better. Also of interest is that the method appears to be rather asymmetric in terms of empirical power, with both cases showing a near 50% drop in power when the true changepoint is moved from 15 to $n-15$, though less of a drop in power is noted when moving from 25 to $n-25$, as expected.

Horváth [1993] concerns himself with a type of simultaneous changepoint problem, specifically focusing on testing for a change in the mean-process or variance-process of a set of n, i.i.d. random normal variates. Specifically Horváth has $\{X_i\}_{i=1}^n$ with $X_i \sim N(\mu_1, \sigma_1^2)$ if $1 \leq i < \tau$ and $X_i \sim N(\mu_2, \sigma_2^2)$ if $\tau \leq i < n$ and wants to test

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_n \quad \text{and} \quad \sigma_1 = \sigma_2 = \cdots = \sigma_n$$

versus

$$H_1 : \quad \mu_1 = \mu_2 = \cdots = \mu_{\tau-1} \neq \mu_\tau = \cdots = \mu_n$$

or

$$\sigma_1 = \sigma_2 = \cdots = \sigma_{\tau-1} \neq \sigma_\tau = \cdots = \sigma_n$$

which clearly assumes if there is a shift in both the mean and variance processes, then they shift simultaneously. Horváth derives the maximum likelihood estimator Λ_n and gives a result for the asymptotic distribution of the square-rooted logarithm of the statistic, $\lambda_n = \sqrt{\log \Lambda_n}$. Numerical results based on Monte Carlo simulations are given as support for using the approximation for moderate sample sizes (n=20, 50, 100 are used in Horváth [1993]) but the probabilities are not given; instead the percentiles are provided. The degree of agreement between the empirical percentiles and those calculated from the asymptotic distribution is satisfactory, but the slow rate of convergence is apparent. Nevertheless, in simulations studies the method was shown to have only slightly lower than nominal type I error rates and reasonable power, even for sample sizes of 40, in most cases [Duggins et al., submitted].

Jandhyala and Fotopoulos [1999] focus on the estimation of the changepoint in the case of a single changepoint and, by way of the discreteness of the time steps, derive upper and lower bounds for the probability that the estimate is exactly j observations

away from the true changepoint. Strictly speaking the authors find the bounds on these probabilities based on the MLE using an infinite sample size. Furthermore, they assume that both the changepoint, τ, and $n-\tau$ are infinitely large. They point out that for their asymptotic bounds on the probability that $\hat{\tau}_\infty - \tau = \pm j$ then require that the changepoint be "away from both tails of the data" [Jandhyala and Fotopoulos, 1999] but give no indication as to how far from the tails is sufficient. Additionally, no discussion of the specific consequences for application of their bounds when this caveat is violated are discussed. Simulation results for the normal and exponential cases are show the resulting probabilities for their bounds as well as simpler approximations to those bounds that they derive. Exactness of the bounds for the exponential case is also established.

The application of parametric RCPA has not been entirely limited to normal distributions since other common distributions are popular for modeling real world phenomena. Jandhyala and Fotopoulos [2007] provide an estimate of the single changepoint occurring in a sequence of random variables following the lognormal distribution. Following closely the pattern the authors used to establish bounds in Jandhyala and Fotopoulos [1999], they establish lower and upper bounds as well as an approximation for $P(\hat{\tau}_\infty - \tau = j)$ and provide simulation results showing the sharpness of their bounds. Again, the results are asymptotic and hold for $\tau \to \infty$ and $n-\tau \to \infty$, with the bounds being valid when τ is reasonably far away from the boundaries. Víšek [1980] uses the likelihood ratio method to search for a change in the parameters of data following a double exponential (Laplace) distribution. By assuming the magnitude of the change in the parameters tends to zero as n increases the asymptotic distribution of the test statistic is established. The case of the gamma distribution is addressed by Ramanayake [2004], though she only considers changes in the shape parameter – however, she provides both a Bayesian and frequentist based statistic. For the Bayesian statistic, critical values are given via simulations and from an Edgeworth expansion. Additionally, power curves are calculated and shown for various scenarios. A standardization of the likelihood based statistic is shown to converge in distribution to a standard normal and both methods are applied to two examples highlighting their usefulness, but also highlighting the low power of the Bayesian method when the changepoint is near the boundary.

While the great majority of the work on changepoint problems focuses on univariate data with no more than two parameters, some work on changepoint analysis in multivariate parameter settings has been done. Gombay and Horváth [1994] provide an extension of Horváth's work from Horváth [1993]. Significance testing for the case of $\mathbf{X}_i \sim F(\cdot|\boldsymbol{\theta}_i)$ where the $\{\mathbf{X}_i\}_{i=1}^n$ are independent is considered when the null hypothesis is that the parameter vectors are all equivalent, while the alternative is that for some k, $1 \leq k < n$, $\boldsymbol{\theta}_1 = \cdots = \boldsymbol{\theta}_k \neq \boldsymbol{\theta}_{k+1} = \cdots = \boldsymbol{\theta}_n$. Gombay and Horváth provide the maximum likelihood statistic as well as the proof that the asymptotic distribution of the centered and scaled statistic is, what the authors refer to as, a double exponential distribution. This double exponential has distribution function $F(t) = \mathrm{e}^{-2\mathrm{e}^{-t}}$ and so is not the traditional double exponential, or Laplace, distribution as given by, say, Casella and Berger [Casella and Berger, 2002]. Though this nomenclature is not uncommon in some changepoint literature, this $F(t)$ is often referred to as double exponential even though it is the one of the Gumbel distributions [Evans et al., 2000].

2.4 Bayesian RCPA

The Bayesian approach to changepoint analysis has not been explored for as long as its frequentist counterpart, but that has not kept it from broaching the issues of estimation and testing. Early work in the area of Bayesian changepoint analysis was done by Chernoff and Zacks [1964] and Sen and Srivastava [1975]. Chernoff and Zacks motivate their interest by describing a situation in which the goal is to correctly estimate the current value of some process that is proceeding along a path that is subject to potential periodic changes in the mean of the process. Hence estimation of the changepoints are necessary to obtain a more accurate estimate. They proceed by assuming the n random variables are independent and follow some parametric distribution and that the amount of change at time i is normally distributed with variance σ^2 and mean μ_i, which itself follows a normal distribution with mean 0 and variance τ^2. By letting the hyperparameter τ^2 tend to zero an estimator for μ_n, the current mean, is derived under quadratic loss when the prior is uniform on the set of real numbers.

In addition to this Bayes estimator a minimum variance linear unbiased (MVLU) estimator is derived as well, which while simpler, is less efficient in the case of more than one or less than $n-1$ changes. Two alternatives are provided, the first by allowing the variance of the mean shifts, σ^2, also tend to zero. Unfortunately this estimator is inefficient in the case of large changes. A second alternative is provided under the assumption that an *a priori* distribution is placed on the time points such that at most one change occurs and is called the AMOC (at most one change) Bayes estimate. Again problems arise if this estimator is used blindly since it performs poorly when more than one change exists, e.g. in the case of two changes where the first change is larger the AMOC Bayes estimate acts as if the first change were the only one present [Chernoff and Zacks, 1964].

In an extension of the work done by Chernoff and Zacks, Sen and Srivastava [1975] provide the Bayesian estimators for the following four cases.

1. A one-sided test for a mean shift when the variance is known, but the initial mean and location of change are unknown.

2. The exact same setting as 1, but the initial mean is known.

3. A two-sided test for a mean shift when the variance is known, but the initial mean and location of change are unknown.

4. The exact same setting as 3, but the initial mean is known.

Bayesian test statistics for Case 1 and Case 2 are given by Chernoff and Zacks [1964] and by Gardner [1969] for Case 3. Sen and Srivastava extend the work done by Gardner to obtain the Bayesian test statistic for 4 as well as deriving the maximum likelihood statistics for, r, the location of the changepoint. Furthermore, Sen and Srivastava provide numerical and graphical comparisons of the Bayesian and maximum likelihood approaches in terms of power [Sen and Srivastava, 1975]. The authors' conclusion is that neither the Bayesian nor MLE method is uniformly superior, with the optimal method depending on the size of the shift and the location of the changepoint.

Despite the thorough treatment given by Sen and Srivastava, their work not only assumes that the variance is constant but also that it is known [Sen and Srivastava, 1975]. Even in the somewhat unlikely case that even a moderate variance change is not present, the assumption that variance is known is quite untenable. Smith [1975] provides a detailed extension to the case where the variance is unknown and even possibly unequal for the case of normal errors. Smith generates a random sample and through the calculation of the posterior probabilities determines the estimate of the changepoint and tests the hypothesis of no change. Lee [1998] derives the posterior distribution for the changepoint for any member of the exponential family and applies the results to examples in the special cases of normal and gamma densities. Qian et al. [2003] use the hierarchical Bayesian approach of Smith [1975] as well as a model deviance based method on a data set from the Everglades where the methods are shown to have similar point estimates with comparable significance levels, but with the Bayesian method having narrower interval estimates.

2.5 Multiple Changepoints in RCPA

One major drawback to these methods is their assumption that at most one changepoint exists in the data. Fixing the number of changepoints has the benefit of greatly simplifying the required analytical procedures while estimating them from the data adds another layer of complexity to an already computationally difficult question. In practice it is not unreasonable to assume that some historical data is available to provide some rationale for setting the number of changepoints. However, it is equally as plausible that a situation may be encountered where no data is available and an any *a priori* statement regarding the number of changepoints may be viewed with skepticism. One popular method for dealing with such situations is a regression tree algorithm, such as CART that uses a binary splitting approach in order to provide such estimates [Breiman et al., 1993]. With such algorithms the number of and location of changepoints can be estimated. These binary splitting methods tend to provide relatively quick results but without providing the optimal split points of the series of data when two or more exist

[Hawkins, 2001].

Bai and Perron [1998] consider the issue of multiple changepoints, which they refer to as structural changes, in a linear model. They posit a model with m such structural changes where $y_t = \mathbf{x}_t^T \boldsymbol{\beta} + \mathbf{z}_t^T \boldsymbol{\delta}_j + u_t$ for $t \in (T_{j-1}, \ldots, T_j)$ with $j = 1, \ldots, m+1$. In their model the regressors in $\mathbf{x}$ are associated with coefficients in $\boldsymbol{\beta}$ that do not change from segment to segment while the regressors in $\mathbf{z}$ are associated with the coefficients in $\boldsymbol{\delta}$ which potentially change as a function of j, the index. The random errors are denoted by u_t. The estimation is via least-squares and the subsequent estimates of the break points and the model parameters are shown to be consistent estimates and their limiting distributions are derived. A test statistic is provided for testing the null hypothesis of no changepoints versus an alternative in which the number of breaks is some specific alternative value, k. Asymptotic critical values are provided for several possible scenarios. The case where m is unknown is also discussed, though briefly. Bai and Perron [2003] investigate the computational aspects of the problem they discussed in Bai and Perron [1998]. In particular they provide an efficient algorithm based on dynamic programming for the computation of the parameter estimates under the assumption that breaks in variance are allowed assuming they are simultaneous with a change in the mean, as is the case with Horváth's approach in Horváth [1993]. Bai and Perron apply their algorithm to several examples, highlighting the usefulness of their least-squares algorithm.

When investigating a potential change in the distribution of collected data most developers of testing criteria have assumed, as Carlstein points out in Carlstein [1988], that there is some shift in the mean of a distribution with a given form. Many times it is even assumed that the shift is in the positive direction and only one-sided alternatives are considered. This is particularly true when looking for multiple change points as they become increasingly difficult to detect if the data were not monotone in its changes, but perhaps oscillated between two states of nature instead. In an extension of this one-sided approach to the multiple changepoint problem, Aly et al. [2003] propose a test for the null hypothesis of no change in distribution, F, of $X_1, \ldots, X_n$ against the ordered alternative

$$H_1 : F_1 = \cdots = F_{\lfloor n\lambda_1 \rfloor} \prec \cdots \prec F_{\lfloor n\lambda_k \rfloor + 1} = \cdots = F_n$$

where k is the number of changepoints, $0 < \lambda_1 < \lambda_2 < \cdots < \lambda_k < 1$ and F_i is the distribution of X_i and $\prec$ indicates a partial ordering of the $\{F_i\}_{i=1}^n$. This partial ordering is presumably based on the fact that F are changing in such a way as to make $\{X_i\}_{i=1}^n$ a stochastically increasing sequence of random variables.

Test statistics are provided for the case where the change is, in fact, only in the location parameter and for the more general case described by H_1. Asymptotic critical values are given for both tests and empirical power results are presented based on Monte Carlo simulations. In both results the assumption that $k \geq 1$ is made and no discussion is given to indicate the consequences of applying these tests when this assumption is not met, i.e. there is no presentation of empirical type I error rates. Furthermore, there is no mention of the situation in which both H_0 and H_1 were false and what the power of these tests may be in those cases, say if the true situation were the oscillation mentioned above or in the cases of a stochastically decreasing, rather than increasing, set of data.

Douglas Hawkins provides an alternative method based on assuming the data comes from any member of the exponential family [Hawkins, 2001]. He develops a dynamic programming algorithm based on solving the "outer" problem of finding the maximum likelihood estimates of the changepoints then the "inner" problem of finding the MLEs of the parameter vector for the distribution of interest, e.g. normal or Poisson. The computing time is established to be nearly linear in the number of changepoints so that fitting a large number of changes is not computationally burdensome. A procedure for testing hypotheses regarding the number of changepoints needed is provided, based on a generalized likelihood ratio test that looks at the difference in log-likelihood between models as compared to the log-likelihood of the "larger" model, i.e. the one with more changepoints. The tests, one for known variance and one for estimated variance, necessarily assume there is constant variance between segments which limits the applicability of the method to problems where only a change in mean is of interest with $\sigma_1 = \sigma_2$. While more general than Horváth's test in Horváth [1993] in that Hawkins can fit multiple changepoints, it is less flexible in that it can not detect a change in the variance and, in fact, assumes one is not present. Furthermore the generalized test statistics of Hawkins have no apparent asymptotic distribution and therefore his suggestion, though

he admits it is unsatisfactory, is the use of heuristic methods such as scree plots for determining the correct number of changepoints.

2.6 Multivariate Changepoints in RCPA

The multivariate changepoint problem has been taken to have different interpretations by different authors. The majority of work in the area described as multivariate changepoints by the authors is in the case of multiple explanatory variables but still a single response, i.e. $y = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, rather than the case where there are multiple responses per experimental unit (not to be confused with subsampling). To draw a distinction here I will denote the two scenarios as *multivariate-E changepoint analysis* when there are multiple predictors but a single response and *multivariate-R changepoint analysis* when there are multiple responses - regardless of the number of predictors. In addition to their work mentioned above, Sen and Srivastava [1973] examine the single changepoint problem in the multivariate-E setting as well.

Specifically they take $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^m$ are independently distributed multivariate normal with means $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n$ respectively and which have $\boldsymbol{\Sigma}$ as their covariance matrix. They consider two cases of testing the hypothesis

$$H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_n = \boldsymbol{\mu}$$

versus the alternative

$$H_1 : \boldsymbol{\mu} = \boldsymbol{\mu}_1 = \ldots = \boldsymbol{\mu}_r \neq \boldsymbol{\mu}_{r+1} = \cdots \boldsymbol{\mu}_n,$$

where in both cases the changepoint r is unknown and where the initial mean, $\boldsymbol{\mu}$, is known in one of the cases and unknown in the other. In addition to providing test statistics for both cases the provide a proof of the fact that the given test statistics are not asymptotically equivalent. In fact, they prove that the ratio of their expected values converges to three as the sample size approaches infinity. Additionally, for dimensions between two and eight ($2 \leq m \leq 8$) tables of the asymptotic percentiles for both the known and unknown mean case are given so that tests may be carried out.

Chen and Gupta [1995] investigate the multivariate-E scenario in a setting with a sequence of n independent m-dimensional Gaussian random vectors with $\mathbf{x}_j$ having parameters $\left(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)$ and intend to consider testing for the presence of q changepoints, with q unknown. Specifically the changepoints, if they exist, denote locations of points where *both* the mean process and covariance process exhibit a change. The authors proceed by first stating the maximum likelihood statistic for testing the presence of a single (simultaneous) changepoint and then, after a transformation, deriving the asymptotic distribution of the statistic of interest. There is no discussion as to whether the size of this test would be preserved when repeatedly applied in the search for the q unknown changepoints, and in particular whether the $p-values$ from the asymptotic distribution would be adversely affected by the repeated application of the procedure segments which are monotonically decreasing in sample size after each successive application. However, the authors do cite Vostrikov's binary segmentation procedure [Vostrikova, 1981].

A Bayesian approach to the multivariate-E changepoint problem in the case of multivariate normal data was investigated by Son and Kim [Son and Kim, 2005]. The authors consider a random sample, $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^p$, which are independent, multivariate normal random variables which have mean $\boldsymbol{\mu}_1$ and covariance matrix $\boldsymbol{\Sigma}_1$ under the null hypothesis of no change. They consider three possible alternatives: a change in $\boldsymbol{\mu}$, a change in $\boldsymbol{\Sigma}$ or a change in both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. In each case the change takes place at some location r, with $1 \leq r < n$, which is unknown and potentially different depending on the alternative hypothesis under consideration. Because improper priors are used by the authors ordinary Bayes factors are not well defined, so intrinsic Bayes factors (IBF) are used. In particular the arithmetic, geometric and median intrinsic Bayes factors (AIBF, GIBF, MIBF) are compared based on their ability to detect a change. The AIBF is determined to be superior in many cases and no worse in others, though the methods do differ in their ability to detect a change in the covariance matrix and all three are poor detectors when the change in the covariance matrix is small in magnitude. Once a changepoint is detected using one of the IBF variants, the estimator can be computed from the posterior distribution of the changepoint. However, this method seems to be extremely liberal in

that the effective type I error rates are near 20%, which would seem to indicate that any interval estimates provided as a result of this method would be much too narrow when compared to the intervals based on the nominal rate.

Ruch et al. [1993] gives one of the few applications in which changepoints are investigated using a multivariate analysis tool, but they are still interested in the multivariate-E problem. Specifically the authors use principal components analysis (PCA) to simplify the changepoint detection problem by utilizing PCA as a dimension-reducing tool. They use the four-parameter changepoint model given in Ruch and Claridge [1992], which is simply a version of the hockey-stick model, in this case allowing for different slopes and intercepts but with forced continuity. In Ruch et al. [1993] the changepoint is estimated via Hudson's algorithm [Hudson, 1966] applied to a particular variable – described as the "dominant" variable by the authors. Once the changepoint is estimated, a multiple linear regression model is fit to the principal components within each regime, as defined by the changepoint estimate on the raw data. The authors fail to adequately justify the selection of the "dominant" factor in the model and rely on the usefulness of a changepoint detected in the original data space to be meaningful in the principal component space – an assumption that seems highly suspect.

2.7 Resampling Methods for RCPA

The intractable nature of asymptotics in changepoint problems is motivation for considering the application of resampling procedures such as the bootstrap, jackknife or permutation procedures. The bootstrap method of Efron [1979], Efron and Gong [1983] is by far the most popular resampling method, though randomization tests are occasionally explored as well. Romano [1989] investigates the bootstrap and randomization tests in a nonparametric setting, establishing that the resulting distributions of a test statistic are asymptotically equivalent for the two methods, under some mild conditions. Romano gives a changepoint example and a test statistic for which either bootstrap or randomization methods could be used but, as with his other examples, no analysis is provided.

Bickel and Freedman [1981] give an in-depth look at the asymptotics of the bootstrap and provide specific conditions for when the bootstrap method admits reasonable solutions. Counterexamples are also provided, including the estimation of θ for data following a $U[0,\theta]$ distribution, showing that the bootstrap can fail even in cases where a well-accepted pivot for the quantity exists. This example is often used to serve as a warning that bootstrap methods are not to be applied blindly since the bootstrap is not a failsafe method. In particular this cautionary tale appears on page 81 of Efron's and Tibshirani's monograph devoted to the bootstrap [Efron and Tibshirani, 1993] as well as by Davison and Hinkley on page 39 of their work on the multitude of applications of the bootstrap technique [Davison and Hinkley, 1997]. Antoch and Hušková [2001] provide similar work on the limiting distribution of test statistics in a permutation framework.

Hušková and Slabý [2001] explore the application of randomization tests to the simple location model

$$X_i = \sum_{j=0}^{q} \mu_j I\{m_j < i \leq m_{j+1}\} + \varepsilon_i, \qquad i = 1, \ldots, n$$

where the m_j are the unknown changepoints, the μ_j are the unknown means and q can be known or unknown. There primary concern is with the testing of the null hypothesis of no change against an alternative of at least one change. A kernel based method is proposed and the limiting distribution of the resulting test statistic is provided, as well as the limit distribution of the permutation distribution. The result is supported via simulation studies in which six different kernels under the null and five different alternative hypotheses all for two sample sizes of $n = 100, 200$. Furthermore, they considered the error distribution in the location model to come from three distributions: standard normal, Laplace and $t(4)$.

Kim et al. [2000] apply permutation test methods to a slightly more complicated version of the changepoint problem, which they term "joinpoint regression" and which is essentially a piecewise linear regression model with no continuity constraint. Their model is

$$\mathrm{E}[y|x] = \beta_0 + \beta_1 x + \delta_1 (x - \tau_1)^+ + \cdots + \delta_k (x - \tau_k)^+$$

where the τ_j are the unknown joinpoints and the function $\delta(\omega)^+ = \begin{cases} \omega & \omega > 0 \\ 0 & \omega \leq 0 \end{cases}$ and they provide a permutation test for testing

$$H_0 : \text{there are } k_0 \text{ joinpoints}$$

versus

$$H_1 : \text{there are } k_1 \text{ joinpoints.}$$

Their test is based on a grid search over all possible sets of k_0 joinpoints from which the parameter estimates that provide the minimum are found. Based on this model the residuals are found and a Monte Carlo method is used to sample from the set of permutations to provide an approximate permutation distribution. A transformation of the F statistic for goodness-of-fit is used as the criterion for determining the significance of a particular realization from the permutation distribution. The $p-value$ for the original data is then computed based on these realizations. The authors provide a generalization to accommodate non-constant variance as well as serially correlated residuals. Simulations and an application to a particular example involve cancer rates were provided for their method.

In a similar vein, Kim et al. [2004] compare the randomization test method to the standard F-test procedure for comparing whether two segmented regression fits are coincidental or if they are parallel but with distinct intercepts and seem to show preference for the randomization procedure based on simulation results indicating that its distribution of $p-values$ is more nearly uniform under the null hypothesis than those of the F-test. Julious [2001] investigates the need for resampling when such analysis is carried out and shows, via simulation, that the standard F approximation is not valid in the case of a single changepoint with a piece-wise linear regression model as given by Hudson [1966].

Crowley [1998] gave a coarse review of the applications resampling methods such as bootstrap, permutation and jackknife methods to data in ecological and evolutionary studies. After giving a layman's interpretation of the methods he discusses the increasing trend in the frequency with which these methods are applied and gives several specific

areas that he designates to be areas of application, including but not limited to, competition and community structure, spatial patterns and processes, and environmental modeling. One underlying theme in his focus on these types of applications is their performance with "small" or "moderate" sample sizes. In particular he discusses their ability to avoid making untenable and unverifiable assumptions, e.g. claiming a normal distribution is appropriate when, in fact, it is not. Furthermore, one issue of importance is with these small sample sizes that are typically collected in such applications as opposed to the limiting distribution results which necessarily have n approaching infinity.

CHAPTER 3

MOTIVATION FOR CURRENT RESEARCH

3.1 Summary of Current Methods

The changepoint problem has clearly been well established as an area of critical importance for applied statistics, which as a result has lead to some rich theoretical developments as well, particularly in the area of asymptotics. While there is great interest in the area, and correspondingly there is a vast amount of literature available, the literature reviewed here has tended to focus on but a few facets of the changepoint problem. Whether the approach was nonparametric or parametric, frequentist or Bayesian, it has tended to follow the same pattern. Namely, assuming a fixed number of changes and providing a test statistic for locating the optimal split(s) with large-sample theory being used to justify the decision to reject or fail to reject a null hypothesis. Furthermore the models of interest tend to be either location models or piecewise linear regression models, either with or without continuity restrictions.

Unfortunately many of the methods developed have drawbacks that prevent them from being widely implemented. Many of the frequentist methods rely on large-sample theory which leaves questions as to the performance of the methods for smaller sample sizes that are more routine in environmental research settings. Bayesian methods require the specification of prior distributions be made and often make use of hyperparameters which can provide an obstacle for a non-statistician attempting to implement these methods and who subsequently then must interpret the results. Both frequentists and Bayesians have tended to focus on shifts in the mean that either assume constant variance, or allow for heteroscedasticity but do not test for such an event. In the

methods where tests for homoscedasticity are applied it is typically in conjunction with changes in the mean, more precisely it is typically assumed that if a variance change occurs, then it occurs simultaneously with a mean change. This is not necessarily the case and applying such an assumption to a data set for which it is not valid may lead to inaccurate estimates of the true changepoint in the mean, the variance, or both.

One of the challenges evident in the study of changepoint problems is that it necessitates a knowledge of extreme value theory since, regardless of the statistic used, the optimal changepoint is traditionally defined as the point at which the statistic is minimized or, more commonly, maximized. As a result, the asymptotics of these problems tend to result in similar distribution functions for many of the statistics suggested. The asymptotic distribution for the case of non-simultaneous changes in mean and variance has not been studied extensively. These distributional complications lead naturally to the question of whether or not resampling techniques can be reliably applied to such problems. It is well known that bootstrap methods are not always applicable, with one of the classic failures being specifically the estimation of extreme values, i.e. minima and maxima.

3.2 Narrowing the Field

There is a wealth of literature on changepoint detection and estimation available making the comparison of all parametric RCPA methodologies a nearly insurmountable task. However, the fundamental principles by which the methods differ do not form nearly as extensive a list. A data generation process (DGP) is assumed, a deterministic model is assumed, the nature of the changepoint is defined and finally a test statistic is proposed for purposes of detection. The choice of DGP often has an influence on the type of test statistic that is proposed while the statement of the deterministic model includes a definition of how many changepoints are allowed to exist, again affecting the test statistic. For example, choosing to look at normally distributed data under a simple mean-shift model leads to the common choices of maximum likelihood based versus least-squares based testing procedures. On the other hand if a Poisson DGP is assumed

and a change in rate is of interest, a model-deviance based approach is more likely to be selected as the analysis method.

In focusing on methods well suited to the environmental and ecological applications, this allows one to focus on the models – both stochastic and deterministic – that are often encountered and then concentrate on analysis methods applicable to such situations. Current literature in such areas would indicate that several popular models of interest are mean-shift, hockey-stick and hormesis models [Barrowman and Myers, 2000, Hagerthey et al., 2008, Horness et al., 1998, McCormick et al., 2009, Paul and McDonald, 2005]. While the differentiation between a hormesis model and its hockey-stick approximation (or vice versa) is difficult without a sufficient amount of data near the region of change, the location of the changepoint in these models is relatively similar. Conversely, the location of the changepoint estimated under a mean-shift model versus a hockey-stick model are likely dissimilar. As a result, methods that are applicable to these two deterministic model choices are of great interest, with preference being given to methods generalizable to other situations. Furthermore, it is very common for assumptions of normality or log-normality to be made in the case of continuous environmental factors like water chemistry values which are used in water quality monitoring. Thus assumptions of normality (or log-normality) for the stochastic model and one of the aforementioned deterministic models are common combinations in environmental and ecological settings.

Many of the authors mentioned in the literature review have provided methods applicable to such situations. In some cases authors such as James et al. [1987] give comparisons of their methodology to several other methods, but seemingly without any thought given to the ease of implementation or simplicity of interpretation of the test-statistic being proposed. In more extreme cases authors such as Qian et al. [2003] and Paul and McDonald [2005] use methods with higher than nominal alpha levels resulting in untrustworthy results. In light of that, Chapter 4 looks at both modeling scenarios described above and compares several of the methods that are popular *in applications* to a proposed method which is corrected to control the overall type I error rate. An example data set from the Everglades, Florida, USA is used to showcase the practical and simple

nature of the proposed corrected method, which through simulations is shown to be nearly uniformly preferable to the competing methods. Chapter 4 represents material submitted to *Ecological Modelling* [Duggins et al., submitted].

3.3 Extensions

Of course, not every situation is well modeled under the conditions described above: namely there may be non-normally distributed data, non-simultaneous changes in mean and variance, dependent responses through some autocorrelation structure (known or unknown) or even multiple responses requiring multivariate analysis. Furthermore researchers may encounter more than one deviation from the simple settings described in Sections 3.1 and 3.2 at one time. It is natural then to begin to think about how the optimal or near-optimal methods from the simple case might apply, if at all, to such situations. If these methods are not applicable, then methods that are applicable must be developed. In light of that reasoning two scenarios which deviated from the above conditions were selected, both motivated by solving a problem relevant to current environmental research areas and for which data sets were available for showcasing the methodologies.

The first of the scenarios investigated is motivated by the analysis rainfall data as it relates to modeling and predicting drought patterns. This rainfall data is clearly subject to autocorrelation and the newly popular standardized precipitation index (SPI) for use in analyzing drought patterns is subject to some correlation structure. Recently several authors have looked at the modeling of SPI using log-linear models [Moreira et al., 2006, Paulo et al., 2005] while other authors have given attention to which time-scale should be used to look at particular types of drought [McKee et al., 1993, Vicente-Serrano and López-Moreno, 2005]. It is suggested in Chapter 5 that changes in drought conditions could be detected through the transition matrix defined on the drought classes based on observations before and after a potential change. Several metrics are introduced which allow for such discrimination in several interesting cases and simulations show this to be a valid method. Rainfall data over the last 50-100 years from several sites in Kenya is

used to demonstrate the usefulness of the procedure in practice. Chapter 5 is a result of joint work with Matthew Williams and reflects material published in Duggins et al. [2010].

The second scenario of interest corresponds to a situation in which several environmental health indicators are measured as responses on a site and which together may represent the ecosystem health at the site. If if the responses are all highly dependent so that there is a great deal of shared information between the response, principal component analysis (PCA) could be used to yield a single component and univariate procedures would most likely be sufficient for detection. Similarly the responses were all independent (or nearly 'so) then univariate procedures as described in Section 4 could be used and there is likely no benefit to using a multivariate procedure such as PCA. However, in the intermediate case where PCA would reveal two or three principal ecosystem health components, there is the open question of how to detect various types of changes such as changes in the constituents of the components, changes in the mean level of a component, changes in the variance, etc.

3.4 Summary

In summary, I begin with the issue of developing a testing procedure for the mean-shift model based on bootstrapping since that provides a way to simultaneously test for the significance of a changepoint whose location was not given exogenously and provide interval estimates as to its location. This is done to avoid two common issues in changepoint analysis: the need for corrections due to multiple comparisons and the difficulty in deriving asymptotic results. This method is compared under the basic two-sample, mean shift case and the hockey-stick model and shown to be competitive, if not superior, to the other methods here. The case of time-series data that has been categorized and used to construct a transition matrix as the observation is considered next. Finally the question of estimation and testing in the multivariate case is examined. An asymptotic test is available for multivariate data, but it is based on the covariance matrix. As such, when n is only marginally larger than p it is supposed that the test has

low power. In light of this, several alternative approaches using principal component analysis are proposed as competitors for such a situation.

CHAPTER 4

EVALUATING THRESHOLDS IN ECOLOGICAL AND ENVIRONMENTAL SETTINGS

4.1 Introduction

Establishing thresholds for ecological and environmental systems is important for setting environmental standards and for finding levels of stressors that result in critical changes in systems [King and Richardson, 2003, Weigel and Robertson, 2007]. One method for finding thresholds is based on the idea of a changepoint. The typical changepoint problem is composed of two stages: determining if a changepoint exists and, if so, estimating the changepoint using both point and interval estimators. Critical aspects of the first stage include determination of underlying model (including the number of changepoints), the algorithm for finding the threshold(s) and testing the significance of the threshold(s). Popular models include the abrupt shift model in King and Richardson [2003], Qian et al. [2003] and the hockey-stick model found in Barrowman and Myers [2000], Horness et al. [1998]. King and Richardson, for example, use changepoint methods to determine phosphorus levels that are associated with critical vegetation change in the Everglades [King and Richardson, 2003]. They used the drop-in-deviance method due to Qian et al. [2003] to find the phosphorus level that results in the greatest change in deviance assuming a Poisson distribution of counts. Testing for the significance of the threshold, relative to the null model of no threshold, is carried out via a test based on the asymptotic distribution of the scaled deviance.

In Weigel and Robertson [2007] the interest is in obtaining a criterion for total phos-

phorus and total nitrogen for the state of Wisconsin. The data consist of samples from 41 non-wadeable rivers and contain information on the stressors and biotic variables. A regression tree was used identify the threshold or breakpoint levels of nutrients for several biological metrics. Essentially what is done in this approach is to split the data into two groups of biological metrics according to values of the total phosphorus or total nitrogen. The breakpoint is then the value of the stressor that results in the most separation in the means of the metric relative to the variation. To test for significance of the split, a two sample (unequal variance) t test is used. Other examples of similar approaches for estimating and testing for thresholds include Wang et al. [2007] and Paul and McDonald [2005].

The test for a changepoint relies on a model of the data (deterministic and stochastic). There are two common deterministic models for studying changepoints. First is the mean-shift model in which the mean is assumed constant until the changepoint occurs. After the changepoint, the mean is different but also constant; thus the model is essentially an intervention model. The system is initially stable and a stressor is introduced which moves the system to a new state. In the second model, a regression is assumed. This model corresponds to a situation where the stressor causes a gradual change. After a certain point, there is a collapse or change in state. Thus, after the changepoint, there is still a regression model but the slope and intercept change. This type of model is often associated with situations in which the biological response declines then flattens. This is the over-harvesting or hockey-stick model that is common with populations that fail [Barrowman and Myers, 2000, Horness et al., 1998].

The stochastic component of the model depends on the type of data that is collected. With biological count data models such as the Poisson or negative binomial model might be used. For chemistry data, it is common to use either a normal model or lognormal model. Different distributions lead to different types of tests that might be entertained for evaluating if a shift actually occurs. For example, with normal data and a simple shift in means, the two-sample t test is a possible choice. With Poisson data, deviance based tests are more appropriate.

There are, of course, many other possible deterministic and stochastic models that

may be entertained. Often, subject matter knowledge will suggest a form for the model, thus giving some indication as to the number of changepoints, but not necessarily their location. Once the model has been selected, changepoints are selected using an optimization criterion. For example, with continuous data, one might consider sum of squares error and select the point that divides the data into two groups, with the groups selected to minimize the within sum of squares. With other types of data (e.g. Poisson), the point that yields the greatest difference in deviance might be used. Once the changepoint is selected, a test is used to determine significance of the changepoint. Unfortunately, many applications of changepoint models do not account for the fact that the location of the changepoint, and hence the groups of observations, are estimated from the observed data. Ignoring this may lead to incorrect statements about type I error rates, resulting in potentially misleading conclusions.

Specifically in this chapter we show that even under the simple mean-shift model the deviance and t test approaches have significant drawbacks that should disqualify them as sound testing techniques. A primary concern with both of these methods is that the effective type I error rate is much larger than the nominal significance level. In addition to these two methods, several other methods are evaluated. Namely a likelihood based test proposed by Horváth [1993] and a bootstrap test version of the t test proposed here. In addition the F test and a bootstrap corrected F test are investigated for the trend-change model.

The focus of the chapter is on frequentist methods rather than Bayesian methods. It is recognized that considerable literature exists on the use of both approaches for evaluating changepoints. Frequentist methods were selected for evaluation because they are the primary ones used for testing for a changepoint in the literature. Bayesian methods are sometimes used for estimating credible sets for the changepoint; however, even in this case, a frequentist test was used to determine significance of the changepoint (e.g. Qian et al. [2003]). Other Bayesian methods for changepoint detection exist, in particular Fan and Chen [2005] and Green [1995] have found success in modeling Poisson processes using Markov chain Monte Carlo (MCMC) methods – with the former focusing on ratios of Bayesian Information Criteria (BIC) and the latter on the posterior density

estimates of the changepoints. However, as with the majority of the Bayesian literature the focus is on estimation of a changepoint, assuming one (or more) exists, rather than testing for the existence of a changepoint.

4.2 Methods

Methods for determining a changepoint differ based on the assumptions made regarding the model, data, criteria and test [Hagerthey et al., 2008, Horváth, 1993, Julious, 2001, Paul and McDonald, 2005, Qian et al., 2003]. Here, we consider several methods based on the assumptions that the data of interest are continuous and normally distributed. While this is sometimes not the case for environmental data, the t test and deviance methods are applicable to this situation and should have good properties. In all methods the data are first ordered according to some additional variable which often corresponds to a spatial, temporal or stressor gradient. In total six methods are investigated for testing the significance of an estimated changepoint: t test, drop-in-deviance chi-squared test, likelihood ratio test, and bootstrapped t test for the mean-shift model and the F test and its bootstrap counterpart for the regression model. We first describe the basic deterministic models and hypotheses.

4.2.1 Models and Hypotheses

While there are many possible choices for the deterministic component, two models were selected for consideration based on their common usage: a mean-shift model and a piecewise-linear regression model (the hockey-stick). Under the mean-shift model discussed above the data consist of n pairs of observations, (x_i, y_i), that are ordered according to the stressor, x. At each level of the stressor it is assumed that the data follows a normal distribution with constant variance σ^2, but after some index value, say τ, there is a change in the mean such that the mean is μ_1 for $x \leq x_\tau$ and μ_2 for $x > x_\tau$. This gives rise to Model (4.1) below and leads to the hypothesis test of interest which is for a shift in the mean.

$$y_i = \begin{cases} \mu_1 + \varepsilon_i & x \leq x_\tau \quad i = 1,2,\ldots,n_1 \\ \mu_2 + \varepsilon_i & x > x_\tau \quad i = n_1+1,\ldots,n \end{cases} \tag{4.1}$$

The hypotheses tested here are thus

$$H_0 : \mu_1 = \mu_2$$

versus

$$H_1 : \mu_1 \neq \mu_2.$$

It is also feasible to assume a shift in the variance, although this if often treated as a secondary effect and built into the test statistic, if it is considered at all.

In the regression model mentioned earlier the data again consist of n pairs of observations which are again ordered according to a stressor. However, the stressor now plays the role of explanatory variable in a linear model. As such, it is still assumed that the data is normally distributed with constant variance around the model. This scenario leads to the following model.

$$y_i = \begin{cases} \beta_{01} + \beta_{11} x_i + \varepsilon_i & x \leq x_\tau \quad i = 1,2,\ldots,n_1 \\ \beta_{02} + \beta_{12} x_i + \varepsilon_i & x > x_\tau \quad i = n_1+1,\ldots,n \end{cases} \tag{4.2}$$

Clearly, the model with stability followed by period of decline/incline after the change-point would be analyzed in a manner analogous to the model with a decline/incline followed by stability given in Model (4.2) above. For a model of this form, the relevant hypotheses to test depends on the form of the null model. In application, the model to which Model (4.2) is compared is often a simple linear means model and hence one would test

$$H_0 : \mu_i = \mu$$

and

$$H_1 : \mu_i = \begin{cases} \beta_{01} + \beta_{11} x_i & \text{if } x_i \leq x_\tau \quad i = 1,2,\ldots,n_1 \\ \beta_{02} + \beta_{12} x_i & \text{if } x_i > x_\tau \quad i = n_1+1,\ldots,n \end{cases}.$$

4.2.2 Test Criteria for the Mean-Shift Model

t test

The first method to be investigated is based on detecting the observation that divides the sample of n observations into two groups for which a two-sided, two-sample t test has the lowest significant p-value for testing the hypotheses mentioned in conjunction with Model (4.1). In order to estimate for which observation this condition is met, the algorithm starts by creating two groups using observations one through three (or some other minimum group size) as the first group and the remaining $n-3$ observations as the second group. The aforementioned test is then conducted on these two groups. The algorithm then proceeds sequentially creating groups of size k and $n-k$ until finally reaching the point at which $k = n-3$ so the first group is of size $n-3$ and the second group is of size three.

As mentioned above, it is possible that the variances before and after the changepoint are unequal, though that is not of primary interest. As such, equal variances are not assumed and the t distribution that best approximates the distribution of the computed test statistic is found via Satterthwaite's formula [Satterthwaite, 1946]. The resulting approximate t distribution is used to obtain the test statistic's associated p-value and the changepoint is estimated to be the value of stressor that results in the smallest significant p-value, i.e. x_k. If the data can not be split such that a significant p-value is found for any value of k, then we fail to reject the null hypothesis, indicating a lack of evidence supporting the presence of a changepoint.

Bootstrapped t test

In direct comparison with the t test method above, a bootstrap alternative is proposed that is a modification of said method. As such, it tests the same hypotheses under the same assumptions. The two-sided, two-sample t test is performed on the observed data giving a test statistic, say t_0, to be used as a reference value for the b bootstrap samples that are eventually created. Bootstrap samples are generated, assuming the null hypothesis of no changepoint to be true, in the usual fashion by sampling with re-

placement from the original data. Specifically, SAS 9.2 was used to select a sequence of n random numbers from a uniform(0,1) distribution via the **rand("uniform")** command in **proc iml** which uses the 32-bit version of the Mersenne twister algorithm of Matsumoto and Nishimura [1998]. These uniform variates are then translated to integer indices by computing $\lfloor n \cdot u_i \rfloor + 1$ where u_i represents the i^{th} uniform random variate and $\lfloor \cdot \rfloor$ is the floor function. These indices are used to select the observation from the original vector of observations that are now associated with the i^{th} position in the vector of bootstrapped observations.

From each of the b such samples the same algorithm described above is used to find the observation that leads to a maximal splitting of the data based on the p-value of the test statistic. This statistic, t_B, is compared to t_0 in order to determine whether or not the sample has resulted in a significant changepoint as compared to the original data. At this point the procedure restarts by generating a new bootstrap sample and repeating the significance algorithm, eventually generating a set $\{t_B\}_j$ for $j = 1, \ldots, b$, each of which is compared to, t_0, the original value. At the conclusion of the bootstrap-based test an empirical p-value, $\tilde{p}_t$, is defined to be the proportion of t_B that exceeded t_0 in magnitude. Note that the estimate of the changepoint is unchanged based on whether the t distribution or bootstrapping is used for determining significance; it is the significance itself that is affected.

Drop-in-Deviance Test

The third method uses a deviance reduction approach to detect the observation that leads to the largest reduction in the overall model deviance [Qian et al., 2003, Ramsey and Schafer, 2002]. The hypotheses being tested are the same and the algorithm again sequentially divides the data in exactly the same manner as the t test approach. The total deviance, assuming normally distributed data, is calculated as

$$D = \mathbf{y}^T \left(\mathbf{I} - \frac{1}{n} \mathbf{J} \right) \mathbf{y} \tag{4.3}$$

and the deviance for the i^{th} group ($i = 1, 2$) is

$$D_{ik} = \mathbf{y}_i^T \left(\mathbf{I}_i - \frac{1}{n} \mathbf{J}_i \right) \mathbf{y}_i \tag{4.4}$$

where the $\mathbf{y}_i$ are vectors of response data with $\mathbf{y}_1$ having size k and $\mathbf{y}_2$ having size $n-k$, $\mathbf{J}$ is an $n \times n$ matrix of ones and $\mathbf{I}$ is the $n \times n$ identity matrix. The deviance reduction is then computed as

$$\Delta_k = D - [D_{1k} + D_{2k}] \tag{4.5}$$

and the changepoint is estimated to be the data point that maximizes Δ_k. Note that this method is then guaranteed to find a point where the reduction is maximized, unlike the method based on the t test, which does not necessarily discover a changepoint. To determine whether the detected changepoint is significant, the fact that $\frac{\Delta_k}{s} \dot{\sim} \chi^2(1)$ under the null hypothesis of no changepoint is used to provide an approximate chi-square test of significance [Venables and Ripley, 1997]. Here s is an estimate of the scale parameter, which under the assumption of normality is the sample standard deviation computed based on the entire set of n observations. Due to this, the deviance method is only valid under the assumption of equal variance, and so even though it tests the same hypotheses as the t test method, it does not necessarily do so under the same assumptions.

Likelihood Ratio Test

The fourth method to be compared is a likelihood ratio technique that again starts with the third observation and proceeds to the antepenultimate observation looking for the point at which the likelihood ratio statistic is maximized. The statistic, due to Horváth [1993] is

$$\lambda_n = \left(\max_{1<k<n-1} \left(n \log \widehat{\sigma}_n^2 - k \log \widehat{\sigma}_k^2 - (n-k) \log \widehat{\sigma}_{n-k}^2 \right) \right)^{1/2} \tag{4.6}$$

which can be seen to be closely related to Bartlett's test for homogeneity of variance [Bartlett, 1937]. Here $\widehat{\sigma}_n^2$ is the sample variance based on the full data set, while $\widehat{\sigma}_k^2$ and $\widehat{\sigma}_{n-k}^2$ are the sample variances based on the first k and last $n-k$ observations, respectively, where the maximum likelihood estimators are used in each case [Horváth, 1993]. Unlike the t test and deviance methods which are designed only to search for a mean shift, the hypotheses being tested with this statistic are

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_n \quad \text{and} \quad \sigma_1 = \sigma_2 = \cdots = \sigma_n$$

$$H_1 : \mu_1 = \mu_2 = \cdots = \mu_k \neq \mu_{k+1} = \cdots = \mu_n$$

or

$$\sigma_1 = \sigma_2 = \cdots = \sigma_k \neq \sigma_{k+1} = \cdots = \sigma_n$$

which show that the likelihood ratio method is designed for the detection of a mean shift, variance shift, or both – assuming that if both occur, they occur at the same point. Once the optimal maximum likelihood ratio statistic is found, the p-value is computed using the asymptotic distribution of a function of λ_n. In particular

$$\sqrt{2\log\log n}\,\lambda_n - (2\log\log n + \log\log\log n) \xrightarrow{\mathscr{D}} W$$

where W follows a type-II Gumbel distribution with distribution function $\exp(-2e^{-x})$ [Horváth, 1993]. Once the asymptotic p-value is computed it is used to determine whether or not the x_k at which the shift is estimated to have occurred is significant.

4.2.3 Test Criteria for the Hockey-Stick Model

F test

A popular approach in threshold detection is to fit a "hockey-stick" style model resulting from allowing one of the two segments to be a simple linear regression model while the second segment is a simple means model [Barrowman and Myers, 2000, Horness et al., 1998]. Hudson [1966] provided an iterative scheme for determining an optimal piecewise linear regression model based on a change in residual sum-of-squares (RSS) between the "null" model - a single linear regression - and the "alternative" model - two joined simple linear regressions. For the purposes of this investigation the piecewise linear model was assumed to be of the "hockey-stick" formulation and without loss of generality the first regime was chosen to be the segment with nonzero slope. Rather than force continuity onto the estimated regressions, the piece-wise linear models in these cases were not constrained to join at the changepoint which is a slight deviation from the approach of Hudson [1966], Julious [2001].

The test statistic given by Julious [2001] for the case of two piece-wise linear segments is attributed to Worsley [1983] and is

$$F_{obs} = \frac{(RSS_0 - RSS_1)/2}{RSS_1/(n-4)}$$

where RSS_0 represents the residual sum-of-squares from the fit of a simple linear regression to the full data set. RSS_1 represents the sum of the residual sum-of-squares for the single changepoint model, i.e. RSS_1 is constructed by fitting all possible two-segment models with minimum group size of two, determining the RSS within each segment and summing them together. The optimal two-segment model is the one for which RSS_1 is minimized.

Bootstrapped F test

Julious [2001] demonstrated that the naïve F test based on the change in RSS is not well modeled by an F distribution when the location of the changepoint is not known *a priori* and proposes bootstrapping to determine the actual significance of any resulting changepoint estimate. To obtain corrected significance levels for the observed F statistics, bootstrap samples were used, specifically fixed bootstrap samples were constructed in a manner similar to the bootstrapping done for the bootstrapped t test. In this case the bootstrap methodology described in Section 4.2.2 is carried out: from the observed data b bootstrap samples of the responses are taken, now with the resulting vector of responses matched with the original stressor values, creating b bootstrap data sets. From each of these bootstrap samples the F statistic is constructed and compared to the F_{obs} from the original data and an empirical p-value, $\tilde{p}_F$, is defined to be the proportion of the bootstrap F statistics that are at least as large as F_{obs}. As with the bootstrapped t test, the location of the changepoint estimate is unaffected by the decision to use bootstrapping, only the significance level is affected.

As mentioned previously the algorithm for detecting the changepoint is slightly different than the one used by Julious [2001] in not requiring the resulting model to be continuous at the changepoint. As a result, and in the interest of extending Julious' work, it was decided to expand the analysis to include variations in sample size and additional

values of the standardized slope. The location of the changepoint along the gradient was still of great importance so this parameter was included here, just as in Julious [2001]. To keep the comparability of the test statistic to Julious' work the means model was estimated using a simple linear regression.

4.2.4 Simulation Design

To achieve the primary goal of establishing the effective type I and type II error rates simulations were carried out on n pairs of observations with responses $(y_1, \ldots, y_n)$ randomly generated from a $N(1,1)$ distribution using SAS 9.2. The regressor x was set to be a random uniform number on the interval (0,1) so that the results to be compared could be viewed on the same x-scale. Since the methods used did not depend on whether the shifts were upward or downward, the simulations are, without loss of generality, designed such that the observations in the first group have a constant mean and variance of one. Any designed changes in the parameters are then applied to the data falling in the second regime.

If no changepoint were truly present simulations were proposed to study the empirical type I error rate. When a changepoint is present in the Model (4.1) each possible situation was of interest: no changepoint in the data, a changepoint in the mean only, a changepoint in the variance only and a changepoint in both the mean and variance. For the hockey-stick model the only investigated change was from a line with positive slope to a line with zero slope. To investigate the power of the aforementioned changepoint detection methods under these scenarios, the following simulations were used.

No Changepoint

In simulations used to investigate type I error rates there was no changepoint introduced and the only factor manipulated for each model was the sample size. Five sample sizes were selected ($n = 20,\ 40,\ 60,\ 80,\ 100$), though cases with small sample sizes are of more interest in an ecological setting. These sample sizes were used under both Model (4.1) and Model (4.2).

Model (4.1) - Mean Change Only

The second set was designed to diagnose the effects of a pure mean shift, thus the variances were held constant with $\sigma_1 = \sigma_2 = 1$. A changepoint was then introduced in the mean at the r^{th} observation for some specific choices of $1 < r < n$. Critical to this assessment were the size of the mean shift, overall sample size and the proportion of observations falling on either side of the changepoint. The sample sizes stated in Section 4.2.4 are again used and the changepoint was inserted at one of the indices $r = 0.10n$, $r = 0.25n$ or $r = 0.50n$; this ratio is referred to as the *split ratio* of the data. For each of these fifteen combinations a mean shift ($\delta = 0.5,\ 1,\ 2, 3$) was used, the values for which were chosen to correspond to shifts of $\frac{1}{2}\sigma_1$, σ_1, $2\sigma_1$ and $3\sigma_1$.

Model (4.1) - Variance Change Only

Similarly a third group of simulations were carried out in which a changepoint was introduced in the variance and the mean was held constant with $\mu_1 = \mu_2 = 1$. The same sample sizes ($n = 20,\ 40,\ 60,\ 80,\ 100$) and split ratios ($0.10n,\ 0.25n,\ 0.50n$) as before were again used. In these simulations the standard deviation of the second group was increased by factors of two ($\sigma_2 = 2$) and three ($\sigma_2 = 3$).

Model (4.1) - Mean and Variance Change

The fourth round of simulations was used to investigate the power when a shift occurred simultaneously in the mean and variance. For these simulations the settings from the previous two scenarios were again used. Thus the sample sizes were still 20, 40, 60, 80 and 100, the split ratios were again $0.10n$, $0.25n$ and $0.50n$, the mean shifts were kept as 0.5, 1, 2 and 3 and the standard deviations were still inflated by factors of two and three. The motivation for considering situations where a change occurs in both parameters comes from a common situation in which the occurrence of a shift in the mean induces a change in the variance as well. This occurs frequently when an ecological stressor increases (or decreases) to a level that forces the response of interest to approach its upper or lower limit. Having reached such a limit the response tends to

exhibit a marked reduction in variance that is a direct result of the mean shift.

Model (4.2) - Slope Change

Simulations used to examine the empirical power of the F based methods for Model (4.2) used the same sample sizes as in the previous studies and the variance was held constant across the two regimes. Julious investigates the change through the difference parameter $d = \frac{\beta_1 - \beta_2}{\sigma}$ where β_1 is the slope in the first regime and β_2 is the slope in the second regime [Julious, 2001]. In the notation presented here this gives $d = \frac{\beta_{11} - \beta_{12}}{\sigma}$ which, with a common standard deviation of $\sigma = 1$ and a second regime with slope zero ($\beta_{12} \equiv 0$), this gives $d = \beta_{11}$. As such, the slopes of the first regime were chosen to range from 2 to 10 in magnitude, with steps of size 2. The changepoint locations were chosen to be $0.25n$, $0.50n$, $0.75n$ to investigate the effects on empirical power when the second regime was both a minority and majority of the data, which was expected to have an effect. The minimum group size was set to be three to reduce the chance of having singular matrices when solving the normal equations.

Further Details

The resulting simulation designs are shown below in Tables 4.1 through 4.5. Each table lists the simulation parameter values used in creating the full factorial design.

Sample Size	20	40	60	80	100

Table 4.1: Layout when there is no changepoint present.

Sample Size	20	40	60	80	100
Split Ratio	0.10	0.25	0.50		
Mean Shift	0.50	1	2	3	
Variance	1				

Table 4.2: Simulation parameter values for a changepoint in only the mean of Model (4.1).

Sample Size	20	40	60	80	100
Split Ratio	0.10	0.25	0.50		
Mean Shift	0				
Variance	4	9			

Table 4.3: Simulation parameter values for a changepoint in only the variance of Model (4.1).

Sample Size	20	40	60	80	100
Split Ratio	0.10	0.25	0.50		
Mean Shift	0.50	1	2	3	
Variance	4	9			

Table 4.4: Simulation parameter values for a changepoint in both the mean and variance of Model (4.1).

Sample Size	20	40	60	80	100
Split Ratio	0.10	0.25	0.50	0.75	0.90
Regime 1 Slope	-2.0	-4.0	-6.0	-8.0	-10.0

Table 4.5: Simulation parameter values for a changepoint in the slope of Model (4.2).

The type I and II error rates are assessed based on the proportion of simulations that lead to a rejection for a given α. The proportion that lead to a rejection after setting $\alpha^* = \frac{\alpha}{n-5}$ based on a Bonferroni correction is also considered for the t test and deviance methods. The correction factor is based on the fact that the algorithm starts at the third observation and ends at the third to last, hence there are $n-5$ tests performed. The necessity for using some correction factor to account for scanning across multiple hypothesis tests is discussed in Davies [1977, 1987, 2002] for which approximations to the significance level were derived under certain conditions. Each of the methods discussed above were each evaluated using $m = 10\,000$ simulations for each scenario and the bootstrap method uses $b = 1\,000$ resamples within each simulation. The choice of $m = 10\,000$ simulations was made in part due to the fact that when using an $\alpha = 0.05$ significance level we have that $\frac{0.05}{\sqrt{10\,000}} = 0.0005$ is approximately equal to the simulation error. Also, note that $\sqrt{\frac{(0.05)(0.95)}{10\,000}} = 0.0022$, indicating that the simulation standard error implies a fairly precise estimate.

Bootstrap sampling can be carried out in two different ways: under the null (assuming no changepoint) or under the alternative (assuming a single changepoint). For the purposes of the investigations here, which are all based on hypothesis testing approaches, bootstrapping was done assuming the null hypothesis were indeed true, thus ignoring whether an observation falls before or after an existing changepoint. Thus the sampling is from the null distribution, i.e. assuming no changepoint is present. The specifics of the bootstrapping procedure for hypothesis testing is described in Section 4.2.2 and Section 4.2.3 for Models (4.1) and (4.2), respectively. If bootstrap methods were to be used to create interval estimates then the sampling could be carried out within each group, i.e. assuming the alternative to be true [Julious, 2001].

4.3 Simulation Results

4.3.1 No Changepoint

Model (4.1)

Consider first the case of Model (4.1) where no changepoint exists in the data, for which Figure 4.1 below shows the observed type I error rate when $\alpha = 0.05$ was used as the cutoff for each of the four methods. Clearly both the t test ($\circ$) and deviance ($\triangle$) methods result in inflated type I error rates, though the deviance reduction method appears to be slightly better than the t test method in this case. In contrast, both the bootstrap t ($\square$) and the likelihood ratio ($\diamond$) methods are slightly conservative, but with empirical rates that are very close to the nominal rate, lying only slightly below the reference line at 0.05. While the two methods were close, the likelihood method's empirical type I error rates ranged from 0.0103 (n=20) to 0.0140 (n=80), somewhat below the nominal $\alpha = 0.05$ level, while the bootstrap t method's values ranged from 0.0382 (n=20) to 0.0466 (n=80).

A Bonferroni correction was applied for the t test and deviance methods to determine if this simple adjustment would allow these methods to be competitive. The correction, as described in Section 4.2.4 above, is based on the $n-5$ comparisons that are made and does in fact lead to a significant reduction in the observed type I error rate

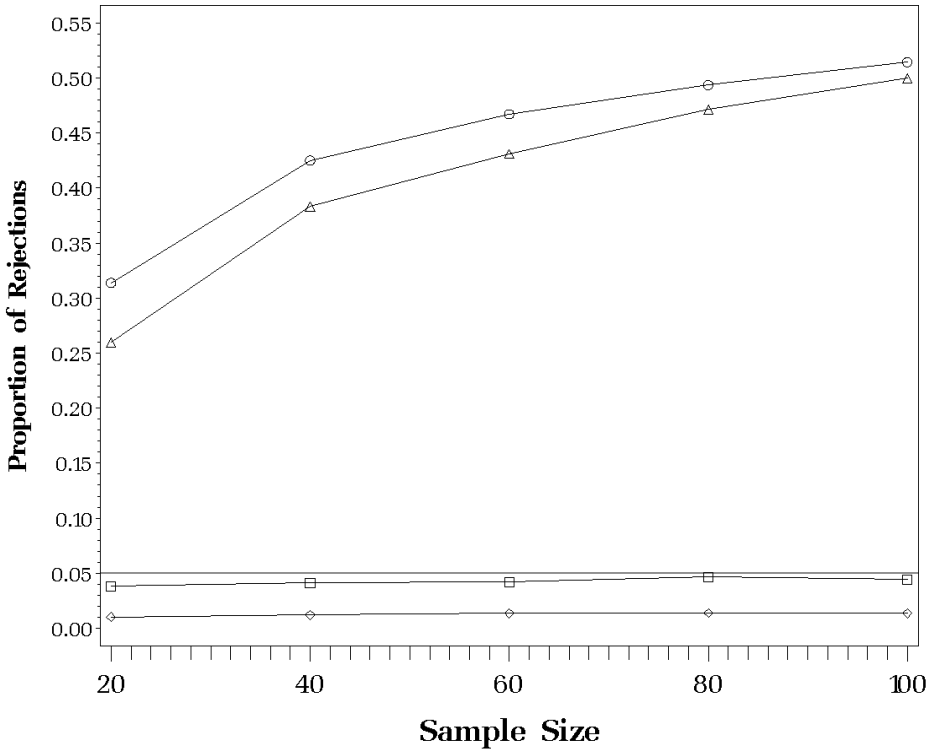

Figure 4.1: Empirical type I error rates for each of the four methods: t test ($\circ$), deviance ($\triangle$), bootstrap t ($\square$), and likelihood ratio ($\diamond$) when the common variance is one ($\sigma^2 = 1$).

but does not provide a satisfactory solution. The Bonferroni adjusted results for the deviance method were over-corrected and became conservative, with error rates similar to those of the likelihood method. The t test method when adjusted similarly, was found to perform at near nominal levels under the algorithm discussed above. However, it was found to be highly sensitive to the minimum allowed group size. If the groups were allowed to have a minimum of two observations, rather than the minimum size of three used here, the adjustment resulted in empirical type I error rates between 0.1125 (n=100) and 0.1315 (n=40) while setting the minimum group size at four resulted in a more conservative range of 0.0270 (n=100) to 0.0420 (n=20).

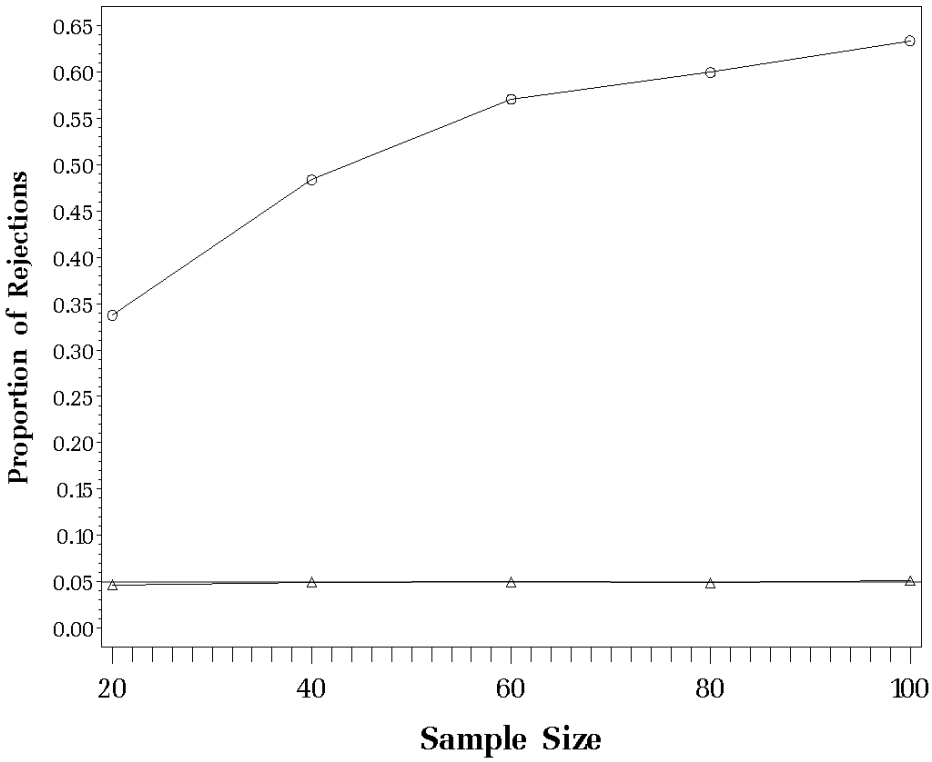

Figure 4.2: Empirical type I error rates for the F test ($\circ$) and bootstrap F ($\triangle$) when the common variance is one ($\sigma^2 = 1$).

Model (4.2)

Similar to the pattern in Figure 4.1, Figure 4.2 indicates that when $\alpha = 0.05$ is used as the nominal type I error rate the uncorrected F test ($\circ$) is inferior as compared to its bootstrap corrected counterpart ($\triangle$). As detailed above, the slope under the null case was fixed at one, as was the variance.

4.3.2 A Mean Change in Model (4.1)

The second case under consideration was the scenario in which a changepoint existed, but only in the mean. The proportion of simulations resulting in a rejection can still be used to compare the methods as this is now an empirical measure of the power of the methods. However, when the type I error rates are not equal, then any comparisons

of the method's power must be done after adjusting for this inequality. To maintain simplicity, we focus only on comparing the likelihood and bootstrap t methods as they have relatively similar results in terms of type I error rates.

The optimal case would obviously be to have the shift occurring at the $r = 0.50n^{th}$ position, leading to the set of observations that are symmetric around the changepoint. Figure 4.3 shows the results from the scenario in which $\delta = 2$ and the split ratio is 0.50. At the smallest sample size investigated (n=20) the bootstrap t method is superior by a wide margin, approximately 0.20 better than the likelihood test. For the remaining sample sizes the likelihood method obtained slightly higher empirical power than the bootstrap t method in this situation, but with both methods within 3% by $n = 40$ (0.9443 for the bootstrap t and 0.9714 for the likelihood). As expected, increasing δ leads to increased power while moving the split ratio away from 0.50 leads to decreased power for either method.

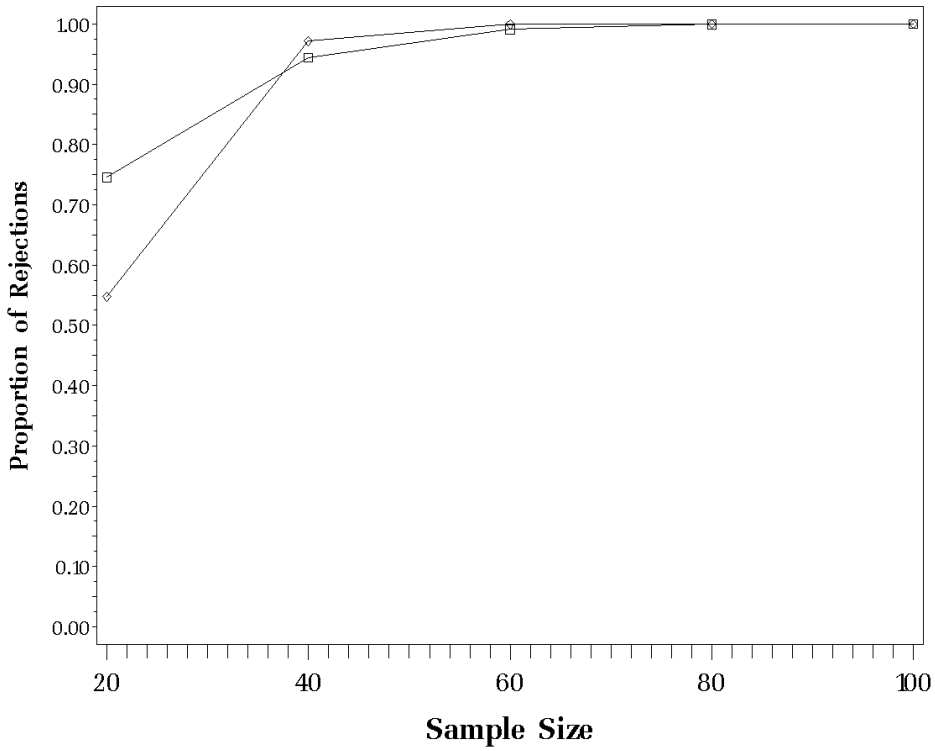

Figure 4.3: Observed power for the bootstrap t (□) and likelihood (◇) methods when the changepoint is at the center ($r = 0.50n$), there is a two-unit shift in the mean ($\delta = 2$) and the common variance is one ($\sigma^2 = 1$).

4.3.3 A Variance Change in Model (4.1)

The implemented bootstrap t method only tests for shifts in the mean, while the likelihood method tests for a shift in either the mean or the variance. As such, the likelihood method is far superior to the bootstrap t method for almost all scenarios investigated since the change occurred in the variance only. For the optimal case of $r = 0.50n$, the likelihood method was vastly superior except for $n = 20$ and $\sigma_2 = 2$, when it only exceeded the bootstrap t method by about 1.6%. Similar results hold for $r = 0.25n$ when $\sigma_2 = 3$, but when $\sigma_2 = 2$ the bootstrap t method outperformed the likelihood method for $n = 20$. In the case of $r = 0.10n$ the bootstrap t method is superior for all but the largest sample size ($n = 100$) when $\sigma_2 = 2$ and the two largest sizes ($n = 80, 100$) when $\sigma_2 = 3$. Similar to the results for a mean shift only, in general the power decreases as the true changepoint is moved away from the center of the data and increases as the magnitude of the shift in the variance increases.

4.3.4 A Mean and Variance Change in Model (4.1)

The final case investigated for Model (4.1) is the power of the likelihood and bootstrap t methods when the shift occurs simultaneously in the mean and variance. As with the previous cases, the preferred situation is for the split ratio to be 0.50 so this case is considered here. Figure 4.4 shows the results from the scenario in which $\delta = 2$, the split ratio is 0.50 and $\sigma_2 = 2$. In this case the likelihood method is consistently more powerful than the bootstrap t method, though never by more than 10% (n=40). For both methods the power is very good for a two-unit shift in the mean ($\delta = 2$).

The remaining scenarios had similar results, with decreasing δ or moving r away from 0.50 tending to lead to lower power regardless of the method. The increase of σ_2 from $\sigma_2 = 2$ to $\sigma_2 = 3$ generally both increased the power of the likelihood method and decreased the power of the bootstrap method.

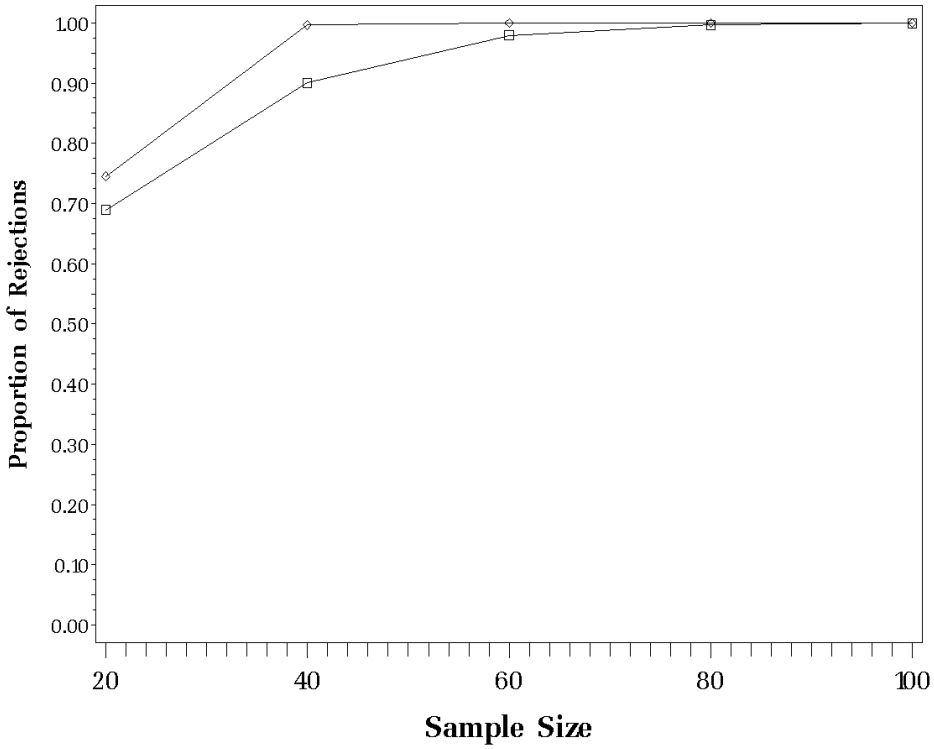

Figure 4.4: Observed power when the changepoint is in the center ($r = 0.50n$), there is a two-unit shift in the mean ($\delta = 2$), and the standard deviation of the second group is two ($\sigma_2 = 2$) for the likelihood ($\diamond$) and bootstrap t ($\square$) methods.

4.3.5 Change in Slope for Model (4.2)

When the bootstrap corrected F test is used to determine a change from a decreasing (or increasing) trend in the first regime to a constant trend in the second regime the empirical power depended on the size of the standardized slope in the first regime and the location of the changepoint, as one would expect. As Figure 4.5 illustrates, the empirical power of this test is roughly symmetric with respect to the location of the changepoint, with the best performance occurring (within the mesh used here) at $\tau = 0.50n$.

While this trend continues for all standardized slopes investigated, Figure 4.6 shows the results at the largest standardized slope under investigation, $\beta_{11} = -10$. In this extreme case the empirical powers are more acceptable, with the $\tau = 0.50n$ case exceeding 90% at $n = 40$ and the remaining cases showing reasonable power for $n \geq 60$.

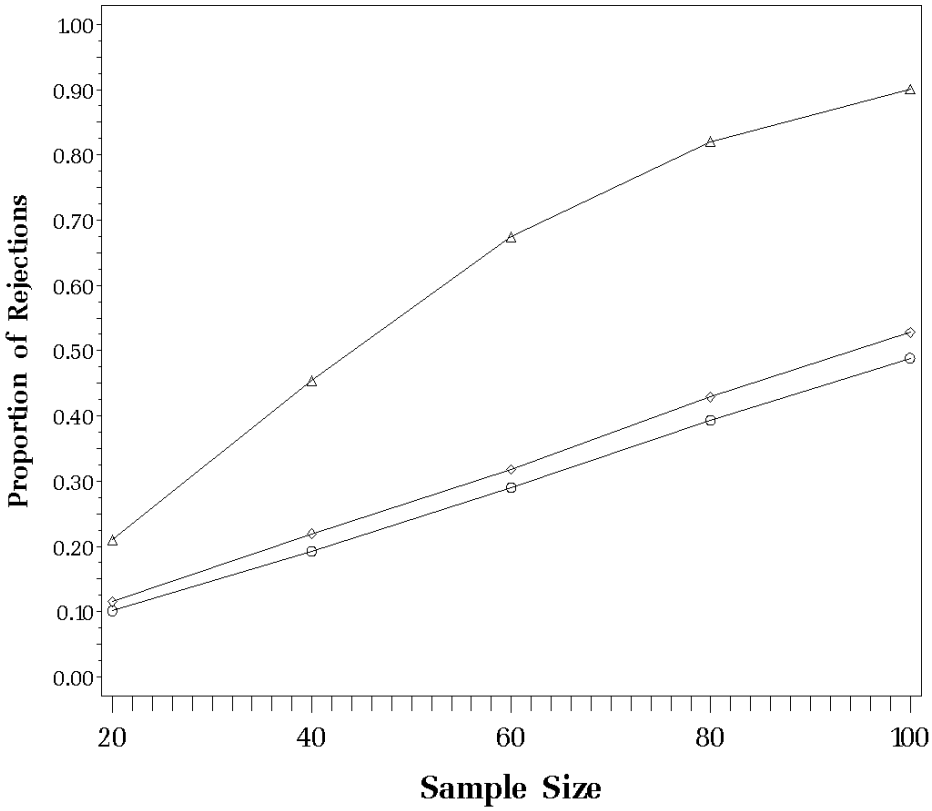

Figure 4.5: Observed power when the standardized difference in regime slopes is -6.0 and the common standard deviation is one ($\sigma = 1$) for changepoint locations of $0.25n$ ($\circ$), $0.50n$ ($\triangle$), $0.75n$ ($\diamond$).

4.4 An Example

The Everglades has experienced adverse effects due to anthropogenic influences that have led to higher levels of phosphorus which can trigger a regime shift [Hagerthey et al., 2008]. Data was collected from surficial soils along a well documented phosphorus enrichment gradient in Water Conservation Area-2A (WCA2A) in the northern Everglades on several environmental indicators including a suite of soil nutrients. The loss of the calcareous periphyton mat is one of the first changes that occurs in response to increased phosphorus loads [Hagerthey et al., 2008, McCormick et al., 2009]. We examined the relationship between soil percent ash and soil total phosphorus (TP), using ash as an indicator of the switch from a calcareous to non-calcareous system. The likelihood method and both bootstrap methods as detailed above were applied in an effort to de-

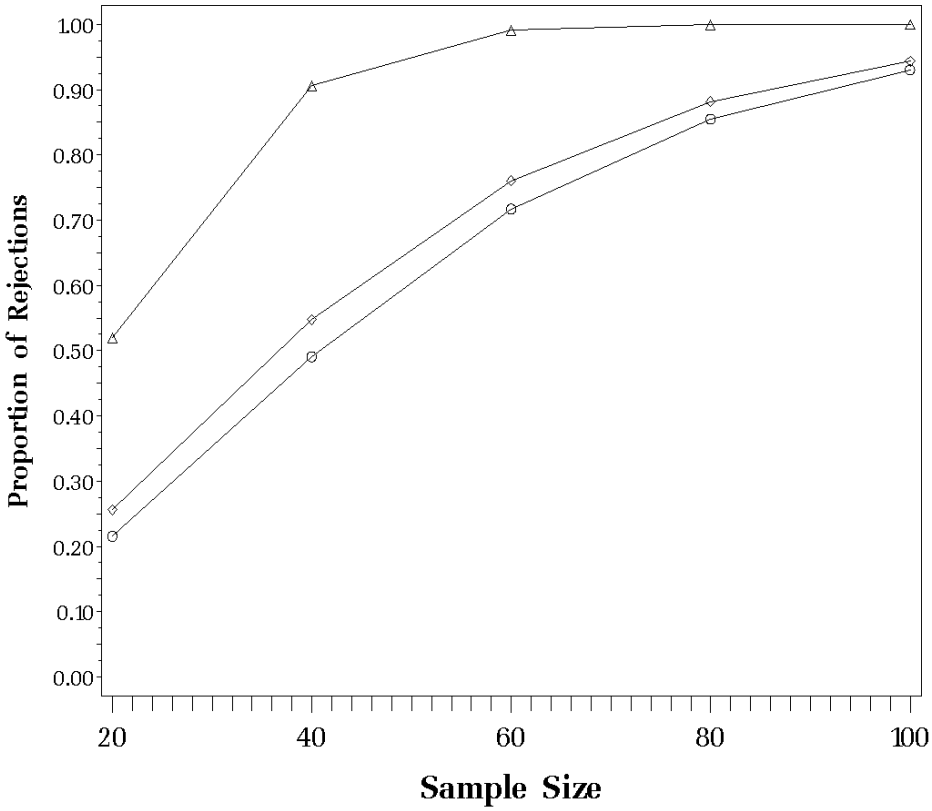

Figure 4.6: Observed power when the standardized difference in regime slopes is -10.0 and the common standard deviation is one ($\sigma = 1$) **for changepoint locations of** $0.25n$ (○), $0.50n$ (△), $0.75n$ (◇)**.**

termine whether or not a changepoint existed, and if so, at what point it was thought to lie in the total phosphorus gradient. The two-sample t test and deviance methods were also applied in order to make a comparison. The data consisted of 53 observations, shown in Figure 4.7 where the percent ash is plotted versus the amount of total phosphorus.

The deviance method detected a changepoint at the 12^{th} observation (at which TP was 292 mg/kg) and for which the reported p-value was approximately 0 when using the default precision level on SAS 9.2. The t test method detected a changepoint at the 9^{th} observation, corresponding to a TP of 260 mg/kg and for which the p-value was 2.010×10^{-8}. However, the bootstrap t method, based on $b = 1\,000$ bootstraps, gave a p-value of $\tilde{p}_t = 0.005$. The likelihood method identified the 22^{nd} observation (TP=502

mg/kg) as the location of the changepoint with a p-value of $p_\lambda = 0.000029$. The F methods also estimated the changepoint to be TP=292 mg/kg (the 12^{th} observation) with a p-value of $p_F = 4.028 \times 10^{-6}$ and $\tilde{p}_F = 0$ from the $b = 1000$ bootstraps. Figure 4.7 shows the data along with the models under a mean-shift model using the changepoint of $TP = 260$ mg/kg from the bootstrap t test and under the hockey-stick model with the estimated changepoint of $TP = 502$ mg/kg. It is certainly worth noting that regardless of the assumed deterministic model, mean-shift or hockey-stick, the grouping of observations with phosphorus levels ranging roughly from 300 mg/kg to 500 mg/kg and with percent ash approximately between 45 and 60 seem atypical.

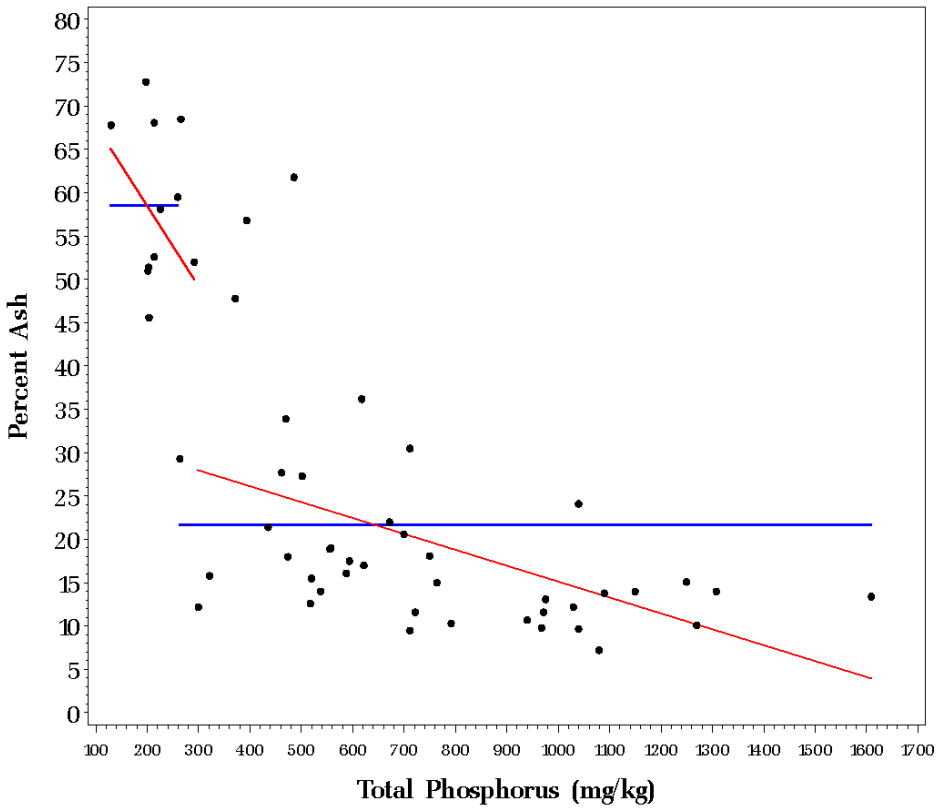

Figure 4.7: Plot of the ASH and TP values along with the estimated mean-shift and hockey-stick models.

Interestingly, each of the different changepoints have potential ecological significance reflecting a transition from a sawgrass system to a cattail dominated system. The higher changepoint supports the observation that when the soils average TP contents

exceed 500 mg/kg, the percent cover of slough species is dramatically decreased as the ecosystem transitions to one that is ultimately cattail dominated [Hagerthey et al., 2008]. In contrast, the lower TP content of 200-300 mg/kg is the average soil TP concentration at sites that are dominated by calcareous periphyton.

4.5 Discussion

From the simulation results the inadequacy of the t test and deviance methods for detecting a simple mean shift is apparent, as was expected. Neither lead to effective type I error rates near the nominal rate, chosen here to be $\alpha = 0.05$. A Bonferroni-style correction is only marginally useful in that it didn't completely correct the error rate for the deviance approach and was sensitive to the minimum group size for the basic t test. The deviance method also suffered in that it checks for a change in the mean, but assumes the variance remains unchanged before and after the changepoint. Both the likelihood method of Horváth [1993] and the proposed bootstrap t techniques gave reasonable false positive rates while not being overly sensitive to minimum group size and not assuming equal variance. In the hockey-stick model the F test has been shown to be poorly approximated by an F distribution, while the bootstrap F test of Julious has close to nominal type I error rates [Julious, 2001]. It is difficult to select a "best" approach as it depends on both the model and the location of the changepoint.

With respect to the mean-shift model, the likelihood method is superior to the bootstrap method when the sample size is small and the change is near the middle of the data, however with small sample size and a changepoint near the extremes of the gradient the bootstrap t method is superior. For larger sample sizes ($n \geq 60$) the likelihood and bootstrap t methods are nearly equivalent in terms of the power, as measured by the proportion of simulations for which the null hypothesis was rejected when it was, in fact, false. However, as mentioned above, the likelihood method is based on the maximum likelihood estimates of the group variances and with very small group sizes the possibility of observing k identical observations increases, especially in ecological and environmental applications where the response may be near the detection limit. As a re-

sult the likelihood based statistic, involving logarithms of the variance estimates, may be uncomputable in some cases since the variance estimate that would result in this situation would be zero. Observing k identical values is of course possible for the bootstrap technique as well, but as it is based on the two-sample t statistic as described above a variance estimate of zero in one regime does not prevent its computation.

If the underlying model is truly a hockey-stick model the bootstrap corrected F test has relatively high empirical power for only when the sample size is large and the changepoint is near the middle of the data. However, in light of the simulation results the importance of both obtaining baseline data and choosing an appropriate model are highlighted. The results clearly suggest that both techniques (likelihood and bootstrap t) have the highest power when the changepoint is in the middle of the data, all else being equal in the means model. However, it appears the critical factor is the amount of data present before and after the changepoint and not the relative location since with the larger sample sizes ($n \geq 60$) both methods have reasonable power for reasonably large changes: $\delta = 2$ in the mean-shift case and $\beta_{11} = 6$ in the hockey-stick case. Since the location of the changepoint is not assumed to be known *a priori* in these cases, this emphasizes the need for baseline data to help establish the behavior of the system in question before the change occurs.

A separate, but related, issue relevant in changepoint problems as discussed here is the selection of an appropriate model. The model ultimately chosen is of critical importance in that it is related to the necessary sample size, the hypotheses of interest and the selection of an appropriate test statistic. The underlying model assumed here, while a simplistic abrupt shift model, admits several hypotheses and therefore a choice of test statistic must be made. If the shift is assumed to be in the mean only and both groups, before and after the changepoint, can be assumed to have equal variance then approaches based on model deviance, a t statistic, or the likelihood method given here would all be reasonable choices. However, if the variances are not assumed to be equal, then the deviance based approach is no longer useful. However, the bootstrap t and likelihood methods don't test the same hypotheses either, since the bootstrap t allows for unequal variance but the likelihood tests for a change in variance - assuming it hap-

pens at the same location as the change in mean. The consequences of this can be seen in the example presented here. The likelihood method provided an estimate of the changepoint (502 mg/kg) that was much larger than the 260 mg/kg estimated by the bootstrap method, most likely due the fact that likelihood method is detecting a variance shift as well as a mean shift. As a result the likelihood method is most significant when the change in the mean and variance together are most significant, while the bootstrap t is only concerned about a mean shift.

The situation is somewhat more complicated if the shifts in mean and variance are not simultaneous, particularly if the mean shift were to precede the variance shift. If a more complicated model such as a piece-wise linear model were appropriate the aforementioned concerns remain, but the sample size becomes increasingly important since multiple parameters must be estimated before and after the changepoint. In such cases the bootstrap technique could be modified to accommodate such structure, even replacing the assumption of normal data with some other distribution would be feasible. However, the likelihood approach is based on the asymptotic distribution of the test statistic and as such would need to be carefully evaluated for applicability if the hypotheses, model or both were to change significantly.

In particular, the data set used here illustrates the importance of model selection, including the specification of the heterogeneity/homogeneity assumption being made as it can be practically significant in terms of changepoint estimation and model fitting. Of the models under consideration here neither model appears to provide a superior fit to the data. However, given the intent of setting a standard based on a change in the mean of the process, the mean-shift model seems to allow for a more satisfactory result. This is in part due to the fact that the two models estimate fundamentally different types of changepoints in an ecological sense. The mean-shift model assumes there is a period where the population is stable and a period where the population is stressed, with the changepoint indicating the level at which the stress occurs while the hockey-stick model assumes the first regime is during the crash (after the stress occurs) and the second regime represents a collapsed state. Hence in the hockey-stick model commonly used the changepoint estimated has a somewhat different interpretation than does the

one resulting from the mean-shift model. These same considerations apply to Bayesian models as well as the frequentist models discussed above.

4.6 Conclusions

Often environmental changes arise due to direct or indirect anthropogenic effects that can vary based on spatial, temporal or spatio-temporal factors. As such, large data sets often have a spatial, temporal or spatio-temporal component that needs to be addressed when carrying out any analysis, changepoint analysis in particular. A possible extension that has potential in the area of ecological modeling is incorporating changepoint methodology with quantile regression techniques in order to detect changepoints in the quantiles of a distribution. Often thresholds used in setting quality standards are obtained based on a particular quantile of the distribution of responses under certain conditions. As such, if methods for changepoint detection in quantile regression were readily available it could potentially affect the process by which such thresholds are established. In any case, the development of a bootstrap method that can detect both multiple types and locations of changepoints should be developed as the bootstrap methodology does not depend on the determination of an asymptotic distribution. This methodology would be of particular interest in finding a more suitable model for data such as that presented here, where a three-regime model of stable-collapsing-collapsed might be of even more interest.

CHAPTER 5

CHANGEPOINT DETECTION IN SPI TRANSITION PROBABILITIES

5.1 Introduction

The Standardized Precipitation Index (SPI) was developed by McKee et al. [1993] to identify drought conditions through historical precipitation data. The SPI was designed to be flexible to location and suitable for droughts at different hydrological levels. SPI is based on fitting a probability distribution to precipitation data, followed by a transformation to a standard normal distribution. The SPI values are based on the precipitation for the past several months, with common scales of 1, 3, 6, 12, 24 and 48 month(s). Guttman [1998] demonstrated that SPI compares favorably to the more prominent Palmer Index [Palmer, 1965, 1968]. Both indices are used for drought monitoring by organizations such as the U.S. National Oceanic and Atmospheric Administration (NOAA), the Pakistan Meteorological Department, the International Research Institute for Climate and Society (IRI), and the Famine Early Warning Systems Network (FEWS). Recent articles about SPI have been published in several journals including the *Journal of Hydrology, the International Journal of Climatology, Water Resources Management, Climate Change, Theoretical and Applied Climatology, the International Journal of Applied Earth Observation and Geoinformation, Atmosfera,* and *Natural Hazards and Earth Systems Sciences.*

McKee et al. [1993] define four drought categories based on the SPI values. Values in $(-1,0]$ represent mild drought, while values in $(-1.5,-1]$, $(-2,-1.5]$, and $(-\infty,-2]$ rep-

resent moderate, severe, and extreme drought, respectively. With these categories continuous SPI data can be converted into a series of drought states. Paulo et al. [2005] use these drought state series to construct 4×4 transition tables for the drought state at time t vs. time $t+1$. These tables are then analyzed using log-linear models or Markov chains with the goal of early warning detection of changepoints. Moreira et al. [2006] extend the use of transition tables and log-linear models to identify changes in the structure of these drought states by partitioning the SPI data into three periods. The categories used for these tables are slightly different: values in $[0,\infty)$, $[-1,0)$, $[-1.5,-1)$, and $(-\infty,-1.5)$ are used. These correspond to categories of no drought, mild drought, severe drought, and extreme drought, respectively. Each period is used in the construction of a 4×4 transition table, and tables are used to identify differences between periods. The main method of analysis in Moreira et al. [2006] is the use of confidence intervals for each cell in the 4×4 table for the odds-ratio of one period to another.

While reasonable as an exploratory device, this confidence interval approach in not equivalent to formal statistical testing. For example, 95% confidence intervals for two parameters p_1 and p_2 could overlap, but an $\alpha=0.05$ level test may still reject the null hypothesis that $p_1=p_2$ [Austin and Hux, 2002, Ryan and Leadbetter, 2002]. This phenomenon is well documented in the literature, in particular by Schenker and Gentleman [2001] who point out that investigating overlapping confidence intervals to determine significance rather than traditional significance testing is problematic for two reasons. The non-standard (overlapping confidence interval) method has a tendency to fail to reject too often when a difference truly exists and to reject the null less often than expected when no difference exists where, in both cases, the comparison is to the standard method. In Moreira et al. [2006] the analysis results in $3\times16=48$ different 95% confidence intervals and so there is considerable concern that the aforementioned drawbacks to this overlapping confidence interval approach may be influencing the final results.

Our goal is to extend the use of transition tables for SPI drought states to test for unknown changepoints in the SPI data. For simplicity we only focus on a single, abrupt change. Such a change would represent the time at which, before and after this time,

the drought state transition tables were most significantly different. Dealing with SPI data via log-linear models presents a challenge in that, due to the normalization process, certain cells are expected to have low cell counts. Since the normalization is with respect to the full data set it is possible for some transition matrices before and after a changepoint to even have zero counts. This leads to the usual problems of estimating odds-ratios in the presence of zero-counts [Agresti, 2002] but with the complication that collapsing columns and rows to eliminate the counts is an unsatisfactory solution since these rows/columns will be representing the most extreme drought conditions, where valid inferences are likely to be the most critical. Furthermore, there is inherent uncertainty when there is an *unknown* changepoint because subsequent analyses that treat the estimated changepoint as a known quantity will typically have effective type I error rates higher than the nominal level [Duggins et al., submitted].

To address these issues, we use a resampling algorithm where the resamples are taken from multinomial distributions fit to the transition tables. We use this technique, much like a bootstrap, to simulate appropriate critical values for our test [Wernecke, 1993]. We extend this technique to include maximizing over all potential changepoints. We show that for moderate to large data sets (60 - 120 years) these methods yield nominal type I error rates while achieving ample power. We further apply the method to rainfall data from Kenya in order to determine the time of a possible changepoint in the drought transition patterns for each of the four investigated sites with the goal of providing more accurate estimates of the current transition probabilities. Thus the method has potential benefits in both the case where a likely changepoint is detected and in the case where no changepoint is detected. In the former, the exclusion of pre-changepoint data when estimating the transition probabilities will allow for improved drought prediction capabilities while in the latter the analyst can safely use all available data during predictions. In essence, this test allows the drought managers to screen their data to determine how much of the available historical record should be pooled to estimate the current transition probabilities.

5.2 Methods

Changepoints in rainfall data can be evaluated with various techniques depending on whether the interest is odds-ratios between time periods, time-to-drought, transition probabilities between drought classes, etc. In particular, the focus of this chapter is on detecting changes in the probability structure of the empirical transition matrix without the need for any assumptions of an underlying model. To construct the empirical transition matrix for the data, the drought classes defined in Moreira et al. [2006] were used to classify the n observed SPI values, one at each time point, as 1 (no drought), 2 (mild drought), 3 (severe drought), and 4 (extreme drought) and then the number of transitions from each drought class to each other drought class were counted. The resulting counts were then divided by $n-1$, since the last observed SPI value does not transition to another state, yielding the empirical transition matrix. This analysis was based on the 12-month SPI values since they address more intermediate-range drought behavior, such as droughts affecting reservoirs, whereas the 3- or 6-month SPI lead to more frequent events of shorter duration. In contrast, the use of the 24- or 48-month SPI values would show fewer droughts of longer duration [McKee et al., 1993, Vicente-Serrano and López-Moreno, 2005]. Regardless of the time scale used, since the SPI values are standardized the expected relative frequencies for drought categories 1 - 4 are 0.500, 0.341, 0.092, 0.067, respectively.

In an attempt to ascertain whether a changepoint in the transition probabilities existed for some time $t = \tau$, a scanning algorithm was applied to the data and various metrics were employed to measure any discrepancy that might exist before and after time τ. The algorithm begins at a user-defined left boundary t_L and proceeds to a user-defined right boundary of t_R with $t_L < t_R$. At the initial step $k = t_L$ transitions are used to create one transition matrix while the remaining $n-1-k$ transitions are used to create a second matrix. The algorithm proceeds by increasing the sample size used for the first transition matrix, and thus decreasing the size of the second matrix, by one until the first matrix is based on $n-1-t_R$ transitions and the second matrix is based on the final t_R transitions. At each step the metric(s) of interest are used to measure the "distance" be-

tween the two empirical transition matrices. The changepoint estimate, $\hat{\tau}$, is the value of k for which the metric is maximized.

In order to ensure non-zero marginals the values of t_L and t_R must be set reasonably far away from the endpoints of the series of data. It is suggested that, when possible, 30 years of data should be used as a buffer on each end of the time-series, i.e. $t_L = 360$ and $t_R = n - 361$ and that guideline was employed here. By design the scanning algorithm is symmetric with respect to the direction of the search (right-to-left versus left-to-right) but is dependent on the proportion of the data falling before and after the changepoint. Specifically, if some proportion p of the transitions occur before time $t = \tau$ the power is not the same as if those $100p\%$ of the transitions had occurred after the changepoint when p is different from one-half.

Table 5.1 shows the generic layout of the transition matrices before (left) and after (right) a potential changepoint. The drought classes are as described above, with 1 corresponding to no drought and 4 corresponding to extreme drought. The probabilities $p_{ij}^{(k)}$ represent the probability of transitioning to state j given the current state is i, while in time period k. As such, the probabilities in each row are conditional probabilities and so they sum to one. In this case where at most one changepoint is assumed to be present, $k = 1$ represents transitions before the change and $k = 2$ represents those transitions occurring after the changepoint.

	1	2	3	4
1	$p_{11}^{(1)}$	$p_{12}^{(1)}$	$p_{13}^{(1)}$	$p_{14}^{(1)}$
2	$p_{21}^{(1)}$	$p_{22}^{(1)}$	$p_{23}^{(1)}$	$p_{24}^{(1)}$
3	$p_{31}^{(1)}$	$p_{32}^{(1)}$	$p_{33}^{(1)}$	$p_{34}^{(1)}$
4	$p_{41}^{(1)}$	$p_{42}^{(1)}$	$p_{43}^{(1)}$	$p_{44}^{(1)}$

	1	2	3	4
1	$p_{11}^{(2)}$	$p_{12}^{(2)}$	$p_{13}^{(2)}$	$p_{14}^{(2)}$
2	$p_{21}^{(2)}$	$p_{22}^{(2)}$	$p_{23}^{(2)}$	$p_{24}^{(2)}$
3	$p_{31}^{(2)}$	$p_{32}^{(2)}$	$p_{33}^{(2)}$	$p_{34}^{(2)}$
4	$p_{41}^{(2)}$	$p_{42}^{(2)}$	$p_{43}^{(2)}$	$p_{44}^{(2)}$

Table 5.1: Transition matrices before (left) and after (right) a change occurs at $t = \tau$. Row labels indicate the drought condition at time t while columns represent the drought condition at time $t+1$.

The scanning algorithm is designed to estimate the changepoint based on maximizing the "distance" between the two transition matrices. To quantify this concept metrics based on absolute, squared or square-root differences, both scaled and unscaled, were applied to the entire matrices, the diagonals only and the off-diagonals only. Metrics applied to the entire transition matrix were uniformly superior and the four metrics of primary interest are given below. $\mathscr{D}_1$ is the L_1, or Manhattan, distance and $\mathscr{D}_3$ is the square of the L_2, or Euclidean, distance. $\mathscr{D}_2$ and $\mathscr{D}_4$ are scaled versions of $\mathscr{D}_1$ and $\mathscr{D}_3$, respectively. These metrics were chosen because of their popularity in applications where measures of dissimilarity are of primary interest, such as in multivariate analysis [Krzanowski, 1990].

$$\mathscr{D}_1 = \sum_{i=1}^{4}\sum_{j=1}^{4}\left|\widehat{p}_{ij}^{(1)} - \widehat{p}_{ij}^{(2)}\right| \qquad \mathscr{D}_2 = \sum_{i=1}^{4}\sum_{j=1}^{4}\frac{\left|\widehat{p}_{ij}^{(1)} - \widehat{p}_{ij}^{(2)}\right|}{\bar{p}_{ij}}$$

$$\mathscr{D}_3 = \sum_{i=1}^{4}\sum_{j=1}^{4}\left(\widehat{p}_{ij}^{(1)} - \widehat{p}_{ij}^{(2)}\right)^2 \qquad \mathscr{D}_4 = \sum_{i=1}^{4}\sum_{j=1}^{4}\frac{\left(\widehat{p}_{ij}^{(1)} - \widehat{p}_{ij}^{(2)}\right)^2}{\bar{p}_{ij}}$$

Here $\bar{p}_{ij}$ is their weighted average of $\widehat{p}_{ij}^{(1)}$ and $\widehat{p}_{ij}^{(2)}$,

$$\bar{p}_{ij} = \frac{n_1\widehat{p}_{ij}^{(1)} + n_2\widehat{p}_{ij}^{(2)}}{n_1 + n_2}.$$

The changepoint is estimated to be the time, $\hat{\tau}$, that maximizes a particular metric when carrying out the scanning algorithm described above. In order to determine the statistical significance of the estimate an empirical $p-value$ was constructed by resampling the data's transition matrix. Specifically, $b = 1\,000$ samples were drawn by using the transition matrix to define a multinomial distribution for the sixteen possible transitions. For each resample the values of $\mathscr{D}_1$ through $\mathscr{D}_4$ were calculated, creating a sampling distribution used to compute the $p-value$ for each of the metrics. This $p-value$ is the proportion of the 1 000 resamples that result in a value of the metric at least as extreme as the value based on the original data. In order to evaluate the performance of each metric considered, both under the null and the alternative, simulations were run to determine the empirical type I error rate and empirical power.

In order to approximate the type I error rates, each metric was evaluated using $m = 20\,000$ simulations assuming the null hypothesis of no changepoint by using an ob-

served set of SPI data to create a "null" transition matrix. The particular data selected was chosen because it came from a site with a long data set (over 100 years of monthly observations) and due to the normalization process of the SPI values, all sites have similar transition matrices when based on the full set of transitions. Each simulation began by selecting $n = 1200$ transitions (corresponding to 100 years of monthly data), using the null transition matrix to define the multinomial distribution with sixteen cells from which the transitions were drawn. Using the scanning algorithm each metric was calculated on the null data and the maximum value of each was determined. The empirical type I error rate is then the proportion of the 20000 simulations that result in a $p-value$ less than the nominal error rate, say $\alpha = 0.05$ or $\alpha = 0.10$.

The empirical power was evaluated in a similar manner, excepting that the data for the simulations were generated by introducing a changepoint at a certain time, $t = \tau$ and $m = 10000$ simulations were used. Since there are an infinite number of ways in which the null hypothesis can be false, two scenarios of particular interest were chosen and investigated. One scenario of great interest is that of a shift to more prolonged droughts which would be signaled by a relative increase in the proportions in the lower quadrant of the transition matrix. These represent measurements in which both the initial state (state at current time) and final state (state at time $t+1$) are severe drought (3) or extreme drought (4). These are denoted as *Type-A* cells. A second scenario of interest is a shift to more frequent droughts, represented by a relative increase in the proportions where the initial state is severe (3) or extreme (4) drought but the final state is moderate drought (2), or vice versa. These are called *Type-B* cells.

The smaller number of simulations for the alternative case as compared to the null case was a conscious decision based on the error associated with the quantities of interest. Related to the idea of importance sampling (see, for example Johns [1998]) it is well-known that estimating small proportions requires relatively more simulations than estimating larger proportions. In context then, estimating the type I error rate (e.g. $\alpha = 0.05$) requires more simulations than does the power, which in general is expected to be larger than α. In this case choosing $\alpha = 0.05$ and $m = 20000$ reduced the standard error by approximately 30%, from 0.0022 to 0.0015. This results in a standard error whose relative

size as compared to the parameter of interest is more in line with the alternative cases, e.g. even a power of 20% has a standard error of 0.004 with only $m = 10\,000$ simulations. Thus it was determined that more than 10 000 simulations should be used for the null case. The decision to use twice as many was motivated from a practicality standpoint since the simulations for the type-A and type-B cells could be run at $m = 10\,000$ each, yielding 20 000 runs with no change present.

Regardless of the type of effect to be detected, decisions regarding the location of the changepoint in the series and the fraction of the counts to be placed before the changepoint must be made. The true changepoint was considered at two locations $\tau = 600$, corresponding to the midway point in the time series, and $\tau = 480$ which is 10 years earlier. Because the buffer was set at 360 and the method is symmetric in the location of τ, other choices were not considered. In splitting the amount of transitions that are to occur before and after time τ, the decision whether or not to use different ratios of observations in the four type-A or type-B cells needed to be made. For simplicity it was assumed that if a shift occurred, that all cells of a particular type were affected to the same degree, e.g. if a type-A cell such as (3,3) is to have 10% of its occurrences before τ and 90% after, then so are the other three type-A cells. The transitions were divided based on the log-odds, $\log \mathcal{O}$, of the proportion of transitions falling after the changepoint and was allowed to vary from -2.5 to 2.5 with steps of size 0.25 with log-odds of -2.5, 0 and 2.5 corresponding to approximately 7% of such transitions falling after τ and 93% before, the null case of 50% before and after, and approximately 93% after and 7% before τ. In this way the effect of a changepoint not at the midpoint of the series could be evaluated.

5.3 Results

5.3.1 Simulations

The methods were first evaluated under the null hypothesis of no changepoint in order to establish their effective type I error rates. Table 5.2 below shows the empirical

error rates, for $\alpha = 0.05$ and $\alpha = 0.10$, as measured by the proportion of simulations leading to a rejection. As discussed above the results in the null case are from two runs of 10 000 simulations, hence the sampling errors are $\sqrt{\frac{(0.05)(0.95)}{20000}} = 0.0015$ and $\sqrt{\frac{(0.10)(0.90)}{20000}} = 0.0021$ for the $\alpha = 0.05$ and $\alpha = 0.10$ simulations, respectively. As a result it can be seen that $\mathscr{D}_2$ is always too liberal and $\mathscr{D}_3$ is always too conservative, while $\mathscr{D}_1$ and $\mathscr{D}_4$ seem to be neither too liberal nor too conservative.

Metric	$\alpha = 0.05$	$\alpha = 0.10$
$\mathscr{D}_1$	0.0472	0.0986
$\mathscr{D}_2$	0.0570	0.1273
$\mathscr{D}_3$	0.0448	0.0914
$\mathscr{D}_4$	0.0472	0.0989

Table 5.2: Empirical type I error rates when $\alpha = 0.05$ and $\alpha = 0.10$.

From a detection standpoint, the optimal location for a shift to occur is in the middle of the available data, or $\tau = 600$ for this specific case. Thus, the power of the above metrics were first evaluated assuming that a change in the transition probabilities had occurred at time $t = 600$. Figure 5.1 shows the power, computed as the proportion of simulations leading to a rejection, plotted against the log-odds of a transition occurring after the changepoint for the case where $\alpha = 0.05$ and the change is restricted to type-A cells. From Figure 5.1 it is clear that $\mathscr{D}_2$ (△), the metric based on a scaled absolute distance, was inferior to $\mathscr{D}_1$ (□), $\mathscr{D}_3$ (◇) and $\mathscr{D}_4$ (○) which are based on a raw absolute distance, raw squared distance and a scaled squared distance, respectively. Excluding $\mathscr{D}_2$, the metrics are nearly indistinguishable for this case, for which similar results hold when $\alpha = 0.10$ is used.

For the case of changes only in the type-B cells, the empirical power curves are displayed in Figure 5.2. Here the four metrics show less separation, with $\mathscr{D}_2$ (△) no longer having the lowest power. In fact, in this case $\mathscr{D}_3$ (◇) is now the least powerful metric. As with the case of type-A cells, similar results hold when α is increased from five to ten percent.

Figure 5.3 gives the power curves for type-A cells when $\alpha = 0.05$ but the true changepoint has been shifted to $\tau = 480$, equivalent to a ten-year shift in the changepoint's

location. As expected the graphs are no longer symmetric, with the minimum shifted to $\log\frac{0.60}{0.40} = \log 1.5 = 0.4055$ given that the null case of no change corresponds to having 40% of the data before $\tau = 480$ since $\frac{\tau}{n} = \frac{480}{1200} = 0.40$. Again the scaled absolute distance ($\mathscr{D}_2$, $\triangle$) performs poorly, attaining significantly less power than the other metrics under consideration.

As with the previous case the asymmetry is again present in Figure 5.4 and as seen in Figure 5.3 the methods show less differentiation in power with $\mathscr{D}_2$ ($\triangle$) and $\mathscr{D}_3$ ($\diamond$) seemingly inferior to $\mathscr{D}_1$ ($\square$) and $\mathscr{D}_4$ ($\circ$). Increasing $\alpha = 0.05$ to $\alpha = 0.10$ has the expected effect of increasing the power of the methods, but as before does not affect the relative superiority of the metrics in terms of power.

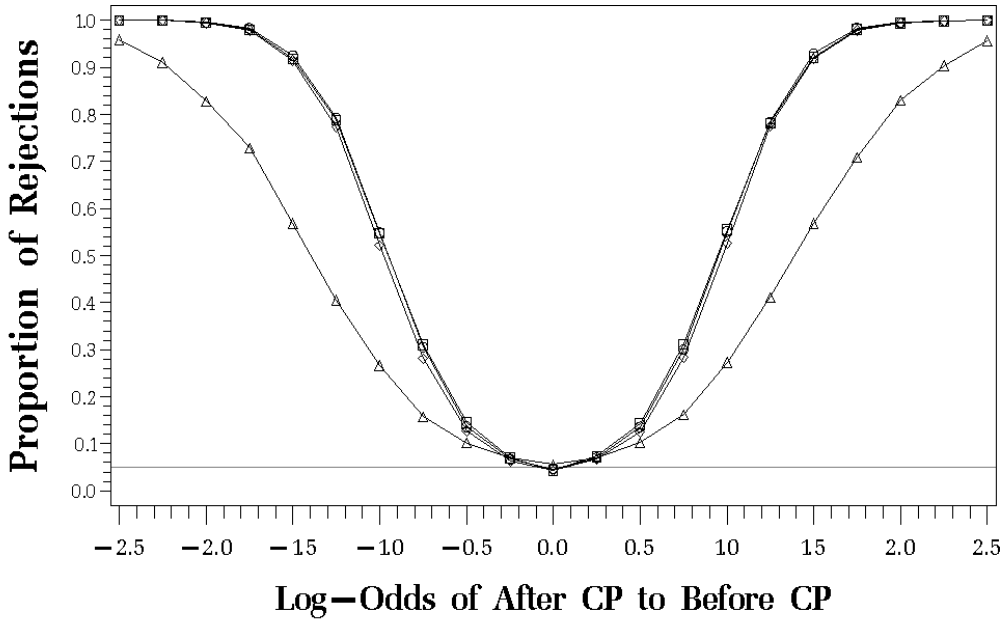

Figure 5.1: Empirical power for the case where $\alpha = 0.05$, the changepoint is in the middle of the data ($\tau = 600$) and changes occur only in type-A cells for $\mathscr{D}_1$ ($\square$), $\mathscr{D}_2$ ($\triangle$), $\mathscr{D}_3$ ($\diamond$) and $\mathscr{D}_4$ ($\circ$).

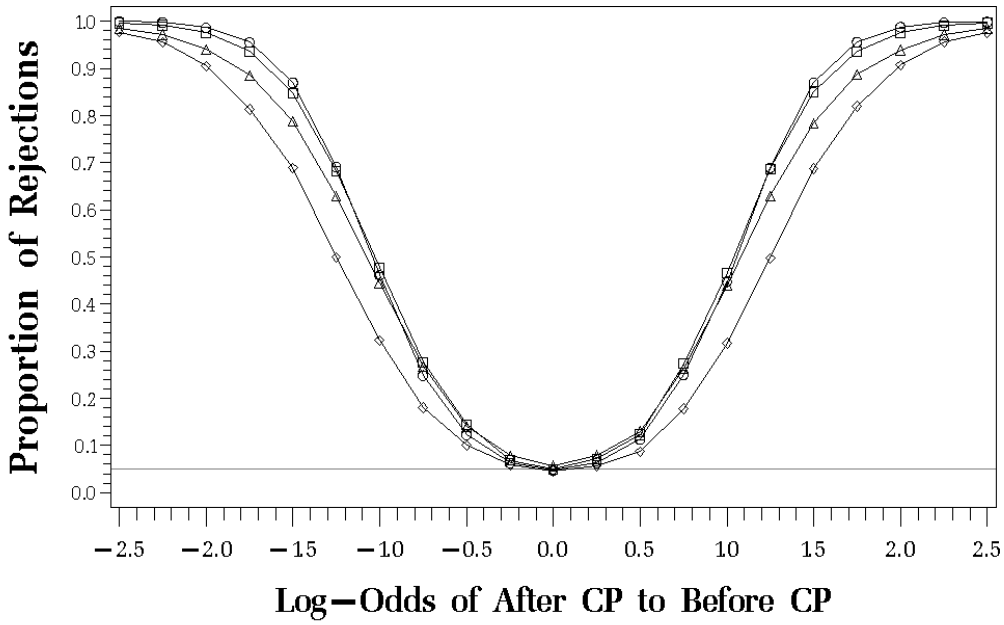

Figure 5.2: Empirical power for the case where $\alpha = 0.05$, the changepoint is in the middle of the data ($\tau = 600$) and changes occur only in type-B cells for $\mathscr{D}_1$ (□), $\mathscr{D}_2$ (△), $\mathscr{D}_3$ (◇) and $\mathscr{D}_4$ (○).

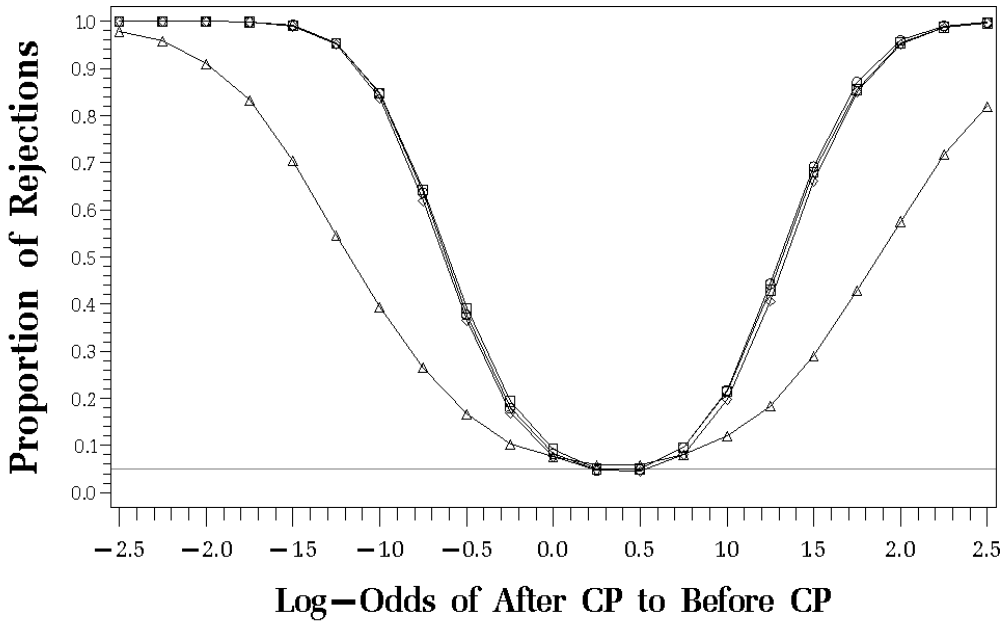

Figure 5.3: Empirical power for the case where $\alpha = 0.05$, the changepoint is offset from the middle of the data by ten years ($\tau = 480$) and changes occur only in type-A cells for $\mathscr{D}_1$ (□), $\mathscr{D}_2$ (△), $\mathscr{D}_3$ (◇) and $\mathscr{D}_4$ (○).

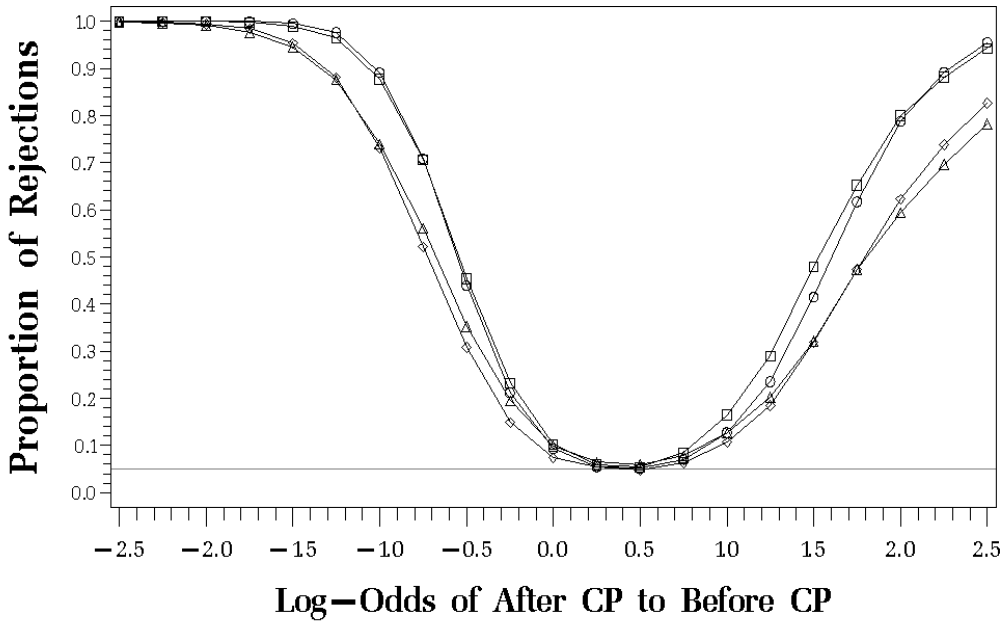

Figure 5.4: Empirical power for the case where $\alpha = 0.05$, the changepoint is offset from the middle of the data by ten years ($\tau = 480$) and changes occur only in type-B cells for $\mathscr{D}_1$ (□), $\mathscr{D}_2$ (△), $\mathscr{D}_3$ (◇) and $\mathscr{D}_4$ (○).

5.3.2 Examples

The four sites investigated are located in Kenya, and vary in elevation and length of record. These sites were chosen in particular because simple plots of 12-month SPI values (Figures 5.5-5.8) suggest that there is no change at Nairobi, possible changes at Kisumu and Voi, and an obvious change at Eldoret. Table 5.3 gives some potentially relevant characteristics about each site, specifically the length (years in which at least one data point is available), elevation (in meters), latitude and longitude. Each series starts in December and are of varying length. Eldoret is the shortest series with 40.5 years ($n = 487$), Kisumu's series is 86 years long ($n = 1033$), Voi's is 91 years ($n = 1093$) and Nairobi's is the longest of the four with 102 years of data ($n = 1225$).

The time-series plots of the 12-month SPI values for each of the four sites are given in Figures 5.5 through 5.8. Figures 5.7 and 5.8 do not seem to give any reason to believe that a change exists while Figure 5.5 shows a plot with a possible changepoint and Figure 5.6 shows a clear changepoint. For each site the scanning algorithm described above

Site	Length	Latitude	Longitude	Elevation
Nairobi	1894-1996	-01.317	+036.917	+1624.0
Eldoret	1961-2005	+00.533	+035.283	+2104.0
Kisumu	1903-1991	-00.100	+034.750	+1146.0
Voi	1904-1996	-03.400	+038.567	+0579.0

Table 5.3: Name, data set length, elevation (in meters), latitude (+ indicates N) and longitude (+ indicates E) for each of the four sites under consideration.

was applied and the $p-values$ were obtained for each of the four aforementioned metrics. Due to the relative short time-span available for the Eldoret site, a buffer of only 10 years, or 120 time points, was used. For the remaining three sites a buffer of 30 years, or 360 time points, was used to help ensure the transition matrices always had nonzero marginal counts.

Table 5.4 shows the estimated changepoints ($\hat{\tau}$) and the $p-values$ for each of the described metrics. Estimated changepoints are reported as *year.month* with the actual data point included parenthetically for reference. For example, 1973.02 represents February, 1973 which is the 123^{rd} data point. From the resulting $p-values$ it appears that neither the Voi nor Nairboi sites have a significant changepoint present. However, the Kisumu and Eldoret sites have $p-values$ that indicate a changepoint is present. For Eldoret the estimated changepoint of January/February, 1973 (122/123) is very near the starting value of December, 1972 (120), possibly indicating that the change may have occurred prior to the first allowable changepoint. This changepoint is estimated by $\mathscr{D}_1$, $\mathscr{D}_3$ and $\mathscr{D}_4$ but $\mathscr{D}_2$ shows a significant changepoint occurring in July, 1979 (200) with a $p-value$ of 0.30. From the plot in Figure 5.6 this is plausible. Kisumu exhibits similar results with its changepoint estimated to be $\hat{\tau}=453$, corresponding to August, 1943, based on all but $\mathscr{D}_3$, which produces an estimate of $\hat{\tau}=653$ which is April, 1960. The $p-value$ of 0.088 is not significant at the $\alpha=0.05$ level being used, but is low enough to warrant attention.

Due to Eldoret's estimated changepoint falling so near the boundary of the searchable area, a secondary analysis was performed. The estimated value based on $\mathscr{D}_1$ and $\mathscr{D}_3$, for example, was February, 1973 which is only three months into the search area. For the

Site	Metric	$\hat{\tau}$	$p-value$
Eldoret	$\mathscr{D}_1$	1973.02 (123)	0.000
	$\mathscr{D}_2$	1979.07 (200)	0.030
	$\mathscr{D}_3$	1973.02 (123)	0.000
	$\mathscr{D}_4$	1973.01 (122)	0.001
Voi	$\mathscr{D}_1$	1939.02 (399)	0.559
	$\mathscr{D}_2$	1935.11 (360)	0.167
	$\mathscr{D}_3$	1935.11 (360)	0.681
	$\mathscr{D}_4$	1935.11 (360)	0.103
Kisumu	$\mathscr{D}_1$	1943.08 (453)	0.000
	$\mathscr{D}_2$	1960.04 (653)	0.088
	$\mathscr{D}_3$	1943.08 (453)	0.000
	$\mathscr{D}_4$	1943.08 (453)	0.000
Nairobi	$\mathscr{D}_1$	1946.04 (617)	0.251
	$\mathscr{D}_2$	1965.03 (844)	0.182
	$\mathscr{D}_3$	1946.04 (617)	0.356
	$\mathscr{D}_4$	1959.06 (775)	0.140

Table 5.4: $P-values$ **for** $\mathscr{D}_1$ **-** $\mathscr{D}_4$ **based on** $b = 1\,000$ **resamples.**

second analysis the boundary was reduced from ten to eight years (from 120 to 96 time points). Table 5.5 gives the $p-values$ and estimated changepoints from this secondary analysis. After resetting the boundary $\mathscr{D}_1$ and $\mathscr{D}_3$ give estimates in line with the previous values and bear out the supposition that the changepoint was near the original boundary of 120. Specifically, they give estimates of October and September, 1972 respectively. Furthermore, even $\mathscr{D}_2$ now also gives September, 1972 as the estimated time of change, though with an insignificant $p-value$ of 0.095. In contrast $\mathscr{D}_4$ has moved away from the late 1972/early 1973 time frame and is now estimating the change to be in November, 1970 which is time point 96 - the initial value for the search.

Site	Metric	$\hat{\tau}$	$p-value$
Eldoret	$\mathscr{D}_1$	1972.10 (119)	0.002
	$\mathscr{D}_2$	1972.09 (118)	0.095
	$\mathscr{D}_3$	1972.09 (118)	0.003
	$\mathscr{D}_4$	1970.11 (96)	0.008

Table 5.5: $P-values$ **for** $\mathscr{D}_1$ **-** $\mathscr{D}_4$ **from the Eldoret site based on** $b = 1\,000$ **resamples, using an 8-year rather than 10-year buffer.**

Time-series plots for the four sites, each based on the 12-month SPI. In cases where a significant changepoint was detected using the algorithm presented here a vertical reference line is inserted at a value of $\hat{\tau}$ given in Table 5.4 and Table 5.5.

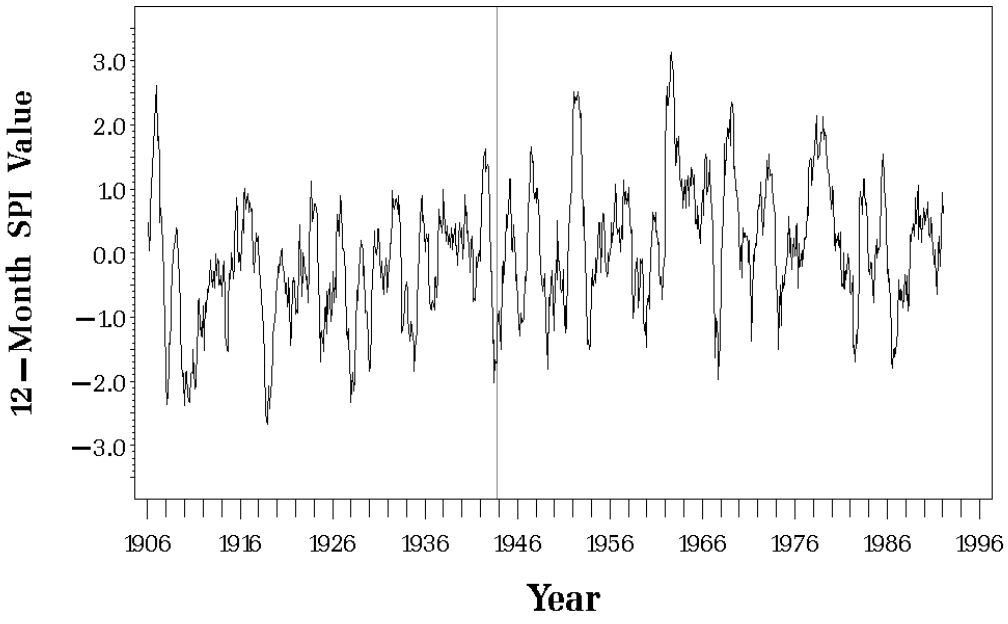

Figure 5.5: Time-series plot of monthly SPI values for Kisumu using the 12-month SPI. The reference line is at the 453^{rd} observation.

5.4 Discussion

From the simulations of type I error rate it was clear that $\mathscr{D}_3$ was the most conservative and $\mathscr{D}_2$ the most liberal methods, with their effective error rates being the approximately 10% too low and too high, respectively. While both $\mathscr{D}_1$ and $\mathscr{D}_4$ were conservative as well, they were much closer to nominal levels according to Table 5.2. More specifically $\mathscr{D}_2$ had empirical rates about 20% larger than nominal while $\mathscr{D}_3$ was about 10% too small, hence they were predisposed to reject too often ($\mathscr{D}_2$) or too rarely ($\mathscr{D}_3$) by amounts deemed to be too large to be negligible. In contrast $\mathscr{D}_1$ and $\mathscr{D}_4$ had empirical rates that were only approximately 3.5% too low, making them very slightly conservative, especially as compared to $\mathscr{D}_2$ and $\mathscr{D}_3$.

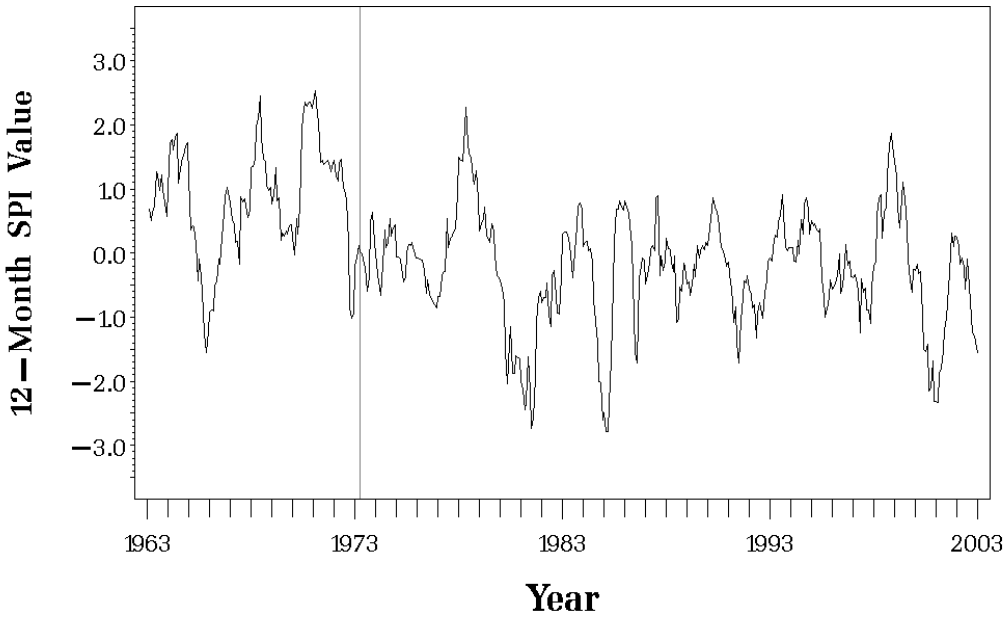

Figure 5.6: Time-series plot of monthly SPI values for Eldoret using the 12-month SPI. The reference line is at the 119^{th} observation.

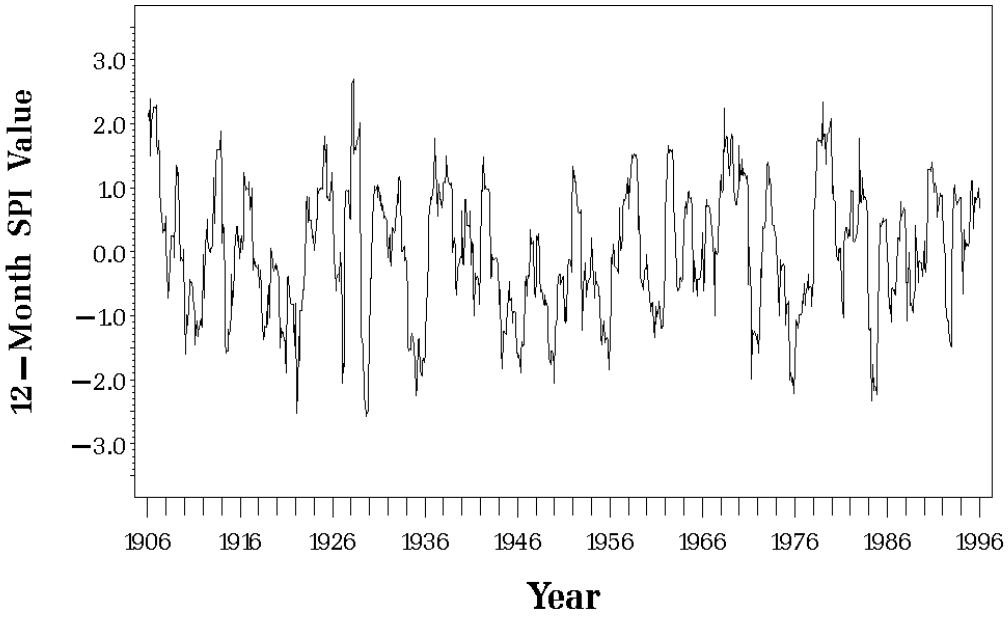

Figure 5.7: Time-series plot of monthly SPI values for Voi using the 12-month SPI.

Turning to the results from the simulations investigating power, the curves in Figures 5.1 through 5.4 show that, for the situations under consideration, $\mathscr{D}_2$ and $\mathscr{D}_3$ were con-

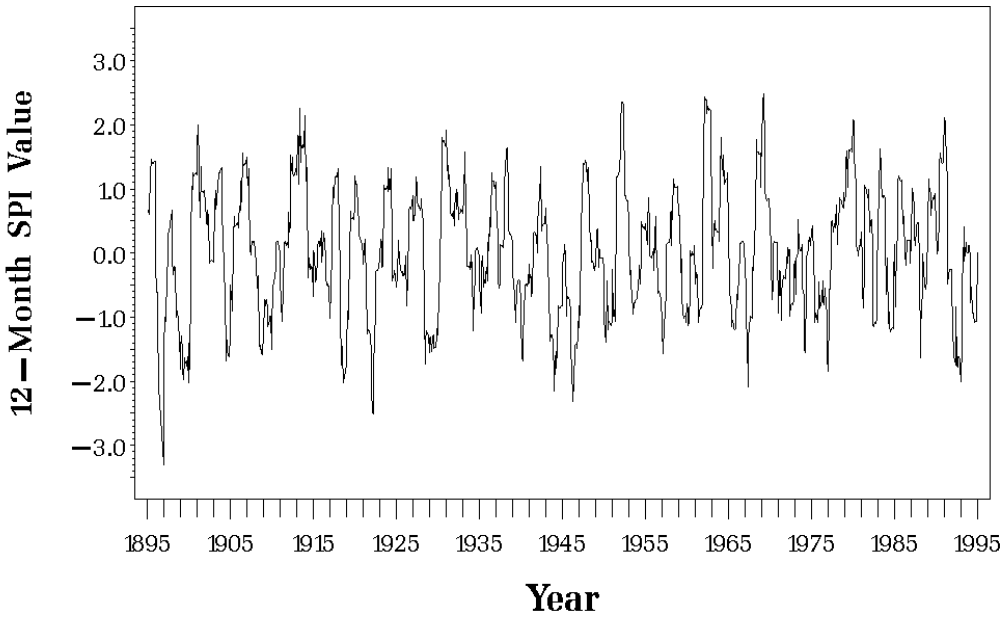

Figure 5.8: Time-series plot of monthly SPI values for Nairobi using the 12-month SPI.

sistently worse in terms of power as well. As with the study of type I error rates, $\mathscr{D}_1$ and $\mathscr{D}_4$ performed best in each scenario, often with clear separation from the other two metrics. This fact may be more easily seen through plots of the *difference* in power between the metrics. Figure 5.9 shows the difference between the metrics of interest in terms of their power. Computation of the median differences indicates that, indeed, $\mathscr{D}_2$ is vastly inferior to the other methods and that $\mathscr{D}_1$ is far superior in that it is almost always the most powerful metric, with the only caveat being values of the log-odds very close to zero.

Figure 5.10 shows a slightly different picture, with there being more clear differences between the metrics. In this case $\mathscr{D}_1$ is still a better choice than $\mathscr{D}_2$ and $\mathscr{D}_3$, but with $\mathscr{D}_4$ occasionally a slight favorite. However, the magnitude of the difference between $\mathscr{D}_1$ and $\mathscr{D}_4$ is relatively small, with a maximum absolute difference of 0.028, and so again $\mathscr{D}_1$ appears to be the superior method. The results for the cases where $\tau = 480$ are almost identical in the case of type-A cells and, while slightly different in terms of how $\mathscr{D}_2$ and $\mathscr{D}_3$ compare, the results for type-B cells are still more than comparable to those in the $\tau = 600$ case. In any case, $\mathscr{D}_2$ and $\mathscr{D}_3$ are seen, both graphically and numerically, to be

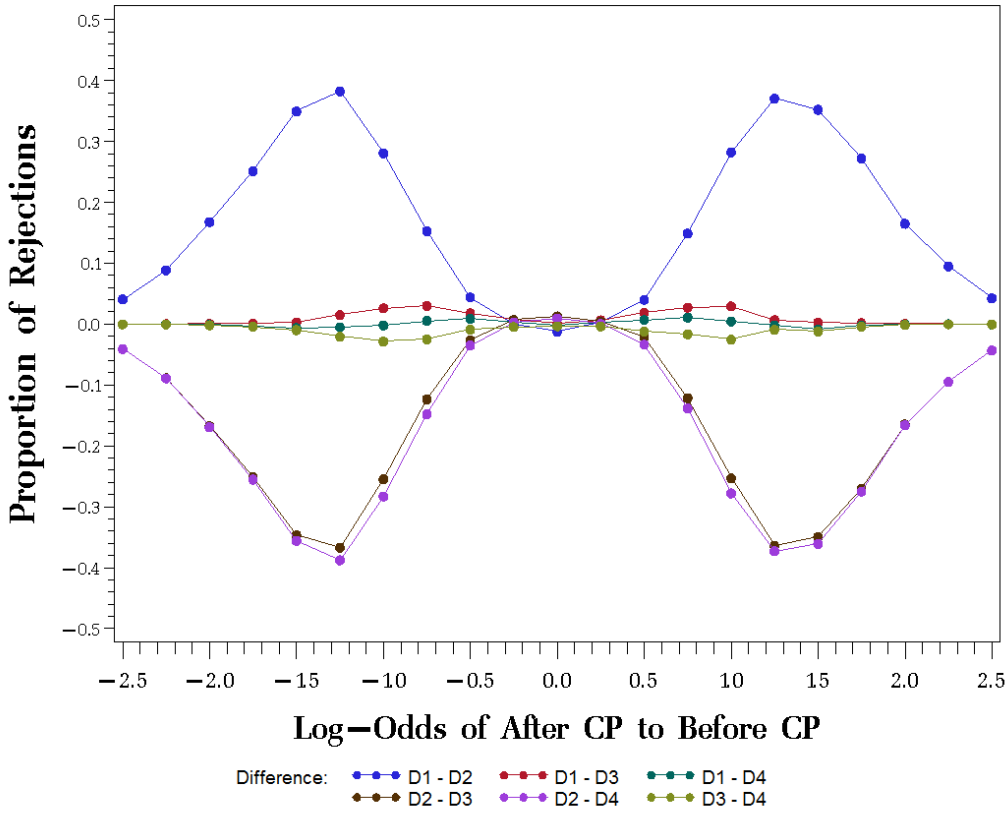

Figure 5.9: Differences between the metrics' empirical power for type-A cells, $\alpha = 0.05$ **and a changepoint in the middle of the data** ($\tau = 600$)

inferior as compared to $\mathscr{D}_1$ and $\mathscr{D}_4$. In addition it would appear that $\mathscr{D}_1$ is the better choice since it tends to outperform $\mathscr{D}_4$, even if only slightly, and would lend itself to simpler interpretations in practice.

When used to detect changepoints in non-simulated data $\mathscr{D}_1$ and $\mathscr{D}_4$ tend to consistently estimate the changepoint, when one is believed to exist. In Voi and Nairobi, where changepoints were not thought to exist, the estimates and their associated $p-values$ would seem to differ. In contrast, $\mathscr{D}_2$ and $\mathscr{D}_3$ do not seem to agree regularly in either case. In fact, they only agree in two cases, namely for Voi where they are associated with wildly different $p-values$ and with the Eldoret location after the endpoint was set at 8 rather than 10 years. For that case even $\mathscr{D}_1$ and $\mathscr{D}_4$ don't agree, though that may be a result of instability due to the likely location of the changepoint being close to the boundary of

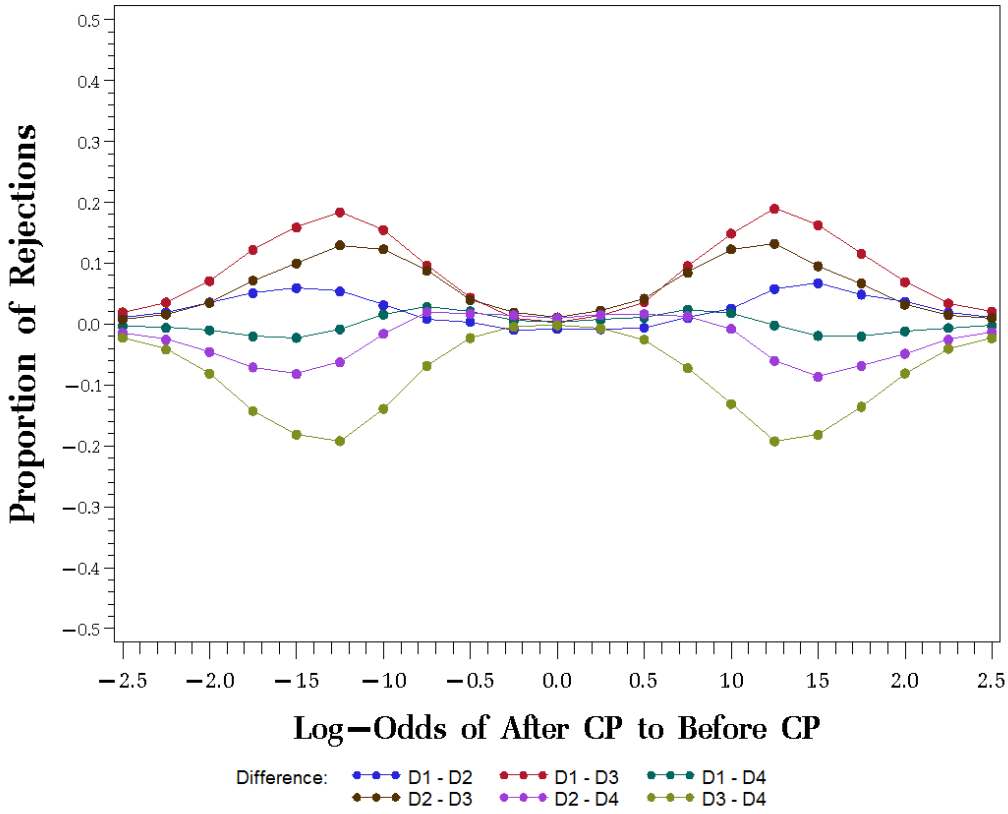

Figure 5.10: Differences between the metrics' empirical power for type-B cells, $\alpha = 0.05$ **and a changepoint in the middle of the data** ($\tau = 600$)

the search region.

As a result of the above discussion $\mathscr{D}_2$ and $\mathscr{D}_3$ would be considered poor choices due to their deviations from the nominal type I error rate and their inferior power. Overall it appears $\mathscr{D}_1$ and $\mathscr{D}_4$ are the superior metrics studied here and, though $\mathscr{D}_1$ seems to edge out $\mathscr{D}_4$ in terms of power, it is recommended that both be used in application since the different methods may be able to detect changepoints differently.

5.5 Conclusions

In contrast to methods that require the fitting of models, the resampling method given here is model independent. It relies only on the empirical transition matrix gener-

ated by using the SPI values to define drought states. Thus a greater flexibility is allowed since time-series models do not have to be fit to the original data nor do log-linear models need to be fit to the counts that generate the transition matrix. Furthermore, since the $p-values$ are determined by resampling, no correction factor is needed to account for the multiple comparisons.

Though this method does not require the fitting of a model, there is still the assumption that at most one changepoint exists in the data. With long-term hydrological data such as rainfall this is possibly unrealistic since climate patterns, both long and short term, may lead to the presence of multiple changepoints. One difficulty in assessing the presence of multiple changepoints is that the need for data on both ends of the search area to provide sufficient counts in the tables would necessitate very long data sets in order to look for multiple changepoints. As a special case of the multiple changepoint scenario is the case where the change is gradual rather than abrupt. In this case there would be two changepoints, one before and one after the gradual change, giving three transition matrices. The scanning algorithm provided here is expected to be generalizable to this situation as well as other cases where changepoints may exist in transition matrices.

CHAPTER 6

DETECTING A CHANGEPOINT IN A MULTIVARIATE MEAN-SHIFT MODEL

6.1 Introduction

IT is often the case in many applications that more than one response variable is measured on each experimental unit. This situation gives rise to a variety of ways in which changepoints can manifest in multivariate analysis – ranging from the idealistic of having all responses follow a similar model and a single, simultaneous (with respect to the gradient along which the data is ordered) changepoint to the complex where the response variables follow disparate models with varied changepoints and even a different number of changepoints across models. The further complication of dependencies among the response variables creates an even greater challenge in developing valid testing procedures for such situations. In cases where dependency is not negligible the first step is to apply a dimension-reduction technique such as partial least squares (PLS) or principal component analysis (PCA). Because the data to be considered here only has a single explanatory variable PLS would necessarily yield only one response factor. Thus, PCA was chosen because of its flexibility in allowing the user to determine the number of components to be retained. In the terminology introduced in Chapter 2 PCA lends itself to multivariate-R changepoint analysis while PLS in this case would be multivariate-E.

If a changepoint is found to exist, follow-up analyses should be carried out in order to diagnose the subset of principal component responses that are actually thought to be changing. This follow-up analysis is much less onerous when PCA has been used to

reduce the dimensionality since typically many less models will need to be considered and so the ramifications from any multiple testing will be greatly reduced. Furthermore if the responses are highly associated then reducing dimensions is appropriate since the changepoint detection may be influenced by the multicollinearity.

6.2 Methods

In order to develop a test for changepoint detection, assumptions regarding the deterministic and stochastic portions of the model must be made. Furthermore decisions regarding the type of data to be investigated and the testing criteria must be made. The p responses to be analyzed here are assumed to be independently and identically distributed according to a multivariate normal distribution, i.e. $\mathbf{y}_i \sim N_p\left(\boldsymbol{\mu}_{y_i}, \boldsymbol{\Sigma}_y\right)$. Since the application of interest is threshold detection each experimental unit is also presumed to be associated with some value of a stressor or gradient variable, x, so $p+1$ measurements are taken on each unit. While it is of interest to determine if the mean vector $\boldsymbol{\mu}_y$ changes after a certain threshold value of x, it is assumed that the underlying covariance structure of the responses is unaffected. As such, $\boldsymbol{\Sigma}_y$ represents the true covariance matrix after adjusting for any changepoints.

Prior to any analysis principal components are extracted and a decision regarding the number to be kept is made. PCA is applied to the correlation matrix of the responses, $\mathbf{R}_y$, and a decision is subsequently made as to the number of components to be kept. This decision may be based on one or several factors, such as the magnitude of the eigenvalue or percent of variance explained [Rencher, 2002, Chapter 12]. Denote this reduced dimensionality by l, so $1 \leq l \leq p$, and the principal components by $\mathbf{W}$. Then since $\mathbf{W} = \mathbf{YL}$ we have $\mathbf{w}_i \sim N_l\left(\boldsymbol{\mu}_{w_i}, \boldsymbol{\Sigma}_w\right)$ where $\mathbf{w}_i$ is the $1 \times l$ vector of principal component scores for the i^{th} unit.

As before $\boldsymbol{\Sigma}_w$ represents the covariance matrix of the principal components within regime and as such is not necessarily diagonal since the principal components are extracted based on the full set of n observations, regardless of regime. The resulting data data, $(\mathbf{x}, \mathbf{W})$, is sorted according to $\mathbf{x}$ so that the application of the search algorithm de-

scribed below results in an estimate of the critical threshold in principal component space. The w subscripts are hereafter omitted, unless they are needed to ensure clarity.

6.2.1 Models and Hypotheses

Under the null hypothesis of no changepoint $\boldsymbol{\mu}$ is assumed to be constant for all n observation vectors. This results in the no changepoint model given in (6.1).

$$\mathbf{w}_i = \boldsymbol{\mu} + \boldsymbol{\varepsilon}_i \quad i = 1, \ldots, n \tag{6.1}$$

However, under the general alternative there exists a changepoint, τ, such that after $x = \tau$ the mean vector $\boldsymbol{\mu}_1$ changes to $\boldsymbol{\mu}_2$ and each of the l principal components are allowed to exhibit some shift. More specifically, under the alternative we have $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ and each of the l components of $\boldsymbol{\delta}$ are allowed to be non-zero, leading to Model (6.2) below.

$$\mathbf{w}_i = \begin{cases} \boldsymbol{\mu}_1 & x_i \leq \tau \\ \boldsymbol{\mu}_2 & x_i > \tau \end{cases} \tag{6.2}$$

Thus the hypotheses of interest are

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$$

versus

$$H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2.$$

In Model (6.2) it is assumed that the covariance matrix, $\boldsymbol{\Sigma}$, is the same before and after any changepoint. Thus Model (6.1) and Model (6.2) differ in that the former only estimates l means while the former estimates $2l$ such means. It should be noted that since under the null hypothesis the changepoint, τ, vanishes the models may not be considered nested.

6.2.2 Test Criteria

In order to diagnose whether or not a changepoint has occurred, a search for a possible changepoint was conducted by sequentially partitioning $\mathbf{W}$ into sub-matrices of size $k \times l$ and $(n-k) \times l$ where k ranged over all allowable changepoint indices. For the purposes of this study, the smallest allowable group size was $k = 3$, giving $n - 2(3) + 1 = n - 5$ possible changepoint locations. At each potential changepoint a likelihood-based statistic was computed and the maximum value over all possible $n-5$ values was recorded. The statistic used was $\mathscr{D} = 2(\ell_f - \ell_r)$ where ℓ_f denotes the log-likelihood under the full model (Model (6.2)). Similarly ℓ_r is the log-likelihood under the reduced model given in Model (6.1). As mentioned above, the reduced model is not nested in the full model, but this terminology is still used because the full model has a larger parameter space $\left(\mathbb{R}^{2l}\right)$ while the reduced model's parameter space is $\mathbb{R}^l$.

Recall that the responses are assumed to follow a multivariate normal distribution with covariance matrix $\mathbf{\Sigma}$. Then under the null hypothesis of no changepoint the log-likelihood is easily computed to be

$$\ell_r = -\frac{nl}{2}\log 2\pi - \frac{n}{2}\log\left|\widehat{\mathbf{\Sigma}}_p\right| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{w}_i - \widehat{\boldsymbol{\mu}}_0)\widehat{\mathbf{\Sigma}}_p^{-1}(\mathbf{w}_i - \widehat{\boldsymbol{\mu}}_0)^T \tag{6.3}$$

where $\widehat{\boldsymbol{\mu}}_0$ consists of the l sample means computed using all n observation vectors and $\widehat{\mathbf{\Sigma}}_p$ is the maximum likelihood estimator of the covariance matrix obtained by pooling the maximum likelihood sample covariance matrix estimators from the two regimes. Thus $\widehat{\mathbf{\Sigma}}$ is updated at each step of the search for the optimal changepoint, namely

$$\widehat{\mathbf{\Sigma}}_p = \frac{k\widehat{\mathbf{\Sigma}}_1 + (n-k)\widehat{\mathbf{\Sigma}}_2}{n} \equiv \frac{n_1\widehat{\mathbf{\Sigma}}_1 + n_2\widehat{\mathbf{\Sigma}}_2}{n}. \tag{6.4}$$

Note $\mathbf{\Sigma}_p$ is not necessarily diagonal since the principal components were computed based on all n vectors but $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ were based on k and $n-k$, respectively.

Analogously ℓ_f, derived under the assumption of normality and presuming that a changepoint is present at $\tau = k$, is given in (6.5).

$$\ell_f = -\frac{nl}{2}\log 2\pi - \frac{n}{2}\log\left|\widehat{\mathbf{\Sigma}}_p\right| - \frac{1}{2}\sum_{i=1}^{k}(\mathbf{w}_i - \widehat{\boldsymbol{\mu}}_1)\widehat{\mathbf{\Sigma}}_p^{-1}(\mathbf{w}_i - \widehat{\boldsymbol{\mu}}_1)^T$$

$$-\frac{1}{2}\sum_{i=k+1}^{n}(\mathbf{w}_i-\widehat{\boldsymbol{\mu}}_2)\widehat{\boldsymbol{\Sigma}}_p^{-1}(\mathbf{w}_i-\widehat{\boldsymbol{\mu}}_2)^T \tag{6.5}$$

Here $\widehat{\boldsymbol{\mu}}_1$ and $\widehat{\boldsymbol{\mu}}_2$ are the estimated mean vectors for the first and second regimes, respectively, and $\widehat{\boldsymbol{\Sigma}}_p$ is as described following Equation (6.3).

As a result the statistic to be used for the discrimination between the groups is given in Equation (6.6).

$$\mathscr{D}=\sum_{i=1}^{n}(\mathbf{w}_i-\widehat{\boldsymbol{\mu}}_0)\widehat{\boldsymbol{\Sigma}}_p^{-1}(\mathbf{w}_i-\widehat{\boldsymbol{\mu}}_0)^T-\sum_{i=1}^{k}(\mathbf{w}_i-\widehat{\boldsymbol{\mu}}_1)\widehat{\boldsymbol{\Sigma}}_p^{-1}(\mathbf{w}_i-\widehat{\boldsymbol{\mu}}_1)^T-\sum_{i=k+1}^{n}(\mathbf{w}_i-\widehat{\boldsymbol{\mu}}_2)\widehat{\boldsymbol{\Sigma}}_p^{-1}(\mathbf{w}_i-\widehat{\boldsymbol{\mu}}_2)^T \tag{6.6}$$

This is different than the likelihood ratio procedure of Chen and Gupta [1995] because in addition to testing for a change in the mean vector across regimes they also test for a change in the covariance matrix. Here the covariance matrix is estimated in the same manner for both the null and the alternative.

Because Model (6.1) is not nested within Model (6.2), even if k were fixed there is no known distribution associated with $\mathscr{D}$. With the added component of maximizing $\mathscr{D}$ over all possible k to estimate τ the resulting statistic belongs to the extreme value family of distributions. As a result, determination as to the significance of any possible changepoint was made through the use of resampling – bootstrapping specifically – rather than asymptotic results.

6.2.3 Empirical Evaluation

Resampling was used to estimate the sampling distribution of $\mathscr{D}$ to establish the significance of any change in the likelihoods that resulted from fitting the full model as opposed to the reduced model. Bootstrapping was carried out by fixing the stressor value, x_i, and using case resampling on the principal component vector $\mathbf{w}_i$. Because the null hypothesis was assumed to be true, the value of x_i was not used in determining the bootstrap samples. Specifically, once x_i was fixed a $U(0,1)$ random variable was generated using SAS 9.2, which incorporates the Mersenne twister of Matsumoto and Nishimura [1998] in its 32-bit form. It was then transformed to integer via $\lfloor n\cdot u_i\rfloor+1$

where u_i represents the i^{th} uniform random variate and $\lfloor \cdot \rfloor$ is the floor function, used to select the observation (row of W) to be included in the bootstrap sample.

After the bootstrap sample was constructed by using these random values to select rows of **W**, the steps described above were applied to the bootstrap data resulting in a test statistic, $\mathscr{D}_j$, for $j = 1, \ldots, b$. For all resampling done here $b = 1\,000$ bootstraps were used. The $1\,000$ $\mathscr{D}_j$ are then used to estimate the empirical distribution of $\mathscr{D}$ and by comparing them to $\mathscr{D}_0$, the value from the original data, the significance of the changepoint is determined. The empirical $p - value$, $\tilde{p}$ is defined to be

$$\tilde{p} = \frac{\#\{\mathscr{D}_j \geq \mathscr{D}_0\}}{1\,000}$$

where $\#\{A\}$ indicates the cardinality of the set A.

6.2.4 Subset Determination

One drawback to the test statistic provided in Equation (6.6) is that it only compares the null model to a model that, with respect to the assumed mean model, would be considered saturated since it allows each of the l components to change. However, if only a few of the l principal components exhibit a change this is of both practical and statistical significance. If $\tilde{p}$ indicates that the H_0 above should be rejected in favor of the alternative, then it is prudent to conduct follow-up analyses with the goal of identifying those components that are likely to have a changepoint and separating them from those which are not likely to have a changepoint.

To isolate which responses are exhibiting a change the likelihood statistic from Equation (6.6) is used with the modification that, rather than the full model given above (i.e. all principal components change) some are allowed to change while the remaining responses constrained to not change. If an all-possible-subsets approach is used on the original data this yields 2^p possible models, while it only results in 2^l models after the PCA has been carried out. Since each model is compared to the null model, there are $2^l - 1$ tests to be carried out. If $l > 2$, then multiple testing concerns become more serious and must be addressed. A simple adjustment can be obtained by using a Bonferroni correction.

In each case the changepoint estimate, $\hat{\tau}$, used in constructing Equation (6.6) is now taken as fixed and the remaining $2^l - 2$ comparisons are done conditional on this being the location of the changepoint. For each sub-model estimates of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are obtained and a test statistic is obtained in a method similar to above, using the previously obtained $\widehat{\boldsymbol{\Sigma}}_p$. As expected, each sub-model has different $\widehat{\boldsymbol{\mu}}_1$ and $\widehat{\boldsymbol{\mu}}_2$ and while the test statistic can still be expressed as in Equation (6.6), $\mathscr{D}$ will have a different value because of the changes in $\widehat{\boldsymbol{\mu}}_1$ and $\widehat{\boldsymbol{\mu}}_2$ between the full and reduced models.

The SAS code developed for this application provides a file that includes the sorted model deviances, $\mathscr{D}$, for the models along with the responses assumed to have a changepoint in the sub-model. Because determination of the best model depends on subject matter expertise as well as statistical inference no pruning of possible models was included.

6.3 An Example

As discussed in Hagerthey et al. [2008] and Duggins et al. [submitted] the Everglades have suffered from an increase in phosphorus levels due to human influence. A potential ecological impact of such an increase is a change in the vegetation make-up of the Everglades – specifically a transition from a sawgrass (*Cladium jamaicense*) to a cattail (*Typha domingensis* Pers. and *T. latifolia* L.) dominated ecosystem DeBusk et al. [2001]. In order to investigate this phenomenon soil chemistry data on the total soil phosphorus (tp), percent of ash in the soil (ash), calcium (ca), total carbon (tc), total organic carbon (toc) and total nitrogen (tn) were collected from the surficial soils of Water Conservation Area 2A (WCA-2A). These five response variables were thought to be surrogates for detecting phosphorus eutrophication and the subsequent loss of the calcareous periphyton mat [Duggins et al., submitted]. In total 53 observations were collected from WCA-2A and the multivariate changepoint detection method introduced above was applied to them.

After applying principal component analysis to the correlation matrix of the five responses it was decided that two components were to be retained. The first two principal

components had $\lambda_1 = 4.86$ and $\lambda_2 = 0.11$ and explained 99.39% of the total variance. The loadings are given in Table 6.1. While the first component's loadings are all approximately equal in magnitude, the second component is mostly driven by the amount of total nitrogen found at the site and was part of the motivation for including a second component with such a low eigenvalue. Table 6.2 gives the correlations between the original variables and the first two principal components.

Response	PC_1 Loadings	PC_2 Loadings
ash	-0.4514	0.1375
ca	-0.4504	0.2083
toc	0.4497	-0.2948
tc	0.4510	-0.2209
tn	0.4333	**0.8956**

Table 6.1: Loadings of the original $p = 5$ responses on the first $l = 2$ principal components.

Response	PC_1 Correlations	PC_2 Correlations
ash	-0.9953	0.0453
ca	-0.9930	0.0687
toc	0.9944	-0.0728
tc	0.9914	-0.0972
tn	0.9554	0.2952

Table 6.2: Correlations of the original $p = 5$ responses with the first $l = 2$ components.

Applying the methodology outlined above to the first two principal components the maximum value of $\mathscr{D}$ occurred at the 21^{st} point, corresponding to a phosphorus value of 486 mg/kg, which had a bootstrap based $p - value$ of 0.000. This changepoint estimate is in keeping with Hagerthey et al. [2008] which suggests that when soils exceed an average total phosphorus of 500 mg/kg the ecosystem becomes cattail dominated. Furthermore this result is similar to the estimate of 502 mg/kg provided in Duggins et al. [submitted] based on the likelihood method due to Horváth [1993]. Since a changepoint is determined to likely be present in the data the next step was to determine if any of the possible sub-models provided a better fit to the data under the assumptions of at most one changepoint (now fixed at $x = 486$) and a mean-shift model.

Table 6.3 gives the values of $\mathscr{D}$ for each of the $2^l = 4$ available models and an empirical $p - value$ computed from the bootstrap samples taken earlier. While the $p - values$ are not valid for testing the location of the changepoint, they are useful for model selection purposes. The first $l = 2$ columns are indicator values, where a "1" indicates that in that particular model that response has a changepoint while a "0" indicates no changepoint is present in that model for that response.

PC_1	PC_2	$\mathscr{D}$	$\tilde{p}$
0	0	0	N/A
0	1	2744.444	0.000
1	0	19918.488	0.000
1	1	19954.419	0.000

Table 6.3: $\mathscr{D}$ and $\tilde{p}$ for each of the $2^l = 4$ models for the Everglades data. For the response columns (PC_1, PC_2) a 1 indicates a changepoint is present in that component, while a 0 indicates no changepoint in that component.

Figure 6.1 shows the principal component scores plotted against the phosphorus gradient and a vertical reference line at the estimated changepoint of $\widehat{\tau} = 486$. From Figure 6.1 it would appear that PC_1 has a very prominent changepoint, while the change in the second principal component may even be negligible. However, as the second line in Table 6.3 indicates, the model with only a change in the second component was significantly better than the null model. Furthermore, applying the univariate changepoint detection procedure of Duggins et al. [submitted] using nitrogen as the only response and keeping total phosphorus as the gradient results in exactly the same changepoint location with a bootstrapped p-value of $\tilde{p} = 0.001$. The means (variances) before and after the change for the two principal components are given in Table 6.4.

6.4 Conclusions

Several key points are worth discussion in light of the proposed new methodology for dealing with changepoints in a multivariate setting, and specifically a multivariate-R setting. While the method was applied to a single example with $p = 5$ and $l = 2$, there

Regime	Component	
	PC_1	PC_2
$x \leq 486$	-2.0986	-0.1362
	(4.1457)	(0.1344)
$x > 486$	1.3772	0.0894
	(0.5381)	(0.0747)

Table 6.4: Means and variances in the two regimes for each of the principal components.

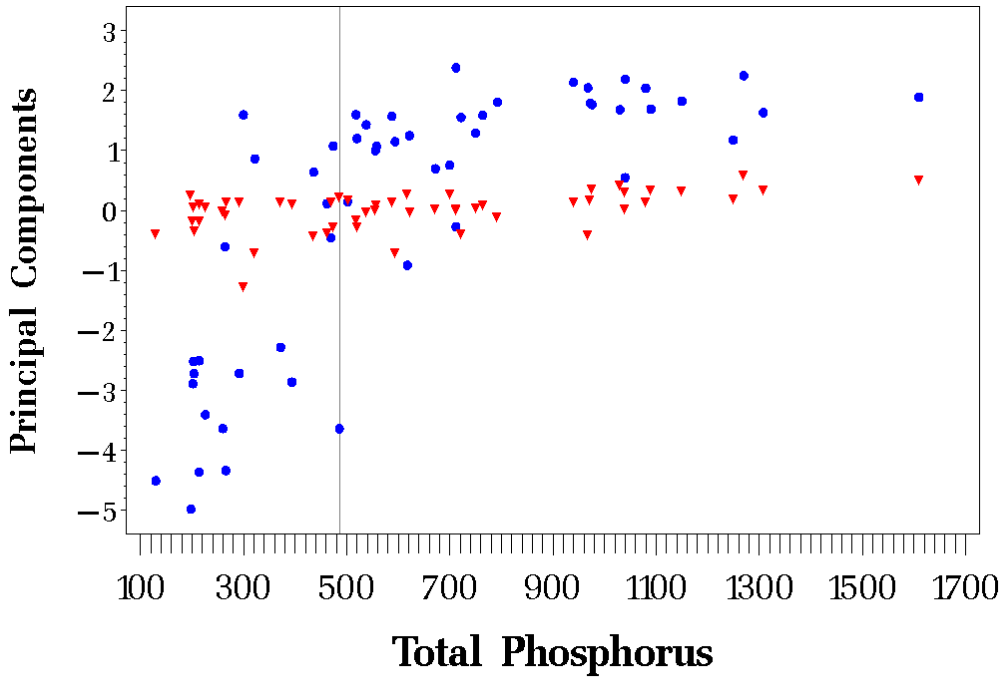

Figure 6.1: Plot of principal component values along with the estimated changepoint of $\widehat{\tau} = 486$. PC_1 is indicated by the blue points circles, while PC_2 is represented by the red triangles.

is no limitation in the number of dimensions that can be handled beyond any practical considerations for data collection and interpretation of the resulting components. However, if l is allowed to increase the issue of multiple testing mentioned earlier must be considered. Without the availability of distributional assumptions to guide the development of a more satisfactory method, the current proposal is a simple Bonferroni adjustment which is known to become overly conservative for a large number of inferences. As such, this method seems to have a wide appeal in that it is applicable to any

multivariate situation where the dimensionality can be reasonable reduced.

The introduction of multiple responses requires assumptions to be made about the relationship between the threshold and each of the responses. It is certainly feasible that while a means model with constant within-regime covariance matrix may be appropriate for some of the p responses, that model may be a poor choice for other responses. This may be due to a variance change in addition to (or in place of) a mean shift, or because the shift is not abrupt so that a piece-wise linear model may be a better fit for some or all of the responses. To address both of these concerns simultaneously the likelihood functions would need to be updated and a new estimator for $\mathbf{\Sigma}$ would need to be used, changing the final form of the statistic to be used for discrimination between possible regimes. The possibilities introduced by having changes in means, variances, covariances or even model structure at the changepoint are too numerous to reasonably quantify and so there should be considerable interest in developing a flexible methodology that can accurately detect changepoints in the presence of one or more of those changes.

The method provided here is a possible first step in that direction since it is new and likelihood based, providing the needed flexibility for dealing with changes in model structure and parameter values simultaneously. The lack of nested models necessitates the need for resampling methods, also demonstrated here.

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

The work presented in this dissertation is of great importance because changepoint problems are becoming more pervasive across multiple disciplines and as such is being applied to a variety of problems via many different techniques. Unfortunately, not all those techniques are acceptable from a statistical point of view. Before this dissertation there was no comparison of several of the more popular methods, even in the simplest of cases. The work presented here begins with that comparison in order to identify where the current methods are lacking and establishing resampling methods, in particular bootstrapping, as a viable alternative to asymptotic theory which is a poor approximation for many of the small sample sizes used in practice, as is evidenced by the examples provided here.

This comparison was done by using a search algorithm based on an old-hat method, the two-sample t-test, which has appeal in that it can be easily implemented by practitioners without the need for much, if any, additional training. Through the simplistic assumptions of Gaussian data undergoing either an abrupt mean-shift or a gradual mean-shift, it was shown that the naïve approaches were woefully inadequate for inferential purposes because of poor type I error rates. Even in the case of such models there is still some interesting work that could be done, in particular extending this result to multiple changepoints.

The changepoint problem is of great interest in environmental research, especially in climate change. Hydrology has begun looking into using changepoint analysis to improve prediction of drought states and durations, but the current analysis suffers from questionable analysis. This dissertation was thus extended to look at detecting changes

in a data with a different structure, in particular time-series data in the form of standardized precipitation index (SPI) values. It was shown that an essentially model-free approach to changepoint detection and drought prediction could be obtained.

This is invaluable in that it can be carried out without the need for analyzing the time-series structure of the data and allows for detection of two important events: increased drought duration and increased drought frequency. We show that the provided method, which is new to the literature, has reasonably high power for detecting changes and several metrics for easy detection are provided. There is great potential for research into one-sided analyses and the development of interval estimates for the changepoint, and ultimately for the predicted probabilities of an event.

The final analysis provided here is again a new development in changepoint analysis, namely the extension of the first analysis presented here to the case where multiple responses are measured. By making similar assumptions regarding the model's structure and data's distribution a two-stage analysis was developed that allowed for dimension reduction through principal component analysis and subsequently the estimation of the changepoint. After a likely changepoint is estimated, there is great interest in whether each of the responses (or components) are believed to have a changepoint and so a model selection procedure is introduced.

As the area of multivariate changepoint analysis is relatively unexplored, this last analysis would seem to hold the most potential for future research. The possibility of multiple changepoints, changes in multiple parameters and varying model structure between each response and the gradient suggest that a highly-flexible but easily implemented program for such analysis would be extremely beneficial for practitioners needing to set thresholds for multiple responses.

In both the univariate and multivariate response setting analyzed here, the departure from normality and independence would be obvious choices for future research as well. Changepoint analysis is closely connected to extreme value analysis such as is common in lifetime analysis which is important in both industrial settings and biosurveillance. In each case the need to detect a changepoint exists, but the data may be well known to be non-normal (e.g. exponential or Weibull) or the observations may be

dependent on one another (profile monitoring). In any of the cases presented here, and most likely in any future work, the need for lack-of-fit testing is paramount as there are many parameters being estimated in the multivariate cases and model selection would seem to be a crucial part of any analysis.

Bibliography

A. Agresti. *Categorical Data Analysis*. John Wiley & Sons, Hoboken, NJ, 2002.

E. Aly, A. Abd-Rabou, and N. M. Al-Kandari. Tests for multiple change points under ordered alternatives. *Metrika*, 57(3):209–221, 2003.

J. Antoch and M. Hušková. Permutation tests in change point analysis. *Statistics and Probability Letters*, 53(1):37–46, 2001.

P. Austin and J. Hux. A brief note on overlapping confidence intervals. *Journal of Vascular Surgery*, 36(1):194–195, 2002.

J. Bai and P. Perron. Estimating and testing linear models with multiple structural changes. *Econometrica*, 66(1):47–78, 1998.

J. Bai and P. Perron. Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18(1):1–22, 2003.

N. Barrowman and R. Myers. Still more spawner-recruitment curves: The hockey stick and its generalizations. *Canadian Journal of Fisheries and Aquatic Sciences*, 57:665–676, 2000.

M. Bartlett. Properties of sufficiency and statistical tests. *Proceedings of the Royal Statistical Society Series A*, 160(2):268Ű282, 1937.

P. Bhattacharya. Some aspects of change-point analysis. *IMS Lecture Notes - Monograph Series*, 23:28–56, 1994.

G. Bhattacharyya and R. Johnson. Nonparametric tests for shift at an unknown time point. *The Annals of Mathematical Statistics*, 39(5):1731–1743, 1968.

P. Bickel and D. Freedman. Some asymptotic theory for the bootstrap. *The Annals of Statistics*, 9(6):1196–1217, 1981.

L. Breiman, J. Friedman, C. Stone, and R. Olshen. *Classification and Regression Trees*. Chapman & Hall, New York, 1993.

R. Brown, J. Durbin, and J. Evans. Techniques for testing the constancy of regression over time (with discussion). *The Journal of the Royal Statistical Society B*, 37(2):149–192, 1975.

E. Carlstein. Nonparametric change-point estimation. *The Annals of Statistics*, 16(1): 188–197, 1988.

G. Casella and R. Berger. *Statistical Inference*. Duxbury, Australia, 2002.

J. Chen and A. Gupta. Lilelihood procedure for testing change point hypothesis for multivariate Gaussian model. *Random Operators and Stochastic Equations*, 3(3):235–244, 1995.

H. Chernoff and S. Zacks. Estimating the current mean of a normal distribution which is subjected to changes in time. *The Annals of Mathematical Statistics*, 35(3):999–1018, 1964.

P. Crowley. Resampling methods for computation-intenstive data analysis in ecology and evolution. *Econometrica*, 66(1):47–78, 1998.

R. Davies. Testing when a nuisance parameter is present only under the alternative. *Biometrika*, 64(2):247–254, 1977.

R. Davies. Testing when a nuisance parameter is present only under the alternatives. *Biometrika*, 74(1):33–43, 1987.

R. Davies. Hypothesis testing when a nuisance parameter is present only under the alternative: Linear model case. *Biometrika*, 89(2):484–489, 2002.

A. Davison and D. Hinkley. *Bootstrap Methods and Their Applications*. Cambridge University Press, Cambridge, UK, 1997.

W. DeBusk, S. Newman, and K. Reddy. Spatio-temporal patterns of soil phosphorus enrichment in Everglades water conservation area 2a. *Journal of Environmental Quality*, 30(4):1438, 2001.

J. Duggins, M. Williams, D. Kim, and E. Smith. Changepoint detection in SPI transition probabilities. *Hydrology*, 388:456–463, 2010.

J. Duggins, D. Kim, S. Newman, and E. Smith. Evaluating thresholds in ecological and environmental settings. *Ecologial Modelling*, submitted.

L. Dümbgen. The asymptotic behavior of some nonparametric change-point estimators. *The Annals of Statistics*, 19(3):1471–1495, 1991.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7 (1):1–26, 1979.

B. Efron and G. Gong. A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37(1):36–48, 1983.

B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, USA, 1993.

M. Evans, N. Hastings, and B. Peacock. *Statistical Distributions*. Wiley, New York, 2000.

T. Fan and W. Chen. Bayesian change points analysis on the seismic activity in northeastern Taiwan. *Journal of Statistical Computation and Simulation*, 75(11):857–868, 2005.

L. Gardner. On detecting changes in the mean of normal variates. *Annals of Mathematical Statistics*, 40(1):116–126, 1969.

E. Gombay and L. Horváth. An application of the maximum likelihood test to the change-point problem. *Stochastic Processes and their Applications*, 50(1):161–171, 1994.

P. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

G. Gurevich and A. Vexler. Guaranteed maximum likelihood splitting tests of a linear regression model. *Statistics*, 40(6):465–484, 2006.

N. Guttman. Comparing the Palmer drought index and the standardized precipitation index. *Journal of the American Water Resources Association*, 34(1):113–121, 1998.

S. Hagerthey, S. Newman, K. Rutchey, E. Smith, and J. Godin. Multiple regime shifts in a subtropical peatland: Establishing community specific thresholds to eutrophication. *Ecological Monographs*, 78(4):547–565, 2008.

D. Hawkins. *Computational Statistics and Data Analysis*, 37(3):323–341, 2001.

D. Hawkins and K. Zamba. Statistical process control for shifts in mean or variance using a changepoint formulation. *Technometrics*, 47(2):164–173, 2005.

D. Hawkins, P. Qiu, and C. Kang. The changepoint model for statistical process control. *Journal of Quality Technology*, 35(4):355–366, 2003.

B. Horness, D. Lomax, L. Johnson, M. Myers, S. Pierce, and T. Collier. Sediment quality thresholds: Estimates from hockey stick regression of liver lesion prevalence in english sole (*Pleuronectes vetulus*). *Environmental Toxicology and Chemistry*, 17:872–882, 1998.

L. Horváth. The maximum likelihood method for testing changes in the parameters of normal observations. *The Annals of Statistics*, 21(2):671–680, 1993.

L. Horváth and P. Kokoszka. Change-point detection with non-parametric regression. *Statistics*, 36(1):9–31, 2002.

D. Hudson. Fitting segmented curves whose join points have to be estimated. *Journal of the American Statistical Association*, 61:1097–1129, 1966.

M. Hušková and A. Slabý. Permutation tests for multiple changes. *Kybernetika*, 37(5): 605–622, 2001.

B. James, K. James, and D. Siegmund. Tests for a change-point. *Biometrika*, 74(1):71–83, 1987.

V. Jandhyala and S. Fotopoulos. Capturing the distributional behaviour of the maximum likelihood estimator of a changepoint. *Biometrika*, 86(1):129–140, 1999.

V. Jandhyala and S. Fotopoulos. Estimating the unknown change point in the parameters of the lognormal distribution. *Environmetrics*, 18:141–155, 2007.

W. Jensen, L. Jones-Farmer, C. Champ, and W. Woodall. Effects of parameter estimation on control chart properties: A literature review. *Journal of Quality Technology*, 38(4): 349–364, 2006.

M. Johns. Importance samping for bootstrap confidence intervals. *Journal of the American Statistical Association*, 83(403):709–714, 1998.

S. Julious. Inference and estimation in a changepoint regression problem. *The Statistician*, 50(1):51–61, 2001.

H. Kim and D. Siegmund. The likelihood ratio test for a change-point in simple linear regression. *Biometrika*, 76(3):409–423, 1989.

H. Kim, M. Fay, E. Feuer, and D. Midthune. Permutation tests for joinpoint regression with applications to cancer rates. *Statistics in Medicine*, 19(3):335–351, 2000.

H. Kim, M. Fay, M. Barrett, and E. Feuer. Comparability of segmented line regression models. *Biometrics*, 60(4):1005–1014, 2004.

R. King and C. Richardson. Integrating bioassessment and ecological risk assessment: An approach to developing numerical water-quality thresholds. *Environmental Management*, 31(6):795–809, 2003.

H. Koul and L. Qian. Asymptotics of maximum likelihood estimator in a two-phase linear regression model. *Journal of Statistical Planning and Inference*, 108(1-2):99–119, 2002.

W. Krzanowski. *Principles of Multivariate Analysis: A User's Perspective*. Oxford University Press, New York, 1990.

T. Lai. Sequential analysis: Some classical problems and new challenges. *Statistica Sinica*, 11(2):303–408, 2001.

C. Lee. Bayesian analysis of a change-point in exponential families with applications. *Computational Statistics and Data Analysis*, 27(2):195–208, 1998.

C. Loader. Change-point estimation using nonparametric regression. *The Annals of Statistics*, 24(4):1667–1678, 1996.

M. Mahmoud, P. Parker, W. Woodall, and D. Hawkins. A change point method for linear profile data. *Quality and Reliability Engineering International*, 23(2):247–268, 2007.

M. Matsumoto and T. Nishimura. Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation*, 8(1):3–30, 1998.

P. McCormick, S. Newman, and L. Vilchek. Landscape responses to wetland eutrophication: Loss of slough habitat in the Florida Everglades, USA. *Hydrobiologia*, 621(1): 105–114, 2009.

T. McKee, N. Doesken, and J. Kleist. The relationship of drought frequency and duration to time scales. In *Eighth Conference on Applied Climatology*, Anaheim, CA, January 1993.

E. Moreira, A. Paulo, L. Pereira, and J. Mexia. Analysis of SPI drought class transitions using loglinear models. *Journal of Hydrology*, 331(1-2):349–359, 2006.

H. Müller. Change-points in nonparametric regression analysis. *The Annals of Statistics*, 20(2):737–761, 1992.

E. Page. A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42(3/4):523–526, 1955.

W. Palmer. Meteorological drought. Research paper no. 45, U.S. Weather Bureau. [NOAA Library and Information Services Division, Washington, D.C. 20852], 1965.

W. Palmer. Keeping track of crop moisture conditions, nationwide: The new crop moisture index. *Weatherwise*, 21:156–161, 1968.

J. Paul and M. McDonald. Development of empirical, geographically specific water quality criteria: A conditional probability analysis approach. *Journal of the American Water Resources Association*, 41(5):1211–1223, 2005.

A. Paulo, E. Ferreira, C. Coelho, and L. Pereira. Drought class transition analysis through Markov and loglinear models, an approach to early warning. *Agricultural Water Management*, 77(1-3):59–81, 2005.

M. Perry and J. Pignatiello Jr. A change point model for the location parameter of exponential family densities. *IIE Transactions*, 40(10):947–956, 2008.

A. Pettitt. A simple cumulative sum type statistic for the change-point problem with zero-one observations. *Biometrika*, 67(1):79–84, 1980.

S. Qian, R. King, and C. Richardson. Two statistical methods for the detection of environmental thresholds. *Ecological Modelling*, 166(1-2):87–97, 2003.

R. Quandt. The estimation of the parameters of a linear regression system obeying two separate regimes. *Journal of the American Statistical Association*, 53(284):873–880, 1958.

R. Quandt. Tests of the hypothesis that a linear regression system obeys two separate regimes. *Journal of the American Statistical Association*, 55(290):324–330, 1960.

A. Ramanayake. Tests for a change point in the shape parameter of gamma random variables. *Communications in Statistics*, 33(4):821–833, 2004.

F. Ramsey and D. Schafer. *The Statistical Sleuth*. Duxbury/Thomson Learning, New York, 2002.

A. Rencher. *Methods of Multivariate Analysis*. John Wiley & Sons, Inc., 2002.

J. Romano. Bootstrap and randomization tests of some nonparametric hypotheses. *The Annals of Statistics*, 17(1):141–159, 1989.

D. Ruch and D. Claridge. A four-parameter change-point model for predicting energy consumption in commercial buildings. *Journal of Solar Energy Engineering*, 114(2): 77–83, 1992.

D. Ruch, L. Chen, J. Haberl, and D. Claridge. A change-point principal component analysis (CP/PCA) method for predicting energy usage in commercial buildings: The PCA model. *Journal of Solar Energy Engineering*, 115(2):77–84, 1993.

G. Ryan and S. Leadbetter. On the misuse of confidence intervals for two means in testing for the significance of the difference between the means. *Journal of Modern Applied Statistical Methods*, 1(2):473–478, 2002.

F. Satterthwaite. An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2(6):110–114, 1946.

N. Schenker and J. Gentleman. On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician*, 55(3):182–186, 2001.

A. Sen and M. Srivastava. On multivariate tests for detecting change in mean. *Sankhyā A*, 35(2):173–186, 1973.

A. Sen and M. Srivastava. On tests for detecting changes in means. *The Annals of Statistics*, 3(1):98–108, 1975.

W. Shewhart. *Statistical Method from the Viewpoint of Quality Control*. Washington, The Graduate School, The Department of Agriculture, 1939. ISBN 0-486-65232-7.

A. Smith. A Bayesian approach to inference about a change-point in a sequence of random variables. *Biometrika*, 62(2):407–416, 1975.

Y. Son and S. Kim. Bayesian single change point detection in a sequence of multivariate normal observations. *Statistics*, 39(5):373–387, 2005.

W. Venables and B. Ripley. *Modern Applied Statistics with S-Plus*. Springer, New York, 1997.

S. Vicente-Serrano and J. López-Moreno. Hydrological response to different time scales of climatological drought: an evaluation of the standardized precipitation index in a mountainous mediterranean basin. *Hydrology and Earth System Sciences*, 9(5):523–533, 2005.

T. Víšek. The likelihood ratio method for testing changes in the parameters of double exponential observations. *Biometrika*, 67(1):79–84, 1980.

L. Vostrikova. Detecting disorder in multidimensional random processes. *Soviet Mathematics Doklady*, 24:55–59, 1981.

L. Wang, D. Richardson, and P. Garrison. Linkages between nutrients and assemblages of macroinvertebrates and fish in wadeable streams: Implications to nutrient criteria development. *Environmental Management*, 39(2):194–212, 2007.

B. Weigel and D. Robertson. Identifying biotic integrity and water chemistry relations in nonwadeable rivers of Wisconsin: Toward the development of nutruient criteria. *Environmental Management*, 40(4):691–708, 2007.

K. Wernecke. Jacknife, bootstrap and cross-validation. An introduction to resampling methods. *Allgemeines Statistisches Archiv*, 77:32–59, 1993.

K. Worsley. Testing for a two-phase multiple regression. *Technometrics*, 25(1):35 – 42, 1983.

K. Zamba and D. Hawkins. A multivariate change-point model for statistical process control. *Technometrics*, 48(4):539–549, 2006.

C. Zou, Y. Liu, P. Qin, and Z. Wang. Empirical likelihood ratio test for the change-point problem. *Statistics & Probability Letters*, 77(4):374–382, 2007.