AN ABSTRACT OF THE DISSERTATION OF

Luna Sun for the degree of Doctor of Philosophy in Statistics presented on June 10, 2014.

Title: Statistical Methods for Serially Correlated Zero-inflated Proportions

Abstract approved: _____

Alix I Gitelman

Proportion data falling in the continuum (0, 1) are very common in practice. It can also happen that an inflated number of zeros (or ones) occur with proportion data. There are extensive studies of zero-inflated data in the literature. Almost all of them, however, focus on zero-inflated count data. Furthermore, because of sampling or experimental procedures, correlation commonly exists in data collected through time and/or space. The contribution of this research is to develop methods for analyzing zero-inflated proportion data with serial correlation.

We first propose multiple hypothesis tests for accessing homogeneity of two zero-inflated Beta populations under the assumption of independence of the observations. Fisher's method is adopted to combine independent likelihood ratio tests and asymptotic independent score tests to assess the equivalence of the populations. We also develop non-parametric and semi-parametric permutation-based tests for simultaneously com-

paring two or three features of unknown populations. In Chapter 3, we develop a Hidden Markov Model with zero-inflated Beta emission densities. We show that the standard EM algorithm for Hidden Markov Model parameter estimation can be applied in this case with emission distributions that are mixtures of discrete and continuous parts. In Chapter 4, we develop a generalized linear mixed model with an autoregressive random effect. This model involves the non-standard distribution (zero-inflated Beta) as well as components to account for the dependence among observations. We use Bayesian methodology for generalized linear mixed model parameter estimation and statistical inferences. We examine our methods by simulation and by analyzing a real marine science dataset where interest lies in distinguishing two serially correlated samples. We provide code for simultaneous hypothesis testing and for the Hidden Markov Model under open-source software R, as well as for generalized linear mixed model in WinBUGS.

Statistical Methods for Serially Correlated Zero-inflated Proportions

by

Luna Sun

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented June 10, 2014
Commencement June 2015

Doctor of Philosophy dissertation of Luna Sun presented on June 10, 2014.

APPROVED:

Major Professor, representing Statistics

Chair of the Department of Statistics

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Luna Sun, Author

ACKNOWLEDGEMENTS

I would like to take this special opportunity to express my deepest appreciation to many people I met in the last five years.

First and foremost, I want to thank to my Ph.D. advisor, Dr. Alix Gitelman, for guiding me these five years. She has taught me, both consciously and unconsciously, how to be a good statistician. I appreciate all her contributions of time, work and ideas to make my Ph.D. experience productive and enjoyable. I hope that I could be as lively, charming, insightful and open-minded as she is. I am also thankful for the excellent example she has provided as a successful woman scientist and educator.

I would like to thank Dr. Clifford Pereira, who is my former Ph.D. co-advisor and supervisor in the Statistical Consulting Laboratory in Oregon State University. He is one of the most enthusiastic, humorous, and knowledgeable people I know. Cliff gave me so many suggestions in work, school and life that I will keep them in my mind forever. The process of working with him is always enjoyable and fruitful. I am especially grateful to him for helping me find my real passion in statistics.

I also want to thank my other Ph.D.committee members, Dr. Lisa Madsen, Dr. Charlotte Wickham and Dr. Jeffrey Reimer for their time, interest and helpful suggestions in general.

I am very grateful to Dr. David Birkes. He was my primary resource for getting help in mathematical proofs. He is definitely the smartest person I know and he is always there willing to help with his iconic gentle smile on his face. I thank Dr. Sarah Emerson for having discussion with me about some technical details in my research. I also thank

Dr. Virginia Lesser and Joe Tyburczy for bringing and sharing the motivating dataset.

I thank all the faculty in Department of Statistics for teaching me how to dive and swim in the statistical ocean. It is my fortune to meet all of you.

I would like to thank Gu Mi for being my best friend at Oregon State University and sharing the very best five years with me. I am so glad we made it together. I want to thank my classmates in "Statistics 2011" in Oregon State University as well. I especially thank Evercita Eugenio for being such a cute and nice girl. Her homemade rum cake is always one of my favorite desserts. I also gratefully acknowledge Lu Wang and Rongrong Dong. It was my pleasure to share the same roof with you. I thank my friends at Oregon State University (too many to list here but you know who you are!) for your constant support, encouragement and help.

An special acknowledgment to Department of Statistics of Oregon State University. This is the longest I have ever stayed in a department. The lounge in basement, the computer lab, the library, the kitchen and my office, they are all the places that I spent thousands of hours in and they will all be in my dreams from now on. Everything happened here in the past five years made me who I am. And I am really glad to be the person who I want to be like this.

Lastly, I especially thank my dearest mom and dad. Their unconditional, continuous and unreserved love and care are my strongest support all along. There is no words in the world can convey how much I love them. Without them, I would not have made it this far.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF TABLES

LIST OF TABLES (Continued)

LIST OF TABLES (Continued)

Statistical Methods for Serially Correlated Zero-inflated Proportions

1 Introduction

This work is motivated by a marine science example in which the original question of interest is addressed by comparing two samples. Because of the sampling procedure, however, the observations within and across samples are correlated through time. In addition, there are some uncertainties in the actual determination of the two samples. The majority of observations are zero-inflated proportions, i.e., numbers in $[0, 1)$. The time series of observations were collected at unequally spaced sampling intervals. We provide more details about the motivating dataset in each chapter of the thesis with difference emphasis. We develop statistical methods to study different aspects of this dataset.

Proportion data falling in the continuum (0, 1) are very common in practice. Examples include the percentage of conifer cover in a particular area, the proportion of household income spent on food or the volume of stroke lesion as a percentage of total brain volume. The family of Beta distributions provides broad flexibility for modeling proportion data. Depending on its parameters, the Beta probability density function (pdf) can be right-skewed; left-skewed; "U-" or "J-" shaped; inverted "J-" shaped; or uniform (Ospina and Ferrari, 2012). In some cases, however, an inflated number of zeros and/or ones in a sample of proportions can render the Beta distribution an unsuitable model, since the Beta distribution takes support on the open interval (0, 1). There are extensive studies of zero-inflated data in the literature (Chiogna and Gaetan, 2007;

Barry and Welsh, 2002; Marin et al., 2005). Almost all of them, however, focus on zero-inflated count data (for example zero-inflated Poisson counts). The proportion data that we are considering here are not based on counts, however, but rather on continuous proportions across space and/or time. Ospina and Ferrari (2010) propose a mixed continuous-discrete distribution for data observed on the intervals [0, 1), (0, 1] or [0, 1]. The discrete component of this distribution is defined by a degenerate (point mass) distribution that assigns non-zero probability to 0 and/or 1 depending on whether there is zero- and/or one-inflation. In particular, the interest of our work, zero-inflated Beta (BEZI) has a point mass on zero. As is the Beta family of distributions, the BEZI family is quite flexible in shape.

In the first part of this work, assuming independence, we evaluate hypothesis tests aimed at distinguishing multiple features (i.e., parameters) of two BEZI population: We introduce several new parametric tests for such situations. We also develop non-parametric and semi-parametric tests as alternatives for simultaneously testing multiple features of un-specified populations. In many studies, interest focuses on comparing the distributions of samples or experimental units obtained under different conditions. For example, in an investigation of the proportion of conifer cover in forests, it might be interesting to know whether forests at higher elevation have generally different proportions of conifer than do forests at lower elevations. A two-sample comparison of the means for data such as these may not be sufficient for testing the equality of the two populations. It might be possible, for example, that two samples of forests at different elevations have very similar sample means (i.e., average proportions of conifer cover), but one sample has a larger variance or one sample has a higher proportion of zeros.

In cases like this, the samples may clearly come from two distinct populations, but if our focus is on the means only, these populations will be viewed as the same. Most existing work for testing the homogeneity of two populations with multiple parameters is focused on location and scale parameters. The strategy is first to test for the equality of scale parameters and once the equal scale assumption is found to be tenable, then to test for the equality of the location parameters. Surprisingly, there is not a well-known hypothesis test that considers the equalities of means and variances *at the same time*. Although the usual multiple parameter likelihood ratio test can be used, it appears to be rarely applied in practice. Beyond the likelihood ratio test there appear to be few existing tools that address the equality of location and scale parameters simultaneously, let alone tools for cases with more than just location and scale parameters.

In the second and third parts of this work, we focus on another aspect of the two-sample problem: serial correlation within samples. Dependent data are common in both social and natural science studies because of the methods of data collection. One of the most common dependence is serial dependence. Failing to account for this dependence may lead to misleading conclusion (Montgomery et al., 2008; Ramsey and Schafer, 2002) because researchers may think they have more independent pieces of information than they actually do. As discussed by Cox (1981), time series analysis for dependent data can be categorized as "parameter-driven" or "observational-driven". Parameter-driven models introduce the autocorrelation through a latent process. For example, in Hidden Markov Models (HMM) autocorrelation is determined by a hidden Markov process. Observation-driven models define the autocorrelation in the observations directly; i.e., the distribution of a variable, Y_t, is a function of previous observations, $Y_1, \cdots, Y_{t-1}$.

Some examples are autoregressive moving average models and Markov chain models. In this work, we study both types of models

Many serially dependent observations can be well-modeled using HMM in which observations are random variables whose probability distributions depend on the current state of an unobserved Markov chain (Rabiner et al., 1985). The mathematical theory of HMM was first developed by Baum and his colleagues in a series of classic papers in the late 1960s and early 1970s (Baum and Petrie, 1966; Baum and Eagon, 1967; Baum et al., 1970; Baum, 1972). The model is especially well known for its successful applications in speech recognition (Baker, 1975; Rabiner, 1989; Rabiner and Juang, 1993) and bioinformatics (Bishop and Thompson, 1986; Durbin et al., 1998; Krogh et al., 1994). Other applications include handwriting recognition (Rigoll et al., 1996), gesture recognition (Starner and Pentland, 1995), music score following (Pardo and Birmingham, 2005), image processing (Yamato et al., 1992), partial discharge image classification (Satish and Gururaj, 1993), finance (Mamon and Elliott, 2007), ecology (Baum and Eagon, 1967) and leaning behavior of live and artificial systems (Petrushin, 2000).

Two types of HMM have been extensively studied: Discrete Density Hidden Markov Model (DDHMM) and Continuous Density Hidden Markov Model (CDHMM). A DDHMM has observations chosen from a finite or countable set following discrete distributions, whereas a CDHMM has observations generated from continuous densities. For CDHMM, the most commonly used density is a mixture of Gaussian densities (Rabiner, 1989) because it can well-approximate many continuous density functions (Sorenson and Alspach, 1971). However, when the number of hidden states increases, the number

of parameters in these mixture models increases rapidly, which can cause a high computational load. The situation can become worse when the true emission density is far from Gaussian. Furthermore, when the number of observations is small, the parameters of the mixture density become non-identifiable, making the model infeasible in practice. An alternative is to use a parametric model tailored to the observed data. In Chapter 3, we use a Hidden Markov Model with zero-inflated Beta emission densities to model zero-inflated proportional data with serial correlation.

As an observation-based approach to addressing serial correlation in our BEZI data, we consider a generalized linear mixed model framework. Being a member of the exponential family (Ospina and Ferrari, 2010), BEZI is a candidate for the response distribution in generalized linear models (Nelder and Wedderburn, 1972). Ospina and Ferrari (2012) give a regression model for a general class of zero-or-one inflated beta distributions including BEZI, in which they link explanatory variables to all three parameters in the BEZI density through suitable link functions and assume the response variables are independently distributed after accounting for explanatory variables. For our situation, with serially correlated BEZI data, the generalized linear model is no longer appropriate. As an extension of generalized linear model (McCulloch et al., 2008), the generalized linear mixed model (GLMM) can accommodate not only non-normal responses but also autocorrelation among those responses. GLMM can be viewed as an extension of linear mixed model as well: compared with the generalized linear model, the GLMM adds normally distributed random effects to the linear predictor; compared with the linear mixed model, conditioned on realizations of random effects, the GLMM has non-normally distributed response variable. In Chapter 4, we develop a generalized linear mixed model

with zero-inflated Beta as the response distribution. We use a Bayesian approach for parameter estimation and other inferences. And we focus on the autoregressive random effect.

Our study fills the gap in modeling serially correlated zero-inflated continuous data with a constrained support (i.e., $[0, 1)$). Starting with an assumption of independence, in the first portion of this work we focus on comparing multiple features of zero-inflated proportion data from two populations. In the second and third portions, we are concerned about both the non-normal distribution feature and the correlation among observations. Specifically, the proportion data that we are considering here are not based on counts, but rather on continuous proportions across some unit of space and/or time. The correlation that we are concerned with comes from sampling data through time.

The organization of this dissertation is as follows. In Chapter 2, we propose multiple hypothesis tests for accessing the homogeneity of two populations. We first use Fisher's method to combine independent likelihood ratio tests and asymptotic independent score tests to assess the equivalence of two zero-inflated Beta populations. For each test, test statistics for the three individual parameters are combined into a single statistic to address the overall difference between the two populations. We also develop non-parametric and semi-parametric permutation-based tests for simultaneously comparing two or three features of unknown populations. In Chapter 3, we propose a Hidden Markov Model with zero-inflated Beta emission densities to model zero-inflated proportions with serial correlation. We show that the standard EM algorithm for Hidden Markov Model parameter estimation can be applied in this case with emission distributions that are mixtures of discrete and continuous parts. Also, we present accessible code

for Hidden Markov Model data generating, decoding and estimating under open-source software R. In Chapter 4, we develop a generalized linear mixed model with autoregressive random effect and use Bayesian estimates for inference. Zero-inflated Beta is used as the response distribution conditional on realization of an AR(1) structural random effect. WinBUGS and R code are provided for this model. We also provide some suggestions about choosing prior distributions and monitoring convergence of Markov Chain Monte Carlo. In each of Chapter 2, 3 and 4, we examine our methods both by simulation and analyzing the marine science dataset. A summary of our findings and a discussion about Hidden Markov Models and generalized linear mixed models are provided in Chapter 5, as well as some possibilities for future work.

2 Simultaneous Tests for Homogeneity of Two Zero-inflated (Beta) Populations

2.1 Abstract

Typical practice for testing homogeneity of two populations in terms of location and scale parameters is first to test the equality of the scale parameters and if that assumption is tenable then to test the equality of the location parameters. Few tools have been developed to evaluate both parameters simultaneously, let alone in cases with more than two parameters. Motivated by an example in marine science, we use Fisher's method to combine independent likelihood ratio tests and asymptotic independent score tests to assess the equivalence of two zero-inflated Beta populations (mixtures distributions with three parameters). For each test, test statistics for the three individual parameters are combined into a single statistic to address the overall difference between the two populations. We also develop non-parametric and semi-parametric permutation-based tests for simultaneously comparing two or three features of unknown populations. Simulations show that the likelihood-based tests perform well for large sample sizes and that the statistics based on combining likelihood ratio test statistics outperforms the ones based on combining score test statistics. The permutation-based tests have overall better performance in terms of both power and type I error rate. Our methods are easy to implement, computationally efficient and can be expanded to more than two populations

and to other multiple parameter families. The permutation tests are entirely generic and can be useful in various applications dealing with zero (or other) inflation.

Keywords: Simultaneous comparison, Fisher's method, Likelihood ratio test, Score test, Permutation test, Zero-inflated Beta

2.2 Introduction

Proportion data falling in the continuum (0, 1) are very common in practice. Examples include the proportion of conifer cover in a particular area, the proportion of household income spent on food and the proportion of weekly hours spent on work-related travel. The family of Beta distributions provides broad flexibility for modeling proportions. Depending on its parameters, the Beta probability density function (pdf) can be right-skewed; left-skewed; "U-" or "J-" shaped; inverted "J-" shaped; or uniform (Ospina and Ferrari, 2012). It can happen, however, that an inflated number of zeros and/or ones in a sample of proportions can render the Beta distribution an unsuitable model, since the Beta distribution takes support on the open interval (0, 1). There are extensive studies of zero-inflated data in the literature (Chiogna and Gaetan, 2007; Barry and Welsh, 2002; Marin et al., 2005). Almost all of them, however, focus on zero-inflated count data, for example zero-inflated Poisson counts. The proportion data that we are considering here are not based on counts, however, but rather on continuous proportions across space and/or time. Ospina and Ferrari (2010) propose a mixed continuous-discrete distribution for data observed on the intervals [0, 1), (0, 1] or [0, 1]. The discrete component of this distribution is defined by a degenerate (point mass) distribution that assigns non-zero

probability to 0 and/or 1 depending on whether there is zero- and/or one-inflation. In particular, the zero-inflated Beta (BEZI) has a point mass on zero.

Suppose $Y \sim BEZI(p, \mu, \phi)$. Then the pdf of Y is

$$f_Y(y) = \begin{cases} p & \text{if } y = 0, \\ (1-p)\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} & \text{if } y \in (0,1), \end{cases}$$

where $\Gamma(\cdot)$ is Gamma function, $0 < p < 1$, $0 < \mu < 1$ and $0 < \phi < \infty$. The mean and variance of Y are

$$E(y) = (1-p)\mu$$

and

$$Var(y) = (1-p)\frac{\mu(1-\mu)}{\phi+1} + p(1-p)\mu^2,$$

where μ and ϕ are the mean and precision parameter of Beta component. As is the Beta family of distributions, the BEZI family is quite flexible in shape.

In many studies, interest focuses on comparing the distributions of samples or experimental units obtained under different conditions. For example, in an investigation of the proportion of conifer cover in forests, it might be interesting to know whether forests at higher elevation have generally different proportions of conifer than do forests at lower elevations. A two-sample comparison of the means for data such as these may not be sufficient for testing the equality of the two populations. It might be possible, for example, that two samples of forests at different elevations have very similar sample means (i.e., average proportions of conifer cover), but one sample has a larger variance or one sample has a higher proportion of zeros. In cases like this, the samples may

clearly come from two distinct populations, but if our focus is on the means only, these populations will be viewed as the same. In this paper we evaluate hypothesis tests aimed at distinguishing multiple features (i.e., parameters) of two BEZI populations, and we introduce several new tests for such situations.

Most existing work for testing the homogeneity of two populations with multiple parameters is focused on location and scale parameters. The strategy is first to test for the equality of scale parameters and once the equal scale assumption is found to be tenable, then to test for the equality of the location parameters. In Normal populations, for example, when comparing two means there is usually an assumption about the variances: they could be known or unknown, equal or not equal, and a z-test or t-test can be applied to compare the means depending on the situation. Or, when comparing two variances we could assume the means are known or unknown, equal or not equal, and then use an F-test with appropriate degrees of freedom. Surprisingly, there is not a well-known hypothesis test that considers the equality of means and variances *at the same time*. Although the usual multiple parameter likelihood ratio test (LRT) can be used, it appears to be rarely applied in practice. Beyond the LRT there appear to be few existing tools that address the equality of location and scale parameters simultaneously, let alone tools for cases with more than just location and scale parameters.

Singh (1986) develops a test statistic for testing the equality of means and variances of two Normal populations based on combining two independent likelihood ratio tests: one for testing the equality of means given the variances are equal but unspecified and the other for testing the equality of variances when the means are unspecified and not necessarily equal. In this paper, Singh claims that the test is asymptotically optimal

in the sense of Bahadur efficiency based on facts derived by Littell and Folks (1971, 1973). The idea of combining tests originated with Fisher (1950), though the detailed steps for actually doing the combining were omitted there. Durairajan (1985) also applies Fisher's idea to test the parameters of the inverse Gaussian distribution based on combining two independent likelihood ratio tests.

In an extension of Singh's work, Paul and Jiang (2005) expand the method from Normal distributions to several other two-parameter distributions including the Negative Binomial and Beta-Binomial distributions. They also consider combining asymptotic independent score test statistics. Their simulations provide evidence that Fisher's method performs well even where there is only asymptotic independence. The authors recommend statistics based on combining score tests because the approach appears to maintain the appropriate level of the test in all the situations they investigate. Thiagarajah (2012) considers a combined test procedure for testing the homogeneity of Weibull (or Extreme value) populations with censored data following ideas similar to Singh (1986) and Paul and Jiang (2005). Simulation shows that Fisher's method of combining statistics performs reasonably well under censoring.

In this paper, we adopt Fisher's method to combine the p-values of independent likelihood ratio tests for the three parameters of the BEZI distribution. We also combine asymptotically independent score tests for these three parameters. We discuss the inconsistency of some versions of the score test and evaluate several non-parametric tests often applied in situations of non-Normality. In addition to these standard parametric tests, we develop two new non-parametric and two new semi-parametric permutation tests for testing multiple features of the populations simultaneously. One of the non-parametric

tests considers only location and scale, and the other considers the zero proportion in addition to location and scale. The two semi-parametric tests can be viewed as hurdle (Ridout et al., 1998) models—we first model the presence/absence of zeros, and then we model the non-zeros. Although motived by data from BEZI populations, the non-parametric and semi-parametric approaches are applicable to other inflated populations.

The organization of this paper is as follows. In Section 2, we describe an example from marine science where in the original analysis, non-parametric tests were applied to distinguish two populations, where these non-parametric tests typically look at only one feature of the populations. In this example. the underlying data populations are well-represented by the BEZI distribution. In Section 3, we introduce the parametric simultaneous tests that we developed for assessing the equivalence of the two populations. Then we discuss two non-parametric and two semi-parametric permutation tests as alternatives. In Section 4, we give simulation results for evaluating the performance of our new tests and compare them with standard parametric and non-parametric alternatives. In Section 5, all tests are applied to the marine science data and compared. We discuss our findings and possible extension of this work in Section 6.

2.3 Settlement of Onshore Barnacle Larvae

Tyburczy (2011) compares settlement distributions of onshore barnacle larvae with and without the occurrence of different oceanographic processes. Such processes as large-scale regional relaxation or upwelling and smaller-scale localized diurnal upwelling driven by afternoon sea breezes are thought to influence barnacle settlement. The data

consist of observations on two types of barnacle larva (*Balanus* cf.*glandula* and *Chthamalus* spp.) daily or bi-daily settlement information, and covariate information involving the physical processes (regional relaxation, front passage, and diurnal upwelling) in four study areas (Sandhill Bluff, Terrace Point, Bonny Doon Beach and Lighthouse Point) collected within northern Monterey Bay, CA in 2007 (May-Sept). As described in Tyburczy (2011), the settlement data on intertidal plates was normalized for hours of immersion based on tidal height of the plates and combined with pump samples and larvae counts on larval traps deployed with different depths. The manipulation process compressed the settlement information for *Balanus* cf.*glandula* to the range [0, 1), with more than half of them being zeros. Examples of the barnacle larvea settlement data are shown in Figure 2.1, where all panels are for data from Sandhill Bluff. Each panel contains overlapping histograms of settlement data with and without the occurrence of a particular oceanographic process and the settlement data are the same in these panels. Notice that the two zero counts are also indicated in each panel.

Tyburczy studied the associations between the occurrence of the oceanographic processes and the larvae settlement at each location. His analysis treated the settlement during and between occurrences of the oceanographic process as two samples which were then compared. Analysis of the association between the oceanographic processes and settlement is certainly complicated by non-normality and autocorrelation in the data, as well as uneven sampling intervals. Tyburczy used Superposed Epoch Analysis (Chree, 1913, 1914) on log-transformed settlement data (after adding a small constant) with a modified t-test statistic and the difference between sample means as permutation statistics. These test statistics are $SEA = (\bar{E} - \bar{B})/S$ and $SEAD = \bar{E} - \bar{B}$, where $\bar{E}$ and $\bar{B}$ are

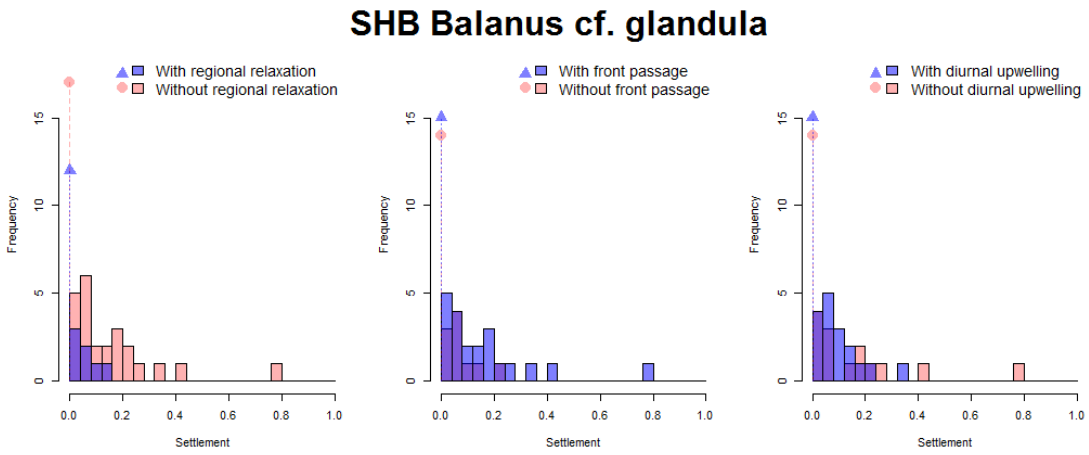


Figure 2.1: Histogram of settlement for *Balanus* cf.*glandula* in Sandhill Bluff (SHB) with and without physical processes (regional relaxation, front passage and diurnal upwelling; left to right).

the sample mean numbers of settlement during the periods when the process did and did not occur, respectively, and S is a modified pooled estimate of standard deviation. Tyburczy also used the Wilcoxon rank-sum test to compare settlement during intervals when each process occurred and those when it did not. After describing our new tests, we will compare them against the tests Tyburczy used by using simulations. We will also compare all tests as they apply to the settlement data.

2.4 Test Statistics for Simultaneous Tests

2.4.1 Fisher's Method Based on Combining Tests

Fisher's method, also known as Fisher's combined probability tests, was initially suggested by Fisher (1950). It combines p-values from several independent tests into one test statistic using the formula

$$X^2 = -2\sum_{i=1}^{k} ln(p_i), \tag{2.1}$$

where p_i is the p-value for the i^{th} hypothesis test. When the p-values tend to be small, X^2 will be large, which suggests that the null hypotheses are not true for every test (Paul and Jiang, 2005). When all the null hypotheses are true and the p_i's (more essentially, their corresponding test statistics) are independent, X^2 will have a χ^2_{2k} distribution where k is the number of tests being combined. This fact can be used to determine the p-value corresponding to X^2.

The key steps for deriving a simultaneous test via Fisher's method are:

1. Partition the general hypotheses into several sub-hypotheses
2. Obtain test statistics for each sub-hypotheses
3. Show the test statistics in 2 are independent (or asymptotically independent)
4. Use expression (2.1) to combine the multiple tests to obtain the test statistic for the general (or simultaneous) hypotheses

We consider $Y_{11}, Y_{12}, \cdots, Y_{1n_1} \sim BEZI(p_1, \mu_1, \phi_1)$ and $Y_{21}, Y_{22}, \cdots, Y_{2n_2} \sim BEZI(p_2, \mu_2, \phi_2)$, where all observations are mutually independent both within and across samples. The general hypotheses we investigate are

$$H_0 : p_1 = p_2; \mu_1 = \mu_2; \phi_1 = \phi_2 \text{ vs } H_1 : \text{At least one equality fails} \tag{2.2}$$

Because the reduced model is nested in the full model, these general hypotheses can be tested using several tests: a standard likelihood ratio test (denoted *LRT* in the simulations and comparisons that follow in Section 2.4.1.1 and Section 2.5), a score test based on maximum likelihood estimates (MLE) under the reduced model ($\hat{S}$ in what follows) or a score test based on MLE under the full model ($\tilde{S}$ in what follows). These three approaches, while reasonably straightforward, do not allow for making direct inference about individual BEZI parameters separately.

The general hypotheses for the two-sample BEZI setting as defined in (2.2) can be partitioned into three sub-hypotheses:

$$H_0^{'} : \mu_1 = \mu_2; \phi_1 = \phi_2; p_1 = p_2 \quad \text{vs} \quad H_1^{'} : \mu_1 = \mu_2; \phi_1 = \phi_2; p_1 \neq p_2 \tag{2.3}$$

$$H_0^{''} : \mu_1 = \mu_2; \phi_1 = \phi_2; p_1 \text{ and } p_2 \text{ arbitrary} \quad \text{vs} \quad H_1^{''} : \mu_1 = \mu_2; \phi_1 \neq \phi_2; p_1 \text{ and } p_2 \text{ arbitrary} \tag{2.4}$$

$$H_0^{'''} : \mu_1 = \mu_2; \phi_1, \phi_2, p_1 \text{ and } p_2 \text{ arbitrary} \quad \text{vs} \quad H_1^{'''} : \mu_1 \neq \mu_2; \phi_1, \phi_2, p_1 \text{ and } p_2 \text{ arbitrary} \tag{2.5}$$

Notice that the null parameter space (i.e., the parameter space under the null hypothesis) for (2.3) is the same as the null parameter space for (2.2); the null parameter space for (2.4) is the same as the entire parameter space (union of the parameter space under the null and the alternative hypotheses) for (2.3); the null parameter space for (2.5) is

the same as the entire parameter space for (2.4); the entire parameter space for (2.5) is the same as that for (2.2). In other words, the parameter spaces have a nested structure. There are other ways to partition the general hypotheses. The ones given here, however, address equality of the different BEZI parameters. Specifically, (2.3) addresses the equality of the proportions of zero and keeps the Beta components equal. Then, the scale parameter (ϕ) of the Beta component is evaluated in (2.4) and finally, the location parameter (μ) in the Beta component in (2.5).

2.4.1.1 Likelihood Ratio Tests

To use Fisher's method of combining independent tests, we need the likelihood ratio test statistics for the sub-hypotheses as defined in (2.3), (2.4) and (2.5). Given two samples from BEZI populations ($BEZI(p_1,\mu_1,\phi_1)$ and $BEZI(p_2,\mu_2,\phi_2)$), the log-likelihood function under each sub-hypothesis can be easily written. Let $\hat{l}_0$, $\hat{l}_1$ and $\hat{l}_2$ denote these log-likelihoods evaluated at the MLE under $H_0^{'}$, $H_0^{''}$, $H_0^{'''}$, respectively; and let $\hat{l}$ denote the log-likelihood evaluated at the MLE under $H_1^{'''}$. Because of the nested relationship of the parameter spaces, the corresponding likelihood ratio test statistics for (2.3), (2.4) and (2.5) are $LRT1 = 2(\hat{l}_1 - \hat{l}_0)$, $LRT2 = 2(\hat{l}_2 - \hat{l}_1)$, and $LRT3 = 2(\hat{l} - \hat{l}_2)$ respectively. When $H_0^{'}$, $H_0^{''}$, $H_0^{'''}$ are true, $LRT1$, $LRT2$ and $LRT3$ are asymptotically distributed as χ_1^2. Notice that $LRT1 + LRT2 + LRT3 = LRT$, where LRT is likelihood ratio test statistic for hypotheses (2.2). The degrees of freedom for the χ^2 distributions have a corresponding equality. Similar to Paul and Jiang (2005), and according to Cochran's Theorem (Cramer, 1946), $LRT1$, $LRT2$ and $LRT3$ are mutually independent.

In another approach, we combine three independent likelihood ratio tests using Fisher's method. If we let $L_1(t_1) = P(LRT1 > t_1 \mid H_0^{'})$, $L_2(t_2) = P(LRT2 > t_2 \mid H_0^{''})$ and $L_3(t_3) = P(LRT3 > t_3 \mid H_0^{'''})$, and let $LRT1_{obs}$, $LRT2_{obs}$ and $LRT3_{obs}$ denote the observed value of $LRT1$, $LRT2$ and $LRT3$, respectively, then

$$MLR = -2ln(L_1(LRT1_{obs}) \cdot L_2(LRT2_{obs}) \cdot L_3(LRT3_{obs})) \sim \chi_6^2 \text{ under } H_0 \text{ in (2.2).}$$

The decision rule will reject H_0 in (2.2) at level α if $MLR \geq \chi_6^2(1-\alpha)$ with $\chi_6^2(1-\alpha)$ as the $(1-\alpha)\%$ quantile of χ_6^2.

2.4.1.2 Score Tests

As another likelihood-based large sample tool, the score test is particularly appealing because it only requires the estimates of parameters under the null hypothesis (i.e., for the reduced model only). Since the BEZI is a full-rank exponential family (Ospina and Ferrari, 2010), it satisfies the regularity condition needed for the score test. For notational simplicity, we re-parameterize p_1, p_2, μ_1, μ_2, ϕ_1 and ϕ_2 taking $p = p_2$, $\delta = p_1 - p_2$, $\mu = \mu_2$, $\beta = \mu_1 - \mu_2$, $\phi = \phi_2$ and $\gamma = \phi_1 - \phi_2$. The corresponding new general hypotheses are:

$$H_0 : \delta = 0; \beta = 0; \gamma = 0 \text{ vs } H_1 : \text{At least one of } \delta, \beta \text{ and } \gamma \text{ differs from } 0 \qquad (2.6)$$

Let ℓ_0 and ℓ_f denote the log-likelihood under the reduced and full models in (2.6). Then the first and second partial derivatives of ℓ_f with respect to $(p, \delta, \mu, \beta, \phi, \gamma)$ can be

obtained. Define $s = \begin{bmatrix} \frac{\partial \ell_f}{\partial \delta} & \frac{\partial \ell_f}{\partial \beta} & \frac{\partial \ell_f}{\partial \gamma} \end{bmatrix}$ and

$$A = E_{H_0} \begin{bmatrix} -\frac{\partial^2 \ell_f}{\partial \delta^2} & -\frac{\partial^2 \ell_f}{\partial \beta \partial \delta} & -\frac{\partial^2 \ell_f}{\partial \gamma \partial \delta} \\ -\frac{\partial^2 \ell_f}{\partial \delta \partial \beta} & -\frac{\partial^2 \ell_f}{\partial \beta^2} & -\frac{\partial^2 \ell_f}{\partial \gamma \partial \beta} \\ -\frac{\partial^2 \ell_f}{\partial \delta \partial \gamma} & -\frac{\partial^2 \ell_f}{\partial \beta \partial \gamma} & -\frac{\partial^2 \ell_f}{\partial \gamma^2} \end{bmatrix}, \; C = E_{H_0} \begin{bmatrix} -\frac{\partial^2 \ell_f}{\partial p \partial \delta} & -\frac{\partial^2 \ell_f}{\partial \mu \partial \delta} & -\frac{\partial^2 \ell_f}{\partial \phi \partial \delta} \\ -\frac{\partial^2 \ell_f}{\partial p \partial \beta} & -\frac{\partial^2 \ell_f}{\partial \mu \partial \beta} & -\frac{\partial^2 \ell_f}{\partial \phi \partial \beta} \\ -\frac{\partial^2 \ell_f}{\partial p \partial \gamma} & -\frac{\partial^2 \ell_f}{\partial \mu \partial \gamma} & -\frac{\partial^2 \ell_f}{\partial \phi \partial \gamma} \end{bmatrix},$$

$$D = E_{H_0} \begin{bmatrix} -\frac{\partial^2 \ell_f}{\partial p^2} & -\frac{\partial^2 \ell_f}{\partial \mu \partial p} & -\frac{\partial^2 \ell_f}{\partial \phi \partial p} \\ -\frac{\partial^2 \ell_f}{\partial p \partial \mu} & -\frac{\partial^2 \ell_f}{\partial \mu^2} & -\frac{\partial^2 \ell_f}{\partial \phi \partial \mu} \\ -\frac{\partial^2 \ell_f}{\partial p \partial \phi} & -\frac{\partial^2 \ell_f}{\partial \mu \partial \phi} & -\frac{\partial^2 \ell_f}{\partial \phi^2} \end{bmatrix}.$$

Let $\hat{s} = s|_{(\delta=0,\beta=0,\gamma=0,\hat{p_0},\hat{\mu_0},\hat{\phi_0})}$ with $(\hat{p_0},\hat{\mu_0},\hat{\phi_0})$ the MLE under H_0 in (2.6). Using the formula for blockwise inversion of matrices, the score test statistics involving the nuisance parameters for testing (2.6) is $S = \hat{s}^T(A - CD^{-1}C^T)^{-1}\hat{s}$. Under H_0 in (2.6), S is asymptotically distributed as χ^2_3 (Pace and Salvan, 1997). Since A, C, D involve the expected information matrix, they may be hard to derive in explicit forms. Others (Mukhopadhyay, 2000) have used the observed (or empirical) information matrix to simplify the problem so that

$$\hat{A}_{(0,0,0,\hat{p_0},\hat{\mu_0},\hat{\phi_0})} = \begin{bmatrix} -\frac{\partial^2 \ell_f}{\partial \delta^2} & -\frac{\partial^2 \ell_f}{\partial \beta \partial \delta} & -\frac{\partial^2 \ell_f}{\partial \gamma \partial \delta} \\ -\frac{\partial^2 \ell_f}{\partial \delta \partial \beta} & -\frac{\partial^2 \ell_f}{\partial \beta^2} & -\frac{\partial^2 \ell_f}{\partial \gamma \partial \beta} \\ -\frac{\partial^2 \ell_f}{\partial \delta \partial \gamma} & -\frac{\partial^2 \ell_f}{\partial \beta \partial \gamma} & -\frac{\partial^2 \ell_f}{\partial \gamma^2} \end{bmatrix}, \; \hat{C}_{(0,0,0,\hat{p_0},\hat{\mu_0},\hat{\phi_0})} = \begin{bmatrix} -\frac{\partial^2 \ell_f}{\partial p \partial \delta} & -\frac{\partial^2 \ell_f}{\partial \mu \partial \delta} & -\frac{\partial^2 \ell_f}{\partial \phi \partial \delta} \\ -\frac{\partial^2 \ell_f}{\partial p \partial \beta} & -\frac{\partial^2 \ell_f}{\partial \mu \partial \beta} & -\frac{\partial^2 \ell_f}{\partial \phi \partial \beta} \\ -\frac{\partial^2 \ell_f}{\partial p \partial \gamma} & -\frac{\partial^2 \ell_f}{\partial \mu \partial \gamma} & -\frac{\partial^2 \ell_f}{\partial \phi \partial \gamma} \end{bmatrix},$$

$$\hat{D}_{(0,0,0,\hat{p_0},\hat{\mu_0},\hat{\phi_0})} = \begin{bmatrix} -\frac{\partial^2 \ell_f}{\partial p^2} & -\frac{\partial^2 \ell_f}{\partial \mu \partial p} & -\frac{\partial^2 \ell_f}{\partial \phi \partial p} \\ -\frac{\partial^2 \ell_f}{\partial p \partial \mu} & -\frac{\partial^2 \ell_f}{\partial \mu^2} & -\frac{\partial^2 \ell_f}{\partial \phi \partial \mu} \\ -\frac{\partial^2 \ell_f}{\partial p \partial \phi} & -\frac{\partial^2 \ell_f}{\partial \mu \partial \phi} & -\frac{\partial^2 \ell_f}{\partial \phi^2} \end{bmatrix}.$$

Then $\hat{S} = \hat{s}^T(\hat{A} - \hat{C}\hat{D}^{-1}\hat{C}^T)^{-1}\hat{s}$ becomes the score test statistic we typically use. Asymptotically, its distribution is χ^2_3 (Mukhopadhyay, 2000).

Because the observed information matrix is not always positive definite, it can happen that the observed $\hat{S}$ is negative, and so, not consistent for the expected information (Verbeke and Molenberghs, 2007). As a consequence, the p-value for hypothesis testing turns out to be 1. Some argue that the negative score statistics implies a rejection of the null hypothesis as the data fit poorly under the reduced model (Morgan et al., 2007). Another way to deal with the negative score statistics, though, is to consider the MLE for $(p, \delta, \mu, \beta, \phi, \gamma)$ from the full model, denoted as $(\tilde{p}, \tilde{\delta}, \tilde{\mu}, \tilde{\beta}, \tilde{\phi}, \tilde{\gamma})$, rather than the MLE of (p, μ, ϕ) from the reduced model. Then the corresponding estimators for A, C, D would be $\tilde{A}, \tilde{C}, \tilde{D}$. A new score test statistic $\tilde{S} = \hat{s}^T(\tilde{A} - \tilde{C}\tilde{D}^{-1}\tilde{C}^T)^{-1}\hat{s}$ can be constructed, and it is also asymptotically χ^2_3 distributed (Conniffe, 2001; Freedman, 2007).

Following Paul and Jiang (2005), in addition to the standard score test, we also consider Fisher's method-based on score test statistics. As in the ordinary score test case, the sub-hypotheses can be re-parameterized as $H_0^{'} : \delta = 0$ *vs* $H_1^{'} : \delta = 0$ with four-dimensional parameter space (p, μ, ϕ are nuisance parameters); $H_0^{''} : \gamma = 0$ *vs* $H_1^{''} : \gamma = 0$ with a five-dimensional parameter space (p, δ, μ, ϕ are nuisance parameters); $H_0^{'''} : \beta = 0$ *vs* $H_1^{'''} : \beta = 0$ with six-dimensional parameter space ($p, \delta, \mu, \gamma, \phi$ are nuisance parameters). The score test statistics for these hypotheses can be obtained following the same steps as before. Under the same pattern of notation, we have $\hat{S_1}$, $\hat{S_2}$ and $\hat{S_3}$ (and also $\tilde{S_1}$, $\tilde{S_2}$ and $\tilde{S_3}$) as test statistics for the three hypotheses, respectively. Each of these follows a χ^2_1 distribution under its respective null hypothesis. To use Fisher's method

for combining p-values, we need to show the independence between $\hat{S}_1$, $\hat{S}_2$ and $\hat{S}_3$ ($\tilde{S}_1$, $\tilde{S}_2$ and $\tilde{S}_3$). No exact result exists; however, we show the asymptotic independence of these statistics in Appendix (**Theorem 1**).

Taking $\widehat{L_1(t_1)} = P(\hat{S}_1 > t_1 \mid H_0^{'})$, $\widehat{L_2(t_2)} = P(\hat{S}_2 > t_2 \mid H_0^{''})$ and $\widehat{L_3(t_3)} = P(\hat{S}_3 > t_3 \mid H_0^{'''})$, and letting $\hat{S}_{1obs}$, $\hat{S}_{2obs}$, $\hat{S}_{3obs}$ denote the observed value of $\hat{S}_1$, $\hat{S}_2$, $\hat{S}_3$, we have

$$\widehat{MScore} = -2ln(\widehat{L_1(\hat{S}_{1obs})} \cdot \widehat{L_2(\hat{S}_{2obs})} \cdot \widehat{L_3(\hat{S}_{3obs})}) \sim \chi_6^2 \text{ under } H_0 \text{ in (2.6).}$$

The decision rule rejects H_0 in (2.6) at level α if $\widehat{MScore} \geq \chi_6^2(1-\alpha)$.

Similarly, let $\widetilde{L_1(t_1)} = P(\tilde{S}_1 > t_1 \mid H_0^{'})$, $\widetilde{L_2(t_2)} = P(\tilde{S}_2 > t_2 \mid H_0^{''})$ and $\widetilde{L_3(t_3)} = P(\tilde{S}_3 > t_3 \mid H_0^{'''})$, and let $\tilde{S}_{1obs}$, $\tilde{S}_{2obs}$, $\tilde{S}_{3obs}$ denote the observed value of $\tilde{S}_1$, $\tilde{S}_2$, $\tilde{S}_3$ then

$$\widetilde{MScore} = -2ln(\widetilde{L_1(\tilde{S}_{1obs})} \cdot \widetilde{L_2(\tilde{S}_{2obs})} \cdot \widetilde{L_3(\tilde{S}_{3obs})}) \sim \chi_6^2 \text{ under } H_0 \text{ in (2.6).}$$

The decision rule rejects H_0 in (2.6) at level α if $\widetilde{MScore} \geq \chi_6^2(1-\alpha)$.

2.4.2 Non-parametric Simultaneous Tests

In this section, we develop two non-parametric simultaneous tests based on permutations to address the question of homogeneity between two BEZI populations. Although we mainly discuss BEZI populations, the application of these new tests is not limited to the BEZI. Our idea is to transform two or three features of the data into one dimensional tests using squared Mahalanobis distance (Mahalanobis, 1936). Compared with Euclidean distance, Mahalanobis distance takes into account covariance among vari-

ables and is scale invariant. We use squared Mahalanobis distance because the square transformation is monotone for non-negative values, and in general, taking the square root is computationally inefficient.

2.4.2.1 Test Based on Location and Scale

We consider $Y_{11},\cdots,Y_{1n_1} \sim f_1$ and $Y_{21},\cdots,Y_{2n_2} \sim f_2$, where f_1 and f_2 come from the same distributional family whose first and second moments exist and are finite. We assume that all the observations are mutually independent within and across samples. Suppose $E(Y_{1i}) = \nu_1$, $Var(Y_{1i}) = \sigma_1^2$, $E(Y_{2i}) = \nu_2$ and $Var(Y_{2i}) = \sigma_2^2$, the hypotheses we investigate are

$$H_0 : f_1 = f_2 \text{ vs } H_1 : f_1 = f_2 \tag{2.7}$$

If BEZI is the underlying distributional family, hypotheses (2.7) are equivalent to hypotheses (2.2) and (2.6).

Given the observations, let $\bar{y}_{1obs}, \bar{y}_{2obs}, s^2_{1obs}, s^2_{2obs}$ denote the respective sample means and variances. We compute

$$D_{obs} = |\bar{y}_{1obs} - \bar{y}_{2obs}| \text{ and } R_{obs} = |log(s^2_{1obs}) - log(s^2_{2obs})|.$$

Next, we permute $Y_{11},\cdots,Y_{1n_1},Y_{21},\cdots,Y_{2n_2}$ p times. For the i^{th} permutation, we essentially re-label Y_{11}, $\cdots$, Y_{1n_1}, $Y_{21},\cdots,Y_{2n_2}$ as $Y_{11(i)}$, $\cdots$, $Y_{1n_{1(i)}}$, $Y_{21(i)}$, $\cdots$, $Y_{2n_{2(i)}}$. For

$i = 1, \cdots, p$, we compute

$$D_{(i)} = |\bar{y}_{1(i)} - \bar{y}_{2(i)}| \text{ and } R_{(i)} = |log(s^2_{1(i)}) - log(s^2_{2(i)})|.$$

Define $\mathbf{D} = \left[D_{obs}\ D_{(1)}\ \cdots\ D_{(p)}\right]^T$ and $\mathbf{R} = \left[R_{obs}\ R_{(1)}\ \cdots\ R_{(p)}\right]^T$, let $\bar{D} = \frac{1}{p+1}\sum_{l=1}^{p+1} D_l$, $\bar{R} = \frac{1}{p+1}\sum_{l=1}^{p+1} R_l$ and $\Sigma = Cov(\mathbf{D}, \mathbf{R})$.

For $l = 1, \cdots, p+1$, let

$$T_l = \begin{bmatrix} D_l - \bar{D} & R_l - \bar{R} \end{bmatrix} \Sigma^{-1} \begin{bmatrix} D_l - \bar{D} & R_l - \bar{R} \end{bmatrix}^T.$$

The decision rule will reject H_0 in (2.7) at level α if T_1 (i.e., T_{obs}) $\geq \mathbf{T}_{1-\alpha}$, the $(1-\alpha)$th quantile of $\mathbf{T} = (T_1, \cdots, T_{p+1})^T$.

2.4.2.2 Test Based on Zero Proportion, Location and Scale

Under the same setting as in Section 2.4.2.1, in addition to ν_1, σ_1^2, ν_2 and σ_2^2, we define p_1 and p_2 as zero proportions in the two populations, respectively. Given samples from two populations, we have $\hat{p}_{1obs}$, $\hat{p}_{2obs}$ $\bar{y}_{1obs}$, $\bar{y}_{2obs}$, s^2_{1obs} and s^2_{2obs} as the respective sample zero proportions, means and variances. We compute

$$Q_{obs} = |\hat{p}_{1obs} - \hat{p}_{2obs}|, D_{obs} = |\bar{y}_{1obs} - \bar{y}_{2obs}| \text{ and } R_{obs} = |log(s^2_{1obs}) - log(s^2_{2obs})|.$$

As before we permute $Y_{11},\cdots,Y_{1n_1},Y_{21},\cdots,Y_{2n_2}$ p times, and for $i=1,\cdots,p$, we compute

$$Q_{(i)}=|\hat{p}_{1(i)}-\hat{p}_{2(i)}|, D_{(i)}=|\bar{y}_{1(i)}-\bar{y}_{2(i)}| \text{ and } R_{(i)}=|log(s^2_{1(i)})-log(s^2_{2(i)})|.$$

Then define

$$\mathbf{Q}=\begin{array}{cccc} Q_{obs} & Q_{(1)} & \cdots & Q_{(p)} \end{array}^T, \mathbf{D}=\begin{array}{cccc} D_{obs} & D_{(1)} & \cdots & D_{(p)} \end{array}^T \text{ and } \mathbf{R}=\begin{array}{cccc} R_{obs} & R_{(1)} & \cdots & R_{(p)} \end{array}^T,$$

and let $\bar{Q}=\frac{1}{p+1}\sum_{l=1}^{p+1}Q_l$, $\bar{D}=\frac{1}{p+1}\sum_{l=1}^{p+1}D_l$, $\bar{R}=\frac{1}{p+1}\sum_{l=1}^{p+1}R_l$ and $\Sigma^*=Cov(\mathbf{Q},\mathbf{D},\mathbf{R})$.

For $l=1,\cdots,p+1$, let

$$V_l=\begin{array}{ccc} Q_l-\bar{Q} & D_l-\bar{D} & R_l-\bar{R} \end{array}\ \Sigma^{*-1}\ \begin{array}{ccc} Q_l-\bar{Q} & D_l-\bar{D} & R_l-\bar{R} \end{array}^T$$

The decision rule will reject H_0 in (2.7) at level α if V_1 (i.e., V_{obs}) $\geq \mathbf{V}_{1-\alpha}$, the $(1-\alpha)$th quantile of $\mathbf{V}=(V_1,\cdots,V_{p+1})^T$.

2.4.3 Semi-parametric Simultaneous Tests

In this section, we develop two semi-parametric tests to compare two zero-inflated populations. Both tests can be viewed as two-stage approaches in which we first analyze the data for presence/absence of zeros, and we then analyze the non-zeros (Liu and Chan, 2008). The proportion of zeros is controlled by a parameter $p\in(0,1)$, whereas the non-zeros are characterized by a density function, $h_Y(y)$. The only requirement is that $h_Y(y)$

has finite second moment. We don't assume any particular functional form for $h_Y(y)$. This idea is similar to that of hurdle models (Ridout et al., 1998). For example, in the Poisson hurdle model, $h_Y(y)$ is zero-truncated Poisson density (Zeileis et al., 2008). $h_Y(y)$ is the non-parametric component; p is the parametric component. Therefore the tests we have in this section are semi-parametric methods.

2.4.3.1 Test Based on Zero Proportion, Location and Scale of Non-zero Component

Consider a zero-inflated population with pdf $f_Y(y)$ such that

$$f_Y(y) = \begin{cases} p & \text{if } y = 0, \\ (1-p)h_Y(y) & \text{if } y \in (0,1), \end{cases}$$

where $E(h_Y(y)) = \eta$ and $Var(h_Y(y)) = \omega^2$. We suppose $Y_{11}, \cdots, Y_{1n_1} \sim f_{Y_1}(y)$ with $(p_1, \eta_1, \omega_1^2)$ and $Y_{21}, \cdots, Y_{2n_2} \sim f_{Y_2}(y)$ with $(p_2, \eta_2, \omega_2^2)$. Assuming independence, the hypotheses we are interested in are

$$H_0 : f_{Y_1} = f_{Y_2} \text{ vs } H_1 : f_{Y_1} = f_{Y_2}.$$

In other words,

$$H_0 : p_1 = p_2 \text{ and } h_{Y_1} = h_{Y_2} \text{ vs } H_1 : \text{ Either equality fails.} \tag{2.8}$$

We use $\hat{p}_1$, $\hat{p}_2$, $\tilde{y}_1$, $\tilde{y}_2$, $\tilde{s}_1^2$, $\tilde{s}_2^2$ as estimates of p_1, p_2, η_1, η_2, ω_1^2 and ω_2^2. Specifically, for $i = 1$ and 2,

$$\hat{p}_i = \frac{\sum_{j=1}^{n_i} I_{(y_{ij}=0)}}{n_i}, \tilde{y}_i = \frac{\sum_{j=1}^{n_i} y_{ij} I_{(y_{ij}>0)}}{\sum_{j=1}^{n_i} I_{(y_{ij}>0)}} \text{ and } \tilde{s}_i^2 = \frac{\sum_{j=1}^{n_i} (y_{ij} - \tilde{y}_i)^2 I_{(y_{ij}>0)}}{\sum_{j=1}^{n_i} I_{(y_{ij}>0)} - 1}.$$

Then we compute

$$Q_{obs} = |\hat{p}_{1obs} - \hat{p}_{2obs}|, \tilde{D}_{obs} = |\tilde{y}_{1obs} - \tilde{y}_{2obs}| \text{ and } \tilde{R}_{obs} = |log(\tilde{s}_{1obs}^2) - log(\tilde{s}_{2obs}^2)|.$$

As before, we permute $Y_{11}, \cdots, Y_{1n_1}, Y_{21}, \cdots, Y_{2n_2}$ p times. And for $i = 1, \cdots, p$, calculate

$$Q_{(i)} = |\hat{p}_{1(i)} - \hat{p}_{2(i)}|, \tilde{D}_{(i)} = |\tilde{y}_{1(i)} - \tilde{y}_{2(i)}| \text{ and } \tilde{R}_{(i)} = |log(\tilde{s}_{1(i)}^2) - log(\tilde{s}_{2(i)}^2)|.$$

Then define

$$\mathbf{Q} = \begin{array}{cccc} Q_{obs} & Q_{(1)} & \cdots & Q_{(p)} \end{array}^T, \tilde{\mathbf{D}} = \begin{array}{cccc} \tilde{D}_{obs} & \tilde{D}_{(1)} & \cdots & \tilde{D}_{(p)} \end{array}^T \text{ and } \tilde{\mathbf{R}} = \begin{array}{cccc} \tilde{R}_{obs} & \tilde{R}_{(1)} & \cdots & \tilde{R}_{(p)} \end{array}^T,$$

let $\bar{Q} = \frac{1}{p+1} \sum_{l=1}^{p+1} Q_l$, $\bar{\tilde{D}} = \frac{1}{p+1} \sum_{l=1}^{p+1} \tilde{D}_l$, $\bar{\tilde{R}} = \frac{1}{p+1} \sum_{l=1}^{p+1} \tilde{R}_l$ and $\tilde{\Sigma} = Cov(\mathbf{Q}, \tilde{\mathbf{D}}, \tilde{\mathbf{R}})$.

For $l = 1, \cdots, p+1$, let

$$U_l = \begin{array}{cccccc} Q_l - \bar{Q} & D_l - \bar{\tilde{D}} & R_l - \bar{\tilde{R}} & \tilde{\Sigma}^{-1} & Q_l - \bar{Q} & D_l - \bar{\tilde{D}} & R_l - \bar{\tilde{R}} \end{array}^T$$

The decision rule will reject (2.8) at level α if U_1 (i.e., U_{obs}) $\geq \mathbf{U}_{1-\alpha}$, the $(1-\alpha)$th quantile of $\mathbf{U} = (U_1, \cdots, U_{p+1})^T$.

2.4.3.2 Hybrid Test

Now we consider a two-step procedure for testing (2.8). The first step is testing whether the zero proportions are the same or not. We use Fisher's exact test (Fisher, 1922) for evaluating whether $p_1 = p_2$ to account for potentially small and unbalanced sample sizes. When the sample size is large and balanced, standard z-test for comparing two proportions using the Normal approximation could be used. If the first step test rejects $p_1 = p_2$, we reject the overall null hypothesis in (2.8) and conclude that the two samples are from different populations. Otherwise, in the second step, we use a permutation test based on sample means and variances of the non-zero component (i.e., h_Y) to see whether we can reject the overall null hypothesis in (2.8). The second step of the procedure is the same as in Section 2.4.2.1 except we use $\tilde{\mathbf{D}}$ and $\tilde{\mathbf{R}}$ defined in Section 2.4.3.1 instead of $\mathbf{D}$ and $\mathbf{R}$.

Because the two-step test (denoted as *Hybrid* in what follows) actually involves two separate tests, we need an adjustment to obtain the correct overall type I error rate. Here we use a Bonferroni correction. This correction assumes independence among test statistics and can reduce the power to detect the true effect (Gelman et al., 2012). For our situation, we prove that the type I error rate under this correction is less than or equal to the nominal level. In other words, the test is a level α test, but a more powerful test exists (see **Theorem 2** in Appendix). As with other multiple comparison procedures, one drawback of this hybrid test is that for a particular set of observations, there is no overall p-value available. The only thing we can say is whether we reject or fail to reject the null hypothesis at a certain nominal level.

2.5 Simulation

To evaluate the performance of the tests we derived in Section 3—in terms of type I error rate and power—we conducted a simulation study. We compared the tests defined in Section 3 with several other tests that typically only focus on a single feature of the populations in a two-sample comparison.

Table 2.1 gives a compilation of all the tests we compare in our simulations. Among them, RS, SEA and SEAD are designed to compare location parameters only; KS assumes a continuous distribution[1]; *LRT*, $\widehat{S}$ and $\widetilde{S}$ are standard likelihood-based tests; the others (*MLR*, $\widehat{Mscore}$, $\widetilde{Mscore}$, T, V, U and *Hybrid*) are our new procedures, designed for simultaneous comparison between two populations.

Table 2.1: Test statistics evaluated in simulation

Acronym	Test
KS	Kolmogorov-Smirnov test
RS	Wilcoxon Rank-Sum test
SEA	Permutation-based test using two-sample t-test statistic
SEAD	Permutation-based test using mean difference
LRT	Standard LRT
MLR	Fisher's method-based LRT
$\widehat{S}$	Standard Score test using MLE under reduced model
$\widetilde{S}$	Standard Score test using MLE under full model
$\widehat{Mscore}$	Fisher's method-based Score test using MLE under reduced model
$\widetilde{Mscore}$	Fisher's method-based Score test using MLE under full model
T	Permutation test based on sample mean and variance
V	Permutation test based on sample proportion, mean and variance
U	Permutation test based on sample proportion, mean and variance of non-zero component
Hybrid	Two-step test based on Fisher's exact test and permutation

[1]For this tests, we assume that the underlying distribution is continuous with support $[0,1)$

In our simulations, we considered several different settings for the BEZI parameters ($p = 0.1, 0.3, 0.6$, $\mu = 0.1, 0.5$, $\phi = 10, 100$), as well as several different sample sizes, since the likelihood-based tests should perform better for large samples. These sample sizes are $n_1 = n_2 = 10, 30, 60, 100$ in the equal sample size situations and $n_1 = 20$ vs $n_2 = 40$, $n_1 = 40$ vs $n_2 = 80$ and $n_1 = 20$ vs $n_2 = 60$ in the unequal sample size situations. For each configuration of parameters and sample sizes, we ran 1000 simulations. For the permutation-based tests, we used 1999 permutations. The nominal significance level is $\alpha = 0.05$. Portions of the results regarding type I error rate and power are summarized in Tables 2.2 and 2.3.

In terms of type I error rate (Table 2.2), for both equal and unequal sample size situations, all the tests perform as well as expected, except for the KS test. Because all the assumptions for the RS, SEA and SEAD are met, they have reasonable type I error rates for all settings. T, V and U have the desired type I error rate as well. The type I error rate of *Hybrid* is often less than the nominal level as expected. As sample size increases, the type I error rates of all the asymptotic tests get closer to the desired nominal level. *LRT* and *MLR* outperform the score tests. A possible reason is that the independence is exact for the likelihood ratio based tests but only asymptotic for the score based tests. In general, compared with the RS, SEA, SEAD and our new permutation tests T, V and U, the likelihood-based tests are slightly inferior, and *Hybrid* is perhaps too conservative (as expected).

In terms of power, a general result is that when the parameters of the Beta component remain the same and the proportion of zeros gets larger, the power of all tests goes down. This is likely because the zeros are not as informative as the non-zeros. The higher the

Table 2.2: Empirical type I error rate (%) of different statistics for testing homogeneity of two BEZI populations when the data are simulated from $BEZI(p_i, \mu_i, \phi_i)$ with sample size n_i, $i = 1, 2$; based on 1,000 replications using level at $\alpha = 0.05$; the bold numbers are ranged from 4.3 to 5.7.

Test statistics	$p_1 = 0.3, \mu_1 = 0.1, \phi_1 = 10$ $p_2 = 0.3, \mu_2 = 0.1, \phi_2 = 10$						
	$n_1 = 10$ $n_2 = 10$	$n_1 = 30$ $n_2 = 30$	$n_1 = 60$ $n_2 = 60$	$n_1 = 100$ $n_2 = 100$	$n_1 = 20$ $n_2 = 40$	$n_1 = 40$ $n_2 = 80$	$n_1 = 20$ $n_2 = 60$
KS	1.2	3.6	**4.7**	4.0	2.7	3.3	2.3
RS	4.2	**4.8**	**4.6**	**5.7**	3.7	**5.1**	**5.6**
SEA	**4.8**	**4.4**	3.8	**4.6**	3.8	4.1	**5.0**
SEAD	**5.0**	**4.4**	4.2	**4.6**	4.1	**4.4**	**4.9**
LRT	8.9	5.9	5.8	**5.0**	**5.4**	**5.6**	6.3
MLR	9.0	5.8	**5.4**	**4.9**	**5.2**	**5.1**	6.6
$\widehat{S}$	12.2	8.5	6.5	**5.2**	6.4	7.0	8.4
$\widetilde{S}$	18.7	8.5	7.1	5.9	10.1	8.2	10.7
$\widehat{Mscore}$	16.2	9.0	6.6	**5.3**	8.6	6.4	11.1
$\widetilde{Mscore}$	20.7	8.3	6.8	6.0	10.4	7.9	9.1
T	**4.5**	**4.9**	**4.7**	**5.4**	**5.2**	4.2	**4.9**
V	**4.7**	**4.7**	**5.5**	**5.7**	**5.1**	**4.8**	**4.9**
U	**4.9**	**4.5**	**5.1**	5.8	**4.4**	**5.7**	**4.5**
Hybrid	2.9	3.1	**4.3**	**4.7**	3.5	**5.3**	3.7

proportion of zeros, the less information in the sample. For both equal and unequal sample size situations, the difference in locations are easily detected by all tests, as shown in the left panels in Table 2.3. When the zero proportion and mean of the Beta component are the same (so that the means of two BEZI populations are the same), but the precisions are different, the RS, SEA and SEAD perform much worse than all of our new procedures (the center panels of Table 2.3). Because SEA and SEAD are designed for comparison of location parameters, when the real difference is only in the scale parameters, their power is very low for both equal and unequal sample sizes (the

Table 2.3: Empirical power (%) of different statistics for testing homogeneity of two BEZI populations when the data are simulated from $BEZI(p_i, \mu_i, \phi_i)$ with sample size n_i, $i = 1, 2$; based on 1,000 replications using level at $\alpha = 0.05$; the bold numbers are greater than or equal to 80.

Test statistics	$p_1 = 0.3, \mu_1 = 0.5, \phi_1 = 10$ $p_2 = 0.3, \mu_2 = 0.1, \phi_2 = 10$			$p_1 = 0.3, \mu_1 = 0.1, \phi_1 = 10$ $p_2 = 0.3, \mu_2 = 0.1, \phi_2 = 100$			$p_1 = 0.1, \mu_1 = 0.1, \phi_1 = 10$ $p_2 = 0.3, \mu_2 = 0.1, \phi_2 = 10$		
	$n_1 = 10$ $n_2 = 10$	$n_1 = 30$ $n_2 = 30$	$n_1 = 60$ $n_2 = 60$	$n_1 = 10$ $n_2 = 10$	$n_1 = 30$ $n_2 = 30$	$n_1 = 60$ $n_2 = 60$	$n_1 = 10$ $n_2 = 10$	$n_1 = 30$ $n_2 = 30$	$n_1 = 60$ $n_2 = 60$
KS	54.1	**99.7**	**100**	6.2	34.3	69.0	3.1	23.7	50.0
RS	44.7	**87.2**	**99.4**	7.3	11.7	14.6	10.6	29.3	47.7
SEA	**83.2**	**99.9**	**100**	7.9	5.8	5.0	8.9	15.1	23.6
SEAD	**83.5**	**99.9**	**100**	7.8	6.0	5.0	8.6	14.7	24.2
LRT	**99.4**	**100**	**100**	76.2	**99.8**	**100**	18.1	36.2	64.9
MLR	**99.2**	**100**	**100**	72.5	**99.5**	**100**	18.3	35.4	61.8
$\widehat{S}$	67.7	**97.7**	**99.9**	42.2	**96.1**	**98.4**	38.6	44.5	68.3
$\widetilde{S}$	**85.2**	**100**	**100**	**91.6**	**100**	**100**	44.2	32.2	60.2
$\widehat{Mscore}$	67.9	**80.0**	**92.4**	42.9	**98.6**	**100**	42.7	44.5	66.5
$\widetilde{Mscore}$	68.2	**92.0**	**99.9**	**90.8**	**100**	**100**	45.0	30.6	57.7
T	**93.2**	**100**	**100**	15.9	43.8	78.3	8.7	15.2	24.5
V	**90.5**	**100**	**100**	13.6	33.2	70.5	13.2	38.6	67.8
U	**97.7**	**100**	**100**	40.0	**93.7**	**99.8**	11.8	36.7	66.2
Hybrid	**95.1**	**100**	**100**	37.1	**93.2**	**99.8**	3.7	30.4	63.6

Test statistics	$p_1 = 0.3, \mu_1 = 0.5, \phi_1 = 10$ $p_2 = 0.3, \mu_2 = 0.1, \phi_2 = 10$			$p_1 = 0.3, \mu_1 = 0.1, \phi_1 = 10$ $p_2 = 0.3, \mu_2 = 0.1, \phi_2 = 100$			$p_1 = 0.1, \mu_1 = 0.1, \phi_1 = 10$ $p_2 = 0.3, \mu_2 = 0.1, \phi_2 = 10$		
	$n_1 = 20$ $n_2 = 40$	$n_1 = 40$ $n_2 = 80$	$n_1 = 20$ $n_2 = 60$	$n_1 = 20$ $n_2 = 40$	$n_1 = 40$ $n_2 = 80$	$n_1 = 20$ $n_2 = 60$	$n_1 = 20$ $n_2 = 40$	$n_1 = 40$ $n_2 = 80$	$n_1 = 20$ $n_2 = 60$
KS	**99.7**	**100**	**99.6**	31.3	59.6	31.3	19.2	35.5	18.0
RS	**83.0**	**97.6**	**83.5**	10.3	15.6	11.4	21.8	41.0	27.3
SEA	**99.9**	**100**	**100**	10.1	10.7	11.7	12.3	22.0	14.3
SEAD	**99.9**	**100**	**100**	10.2	10.2	11.9	12.0	21.4	14.2
LRT	**100**	**100**	**100**	**99.1**	**100**	**99.5**	30.5	58.3	33.8
MLR	**100**	**100**	**100**	**98.8**	**100**	**99.4**	29.4	56.1	33.8
$\widehat{S}$	32.8	18.9	78.1	**86.7**	**83.5**	78.7	41.9	68.3	46.9
$\widetilde{S}$	**98.1**	**100**	**100**	**99.7**	**100**	**99.5**	20.4	30.0	21.4
$\widehat{Mscore}$	56.1	37.7	**94.9**	**97.5**	**100**	**98.0**	41.7	66.6	46.7
$\widetilde{Mscore}$	**90.6**	**100**	**99.6**	**99.20**	**100**	**99.3**	21.0	28.7	21.4
T	**100**	**100**	**100**	45.5	**80.3**	54.6	15.3	23.3	15.6
V	**100**	**100**	**100**	33.2	72.4	41.7	29.1	58.1	30.3
U	**100**	**100**	**100**	**83.8**	**100**	**86.4**	23.6	56.2	23.4
Hybrid	**100**	**100**	**100**	**81.9**	**99.9**	**85.1**	27.8	58.4	28.0

center panels of Table 2.3). The RS has slightly higher power than SEA and SEAD. T and V have marginally smaller power than our other new tests in both equal and unequal sample size cases when only the scales are different. The KS test has better power than the RS, SEA and SEAD, but they all have inferior power relative to the new tests. When the difference exists only between the zero proportions (the right panels of Table 2.3), all tests have relatively low power. In general, the empirical power is less than 0.50 unless the sample sizes are large ($n_1 = n_2 = 60$ and $n_1 = 40$ vs. $n_2 = 80$). The new tests are in general more powerful than KS, RS, SEA and SEAD. When at least two of the parameters are different, all tests have reasonable power. Due to space limitation, we do not include all results here. It is also interesting to notice that the power of the reduced model MLE-based score test is usually smaller than that of the full model MLE-based score test. The presence of negative score values is the likely reason for this. There is no clear distinction between the standard tests and those based on Fisher's method. Again, the likelihood ratio tests perform better than the score tests in terms of power. When the total sample sizes are equal (i.e., comparing $n_1 = n_2 = 30$ and $n_1 = 20$ and $n_2 = 40$), all tests generally have larger power in the equal sample size cases.

On the one hand, the simulations show that LRT, MLR, $\hat{S}$, $\tilde{S}$, $\widehat{MScore}$ and $\widetilde{MScore}$ all have similar power to U and $Hybrid$. T and V have slightly smaller power in a few settings. However, all of these tests are better than RS, SEA and SEAD in terms of power. On the other hand, RS, SEA and SEAD have type I error rates that are similar to those of T, V and U. LRT, MLR, $\hat{S}$, $\tilde{S}$, $\widehat{MScore}$ and $\widetilde{MScore}$ have larger type I error rates, unless the sample size is about 100 in each sample. $Hybrid$ is conservative as expected using the Bonferroni correction. The KS test performs the worst in terms of both type

I error rate and power. Overall, considering both power and type I error rate, U is the best test, followed by $Hybrid$, V and T. When the sample sizes are large, the likelihood ratio tests, LRT and MLR, perform well.

We also examined several settings with parameters near the boundary of the parameter space (p=0.05, μ=0.01 and 0.05, ϕ=0.5 and 1[2]). The patterns of type I error and power results are in general similar to those we report in Tables 2.2 and 2.3. However, the tests require larger sample sizes to achieve the desired nominal level in type I error rate or to have reasonably large power. We also found when parameters are close to their boundaries, although the two populations are different, because of the small magnitude of the difference, the power of the tests is also lower than in the cases reported in Table 2.3.

We implemented all tests in R (R Core Team, 2013) (code available from authors upon request). The tests are all computationally efficient. The computing time for the four permutation-based non-parametric and semi-parametric tests range between 4 and 7 seconds to compute for a single dataset with 1999 permutations (sample sizes vary from 20 to 200). All simulations were performed on an Intel© Core^{TM} 2 Quad CPU Q6600 @ 2.40GHz 2.40GHz processor with 4.00 GB RAM.

[2]We choose these numbers because when p is close to 1 we almost always get zeros in samples, as a result we do not have much information to compare; naturally when μ is small, there is likely to be zero inflation; when ϕ is greater than 10, for fixed μ, ϕ's influence on the variance of BEZI random variable is small.

2.6 Application to the Barnacle Settlement Data

In this section, we apply our new tests to the barnacle settlement data described in Section 2, and we compare our results to those in Tyburczy (2011); namely SEA and SEAD (refer to Section 2).

Table 2.4: Settlement data summary for *Balanus* cf. *glandula* at Sandhill Bluff, with and without physical processes (regional relaxation as relax, front passage as front and diurnal upwelling as dup). Abs/pre presents absent/present of process. MLE is based on zero-inflated Beta distribution. Sample zero proportions are the same as MLE of p's, therefore only present once.

Process name Sample Size (abs/pre)	relax 41 / 19	front 24 / 36	dup 28 / 32
means (abs/pre)	0.092 / 0.022	0.032 / 0.095	0.087 / 0.054
variances (abs/pre)	0.023 / 0.001	0.003 / 0.025	0.029 / 0.007
p (abs/pre)	0.415 / 0.632	0.583 / 0.417	0.500 / 0.468
μ (abs/pre)	0.169 / 0.061	0.079 / 0.176	0.193 / 0.103
ϕ (abs/pre)	5.416 / 65.139	31.068 / 4.914	4.021 / 14.696

Table 2.4 provides sample means and variances and MLE of all parameters under the BEZI distribution assumption for the settlement data at Sandhill Bluff. The sample zero proportions are the MLE of the population zero proportions. For these data, the sample sizes are nearly equal in some cases and twofold different in other cases. There are many zeros in all cases, with many samples having zero proportions greater than 50%. The sample means are less than 0.1 and the sample variances are all small.

As shown in Table 2.5, the KS and *Hybrid* tests conclude that there is no difference between samples. The likelihood-based tests conclude that the samples are different, and these tests have the smallest p-values among all the tests. The p-value that is greater

Table 2.5: Simultaneous test results for settlement of *Balanus* cf. *glandula* at Sandhill Bluff, with and without physical processes (regional relaxation as relax, front passage as front and diurnal upwelling as dup). The number of observations are indicated as sample size. Test names are consistent with Table 2.1. The p-values are two-tailed; those < 0.1 are bold and those < 0.05 are marked with an asterisk ($*$). Hybrid test does not have an overall p-value, we indicate whether it rejects or fails to reject the null hypothesis that the two samples are the same.

Process name Sample Size (abs/pre)	relax 41 / 19	front 24 / 36	dup 28 / 32
KS	0.167	0.269	0.982
RS	**0.045**$*$	0.107	0.981
SEA	**0.044**$*$	**0.058**	0.366
SEAD	**0.039**$*$	**0.055**	0.374
LRT	**0.004**$*$	**0.013**$*$	**0.076**
MLR	**0.004**$*$	**0.011**$*$	**0.079**
$\widehat{S}$	$>$0.999	**0.006**$*$	**0.090**
$\widetilde{S}$	$<$**0.001**$*$	$<$**0.001**$*$	**0.013**$*$
$\widehat{MScore}$	0.483	0.461	**0.084**
$\widetilde{MScore}$	**0.099**	**0.003**$*$	**0.011**$*$
T	**0.037**$*$	0.116	0.824
V	**0.056**	0.184	0.727
U	**0.046**$*$	0.417	0.686
Hybrid	fails to reject	fails to reject	fails to reject

than 0.999 under the regional relaxation (relax) condition for $\hat{S}$ is due to the presence of negative score statistic. The RS, T, V and U indicate that the settlement for Balanus larvae is associated with regional relaxation only. We would conclude from SEA and SEAD that the settlement for Balanus larvae is associated with regional relaxation and front passage.

The possible reason for the discrepancy between the likelihood-based tests and the non- and semi-parametric tests is that these data are actually well-represented by the

BEZI distribution. When the parametric assumption is satisfied, the parametric tests are typically more powerful than the non-parametric and semi-parametric alternatives. Because of the large proportions of zeros, the sample means are pulled close to zero and the sample variances are also small. In other words, the data do not vary a lot. Therefore, the magnitude of difference between two samples are small compared to the potential range. The unbalanced sample sizes between two samples also provide some challenges for all the tests.

2.7 Discussion

We have extended Paul and Jiang's (2005) study to assess homogeneity across two three-parameter populations that are mixtures of discrete and continuous components (namely, the BEZI distribution). This distribution is particularly useful for modeling data taking the form of a proportional change in some continuous measurement, such as space, time or income, with the added complication of zero-inflation. We develop two non-parametric and two semi-parametric simultaneous tests to compare two or three features between the two populations simultaneously. Fisher's method based on likelihood ratio test statistics outperforms the tests based on score test statistics. The non-parametric and semi-parametric simultaneous tests have even more desirable properties than the tests based on Fisher's method. Specifically, their type I error rates are desirable even in small sample situations and their power is comparable with the parametric tests. In addition, these non-parametric and semi-parametric tests can be applied to other zero- and/or one-inflated distributions directly. All methods can be expanded to compare more

than two populations.

In Paul and Jiang's simulation for Normal and Negative Binomial distributions, the standard likelihood ratio test has larger type I error rate than the desired level (0.05), and the standard score test has smaller type I error rate than the desired level. But the tests based on Fisher's method have the correct levels. These distinctions do not appear in our BEZI simulation[3]. The standard tests achieve similar type I error rate and power as the tests based on Fisher's method. More interestingly, the score test statistics should equal the likelihood ratio test asymptotically, so it is not clear why they behaved differently in Paul and Jiang's simulations, even in the large sample size case. Furthermore, Paul and Jiang (2005) do not mention the negative score statistics in their paper, but we came across such occurrences with the barnacle settlement data and many times in our simulation study. Thiagarajah's (2012) simulations show similar patterns as ours: Fisher's method using likelihood ratio tests has smaller type I error rate than Fisher's method using score tests. However, there is no clear distinction between the standard tests and combined tests. Thiagarajah (2012) recommends using combined score tests because of the simplicity in maximum likelihood estimation and inversion of a low-dimension matrix.

One benefit of considering combined tests, rather than a single likelihood ratio test, is that it provides more information about what aspect or aspects of two populations might be different. The sub-tests deal with specific features of the data distributions, so they may be useful when examined individually. Based on our simulation results, within the

[3]There was no evidence of this in zero-inflated Poisson simulations either, and only weak evidence in Normal simulations.

parametric setting, we recommend combining the likelihood ratio tests using Fisher's method. The non-parametric simultaneous tests we developed can be applied to other populations with and without inflation directly, as can the semi-parametric tests. They can also be extended to cases of both zero and one inflation. This is particular appealing in ecological data, where zero inflation is common due to reasons such as species rarity or poor habitat conditions (Chiogna and Gaetan, 2007; Barry and Welsh, 2002). Current statistical methods for ecological analysis often involve count data (Marin et al., 2005). Zero-inflated continuous data are mostly based on the log-normal (Tian, 2005), Gamma (Feuerverger, 1979) and exponential (Wu et al., 2012; Zhang et al., 2010) model. Our study fills the gap in simultaneous comparison of zero-inflated continuous data with a constrained support.

Finally, when sample sizes are large and the data can be well represented by BEZI distributions, likelihood-based tests are superior to the non-parametric and semi-parametric alternatives. In small sample size situations, however, the non-parametric and semi-parametric tests are preferred.

2.8 Appendix

2.8.1 Partition of Multivariate Normal Distribution

Lemma 1 (Partition of multivariate normal distribution)**.** *Suppose* $Y \sim N_p(\mu, V)$ *where* V *is nonsingular. Let* Y, μ *and* V *be similarly partitioned in the form* $Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}$, $\mu =$

$\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}$, $V = \begin{bmatrix}V_{11} & V_{12}\\ V_{21} & V_{22}\end{bmatrix}$. *Assume* Y_1 *is* $q\times 1$ *and that* $s = p-q$ *so that* Y_2 *is* $s\times 1$. *Set* $V_{11.2} = V_{11} - V_{12}V_{22}^{-1}V_{21}$ *and let* $T = Y_1 - V_{12}V_{22}^{-1}Y_2$. *Then*

1. $T \sim N_q(\mu_1 - V_{12}V_{22}^{-1}\mu_2, V_{11.2})$ *and* $V_{11.2}$ *is nonsingular.*

2. T *and* Y_2 *are independent random vectors.*

Proof. 1. Since $T = \begin{bmatrix} I_q & -V_{12}V_{22}^{-1}\end{bmatrix}\begin{bmatrix}Y_1\\ Y_2\end{bmatrix}$,

$$T \sim N_q\left(\begin{bmatrix} I_q & -V_{12}V_{22}^{-1}\end{bmatrix}\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}, \begin{bmatrix} I_q & -V_{12}V_{22}^{-1}\end{bmatrix}\begin{bmatrix}V_{11} & V_{12}\\ V_{21} & V_{22}\end{bmatrix}\begin{bmatrix} I_q^T\\ -V_{22}^{-1}V_{21}\end{bmatrix}\right)$$

$$\xrightarrow{d} N_q(\mu_1 - V_{12}V_{22}^{-1}\mu_2, V_{11.2}).$$

Since

$$V_{11.2} = \begin{bmatrix} I_q & -V_{12}V_{22}^{-1}\end{bmatrix}\begin{bmatrix}V_{11} & V_{12}\\ V_{21} & V_{22}\end{bmatrix}\begin{bmatrix} I_q^T\\ -V_{22}^{-1}V_{21}\end{bmatrix},$$

$$rank\begin{bmatrix} I_q^T\\ -V_{22}^{-1}V_{21}\end{bmatrix} = rank\begin{bmatrix} I_q & -V_{12}V_{22}^{-1}\end{bmatrix} = dim(\mathscr{R}(I_q)) + dim(\mathscr{R}(-V_{12}V_{22}^{-1}))$$

$$-dim(\mathscr{R}(I_q)\cap\mathscr{R}(-V_{12}V_{22}^{-1}))$$

$$= dim(\mathbb{R}^q) + dim(\mathscr{R}(-V_{12}V_{22}^{-1})) - dim(\mathbb{R}^q\cap\mathscr{R}(-V_{12}V_{22}^{-1})) = q.$$

That is,

$$\begin{bmatrix} I_q^T\\ -V_{22}^{-1}V_{21}\end{bmatrix} \text{ is full-column rank.}$$

Therefore, $V_{11.2}$ is positive definite, and so nonsingular.

2. $Cov(T, Y_2) = Cov(Y_1 - V_{12}V_{22}^{-1}Y_2, Y_2) = Cov(Y_1, Y_2) - Cov(V_{12}V_{22}^{-1}Y_2, Y_2)$

$$= V_{12} - V_{12}V_{22}^{-1}Cov(Y_2, Y_2) = V_{12} - V_{12}V_{22}^{-1}V_{22} = 0.$$

Since T and Y_2 are jointly normally distributed, T and Y_2 are independent.

□

2.8.2 Asymptotic Independence among Score Tests

Theorem 1 (Asymptotic Independence among Score Tests)**.** *$\hat{S_1}$, $\hat{S_2}$ and $\hat{S_3}$ (and also $\tilde{S_1}$, $\tilde{S_2}$ and $\tilde{S_3}$) are asymptotically independent.*

Proof. Consider $\hat{S_1}$, $\hat{S_2}$ and $\hat{S_3}$ first: Denote

$$\omega = (p, \mu, \phi)^T,\ \zeta = (\delta, p, \mu, \phi)^T = (\delta, \omega^T)^T,\ \eta = (\gamma, \delta, p, \mu, \phi)^T = (\gamma, \zeta^T)^T = (\gamma, \delta, \omega^T)^T,$$

$$\theta = (\beta, \gamma, \delta, p, \mu, \phi)^T = (\beta, \eta^T)^T = (\beta, \gamma, \zeta^T)^T = (\beta, \gamma, \delta, \omega^T)^T.$$

Recall that

$$\hat{S_1} = \hat{s_1}^T(\hat{A_1} - \hat{C_1}\hat{D_1}^{-1}\hat{C_1}^T)^{-1}\hat{s_1},$$

$$\hat{S_2} = \hat{s_2}^T(\hat{A_2} - \hat{C_2}\hat{D_2}^{-1}\hat{C_2}^T)^{-1}\hat{s_2}$$

and

$$\hat{S_3} = \hat{s_3}^T(\hat{A_3} - \hat{C_3}\hat{D_3}^{-1}\hat{C_3}^T)^{-1}\hat{s_3}.$$

Under H_0,

$$s_1 = \frac{\partial l_{1f}}{\partial \delta} = \frac{\partial l_f}{\partial \delta}, s_2 = \frac{\partial l_{2f}}{\partial \gamma} = \frac{\partial l_f}{\partial \gamma}, \text{ and } s_3 = \frac{\partial l_{3f}}{\partial \beta} = \frac{\partial l_f}{\partial \beta}.$$

And

$$\hat{s_1} = s_1(\hat{\omega}), \hat{s_2} = s_2(\hat{\zeta}), \hat{s_3} = s_3(\hat{\eta}),$$

where $\hat{\omega}$, $\hat{\zeta}$ and $\hat{\eta}$ are maximum likelihood estimates of ω under $H_0^{'}$, ζ under $H_0^{''}$ and η under $H_0^{'''}$, respectively (i.e., $\hat{s_1} = \frac{\partial l_f}{\partial \delta}\Big|_{\omega=\hat{\omega}}$, $\hat{s_2} = \frac{\partial l_f}{\partial \gamma}\Big|_{\zeta=\hat{\zeta}}$ and $\hat{s_3} = \frac{\partial l_f}{\partial \beta}\Big|_{\eta=\hat{\eta}}$). Notice that when H_0 is true, $H_0^{'}$, $H_0^{''}$ and $H_0^{'''}$ are all true. Expanding $\hat{s_1}$, $\hat{s_2}$ and $\hat{s_3}$ around the true parameters θ_0 via Taylor expansion with $l_f = l$:

$$\hat{s_1} = \frac{\partial l}{\partial \delta} - I_{\delta\omega_0} I^{-1}_{\omega_0\omega_0} \frac{\partial l}{\partial \omega_0} + \mathscr{O}_p(1),$$

$$\hat{s_2} = \frac{\partial l}{\partial \gamma} - I_{\gamma\zeta_0} I^{-1}_{\zeta_0\zeta_0} \frac{\partial l}{\partial \zeta_0} + \mathscr{O}_p(1)$$

and

$$\hat{s_3} = \frac{\partial l}{\partial \beta} - I_{\beta\eta_0} I^{-1}_{\eta_0\eta_0} \frac{\partial l}{\partial \eta_0} + \mathscr{O}_p(1),$$

where

$$I_{\delta\omega_0} = E\left(-\frac{\partial^2 l}{\partial\delta\partial\omega^T}\,\Big|\,H_0\right), I_{\omega_0\omega_0} = E\left(-\frac{\partial^2 l}{\partial\omega\partial\omega^T}\,\Big|\,H_0\right), I_{\gamma\zeta_0} = E\left(-\frac{\partial^2 l}{\partial\gamma\partial\zeta^T}\,\Big|\,H_0\right),$$

$$I_{\zeta_0\zeta_0} = E\left(-\frac{\partial^2 l}{\partial\zeta\partial\zeta^T}\,\Big|\,H_0\right), I_{\beta\eta_0} = E\left(-\frac{\partial^2 l}{\partial\beta\partial\eta^T}\,\Big|\,H_0\right), I_{\eta_0\eta_0} = E\left(-\frac{\partial^2 l}{\partial\eta\partial\eta^T}\,\Big|\,H_0\right),$$

and

$$\frac{\partial l}{\partial \omega_0} = \frac{\partial l}{\partial \omega}\Big|_{\omega=\omega_0}, \frac{\partial l}{\partial \zeta_0} = \frac{\partial l}{\partial \zeta}\Big|_{\zeta=\zeta_0}, \frac{\partial l}{\partial \eta_0} = \frac{\partial l}{\partial \eta}\Big|_{\eta=\eta_0}.$$

Let

$$s_{01} = \frac{\partial l}{\partial \delta} - I_{\delta\omega_0} I^{-1}_{\omega_0\omega_0} \frac{\partial l}{\partial \omega_0}, \; s_{02} = \frac{\partial l}{\partial \gamma} - I_{\gamma\zeta_0} I^{-1}_{\zeta_0\zeta_0} \frac{\partial l}{\partial \zeta_0} \text{ and } s_{03} = \frac{\partial l}{\partial \beta} - I_{\beta\eta_0} I^{-1}_{\eta_0\eta_0} \frac{\partial l}{\partial \eta_0}.$$

Since the score function can be written as

$$\frac{\partial l}{\partial \theta} = \begin{pmatrix} \frac{\partial l}{\partial \beta} \\ \frac{\partial l}{\partial \gamma} \\ \frac{\partial l}{\partial \delta} \\ \frac{\partial l}{\partial p} \\ \frac{\partial l}{\partial \mu} \\ \frac{\partial l}{\partial \phi} \end{pmatrix} = \begin{pmatrix} \frac{\partial l}{\partial \beta} \\ \frac{\partial l}{\partial \eta} \end{pmatrix} = \begin{pmatrix} \frac{\partial l}{\partial \beta} \\ \frac{\partial l}{\partial \gamma} \\ \frac{\partial l}{\partial \zeta} \end{pmatrix} = \begin{pmatrix} \frac{\partial l}{\partial \beta} \\ \frac{\partial l}{\partial \gamma} \\ \frac{\partial l}{\partial \delta} \\ \frac{\partial l}{\partial \omega} \end{pmatrix},$$

we have $\frac{\partial l}{\partial \theta} \xrightarrow{d} N(0, I(\theta_0))$. Or

$$\begin{pmatrix} \frac{\partial l}{\partial \beta} \\ \frac{\partial l}{\partial \eta} \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ \\ 0 \end{pmatrix}, -E \begin{pmatrix} \frac{\partial^2 l}{\partial \beta^2} & \frac{\partial^2 l}{\partial \beta \partial \eta^T} \\ \\ \frac{\partial^2 l}{\partial \eta \partial \beta} & \frac{\partial^2 l}{\partial \eta \partial \eta^T} \end{pmatrix} \begin{matrix} \\ \\ H_0 \end{matrix} \right).$$

As the sample size going to $+\infty$, using the result from partition of multivariate normal

distribution, under H_0, s_{03} is independent of $\frac{\partial l}{\partial \eta} = \begin{pmatrix} \frac{\partial l}{\partial \gamma} \\ \\ \frac{\partial l}{\partial \zeta} \end{pmatrix} = \begin{pmatrix} \frac{\partial l}{\partial \gamma} \\ \\ \frac{\partial l}{\partial \delta} \\ \\ \frac{\partial l}{\partial \omega} \end{pmatrix}$. Since s_{01} and s_{02} are only functions of $(\frac{\partial l}{\partial \gamma}, \frac{\partial l}{\partial \zeta})$ and $(\frac{\partial l}{\partial \delta}, \frac{\partial l}{\partial \omega})$, s_{03} is independent of s_{01} and s_{02}. Similarly, it is easy to show s_{01} and s_{02} are independent as well.

For the independence among $\tilde{S_1}$, $\tilde{S_2}$ and $\tilde{S_3}$, similar arguments can be made. □

2.8.3 Size of Hybrid Test

Theorem 2 (Size of **Hybrid** test)**.** *The hypothesis in (2.8) is*

$$H_0 : p_1 = p_2 \text{ and } h_{Y_1} = h_{Y_2} \text{ vs } H_1 : \text{ Either equality fails.}$$

Here, p_1 and p_2 are the population zero proportions for populations 1 and 2, and h_{Y_1} and h_{Y_2} are the pdf for the non-zero components of populations 1 and 2. To test these hypotheses at the level α, the Hybrid procedure involves the following steps:

1. *Test $H_{01} : p_1 = p_2$ versus $H_{11} : p_1 = p_2$.*
2. *H_{01} is rejected if the p-value is $p < \alpha_1$. In this case, reject H_0 in (2.8); otherwise, test $H_{02} : h_{Y_1} = h_{Y_2}$ versus $H_{12} : h_{Y_1} = h_{Y_2}$.*
3. *H_{02} is rejected if the p-value is $p < \alpha_2$. In this case, reject H_0 in (2.8); otherwise, do not reject H_0.*

The size of Hybrid is less than or equal to $\alpha_1 + \alpha_2$.

Proof. Define

$$A = \{(p_1, p_2) \in (0,1)^2 : p_1 = p_2\} \text{ and } B = \{h_{Y_1}, h_{Y_2} \in \mathscr{F} : h_{Y_1} = h_{Y_2}\}$$

where $\mathscr{F}$ is the space of a particular family of pdf. Under the set notation, H_0, H_1, H_{01}, H_{11}, H_{02} and H_{12} can be expressed as $H_0 : A \cap B$, $H_1 : A^c \cup B^c$, $H_{01} : A$, $H_{11} : A^c$, $H_{02} : A \cap B$ and $H_{12} : A \cap B^c$.

Further define $\mathfrak{R}_1$, $\mathfrak{R}_2$ and $\mathfrak{R}$ as rejection region for H_{01}, H_{02} and H_0, then

$$C = \{\mathbf{x} : \mathbf{x} \in \mathfrak{R}_1\} \text{ and } D = \{\mathbf{x} : \mathbf{x} \in \mathfrak{R}_2\},$$

then we have

$$C \cup (C^c \cap D) = \{\mathbf{x} : \mathbf{x} \in \mathfrak{R}\}.$$

It is reasonable to assume A and B are independent because the form of h_Y does not depend on p, neither the other way around. Also, it is reasonable to assume B and C are independent because C only concerns the p's.

Let α_0, α_1 and α_2 denote the type I error rates of H_0 vs H_1, H_{01} vs H_{11} and H_{02} vs H_{12}, respectively.

Then

$$\alpha_0 = P(\text{Reject } H_0 | H_0 \text{ is true}) = P(C \cup (C^c \cap D) | A \cap B),$$

$$\alpha_1 = P(C|A) \text{ and } \alpha_2 = P(D|A \cap B).$$

Since $C \cap (C^c \cap D) = \emptyset$, we have $(C|A \cap B) \cap (C^c \cap D|A \cap B) = \emptyset$. Therefore,

$$P(C \cup (C^c \cap D)|A \cap B) = P(C|A \cap B) + P((C^c \cap D)|A \cap B).$$

Since $C^c \cap D \subset D$, there is $C^c \cap D|A \cap B \subset D|A \cap B$. Therefore,

$$P((C^c \cap D)|A \cap B) \leq P(D|A \cap B) = \alpha_2.$$

$$P(C|A \cap B) = \frac{P(C \cap A \cap B)}{P(A \cap B)} = \frac{P(B|A \cap C) \cdot P(A \cap C)}{P(B|A) \cdot P(A)} = P(C|A) \cdot \frac{P(B|A \cap C)}{P(B|A)}$$

Because we assume A and B are independent and B and C are independent, $P(B|A) = P(B)$ and $P(B|A \cap C) = P(B)$. Hence,

$$P(C|A \cap B) = P(C|A) = \alpha_1$$

As a result,

$$\alpha_0 = P(C \cup (C^c \cap D)|A \cap B) = P(C|A \cap B) + P((C^c \cap D)|A \cap B) \leq \alpha_1 + \alpha_2.$$

Under the Bonferroni correction, $\alpha_1 = \alpha_2 = \alpha/2$. □

3 Zero-inflated Beta Hidden Markov Model

3.1 Abstract

In this paper, we use a Hidden Markov Model (HMM) with zero-inflated Beta (BEZI) emission densities to model zero-inflated proportion data with serial correlation. We show that the standard EM algorithm for HMM parameter estimation can be applied in this case with emission distributions that are mixtures of discrete and continuous parts. We conduct simulations to show the effectiveness of this approach. We find that the initial values, the number of observations and the BEZI density shape all impact the performance of the EM algorithm. We provide some suggestions about the choice of initial values. When decoding a hidden state chain, the Viterbi algorithm gives more accurate identifications than does posterior decoding. However, the Viterbi algorithm is sensitive to BEZI density shapes. For both the EM algorithm and the Viterbi algorithm, asymmetric BEZI densities provide a lesser challenge than other density shapes. We apply the method to an oceanographic dataset.

Keywords: Zero-inflated Beta, Hidden Markov Model, EM Algorithm, Viterbi Algorithm

3.2 Introduction

As discussed by Cox (1981), time series analysis for dependent data can be categorized as observational-driven or parameter-driven. Observation-driven models define the autocorrelation in the observations directly; i.e., the distribution of a variable, Y_t, is a function of previous observations, $Y_1, \cdots, Y_{t-1}$. Examples include autoregressive moving average models and Markov chain models. Parameter-driven models introduce the autocorrelation through a latent process, as for example, in Hidden Markov Models (HMM).

Many serially correlated observations can be well-modeled using HMM in which observations are random variables whose probability distributions depend on the current state of an unobserved Markov chain (Rabiner et al., 1985). The mathematical theory of HMM was first developed by Baum and his colleagues in a series of classic papers in the late 1960s and early 1970s (Baum and Petrie, 1966; Baum and Eagon, 1967; Baum et al., 1970; Baum, 1972). The model is especially well known for its successful applications in speech recognition (Baker, 1975; Rabiner, 1989; Rabiner and Juang, 1993) and bioinformatics (Bishop and Thompson, 1986; Durbin et al., 1998; Krogh et al., 1994). Other applications include handwriting recognition (Rigoll et al., 1996), gesture recognition (Starner and Pentland, 1995), music score following (Pardo and Birmingham, 2005), image processing (Yamato et al., 1992), partial discharge image classification (Satish and Gururaj, 1993), finance (Mamon and Elliott, 2007), ecology (Baum and Eagon, 1967) and leaning behavior of live and artificial systems (Petrushin, 2000).

Depending on the form of the probability density function at each state (called the "emission density" or "emission distribution" in this context), two types of HMM have been extensively studied: Discrete Density Hidden Markov Model (DDHMM) and Continuous Density Hidden Markov Model (CDHMM). A DDHMM has observations chosen from a finite or countable set following discrete emission distributions, whereas a CDHMM has observations generated from continuous emission densities. For CDHMM, the most commonly used density is a mixture of Gaussian densities (Rabiner, 1989) because it can well-approximate many continuous density functions (Sorenson and Alspach, 1971). However, when the number of hidden states increases, the number of parameters in these mixture models increases rapidly; leading to computational load. The situation can become worse when the true emission density is far from Gaussian. Furthermore, when the number of observations is small, the parameters of the mixture density become non-identifiable, making the model infeasible in practice. An alternative is to use a parametric model tailored to the observed data.

The contribution of this paper is to extend the HMM framework to the situation where the emission distribution is non-standard, and in fact, a mixture of discrete and continuous components. Specifically, we consider the zero-inflated Beta density (Ospina and Ferrari, 2010). Data that are proportions falling in the continuum (0, 1) are very common in practice. Examples include the proportion of conifer cover in a particular area, the proportion of household income spent on food and the proportion of weekly hours spent on work-related travel. The family of Beta distributions provides broad flexibility for modeling proportions. In some cases, however, an inflated number of zeros and/or ones in a sample of proportions can render the Beta distribution unsuitable

since it takes support on the open interval (0, 1). Ospina and Ferrari (2010) propose a mixed continuous-discrete distribution for data observed on [0, 1), (0, 1] or [0, 1]. The discrete component is defined by a degenerate (point mass) distribution that assigns non-zero probability to 0 and/or 1 depending on whether there is zero- and/or one-inflation. In particular, the zero-inflated Beta (BEZI), the primary focus for our study here, has a point mass at zero.

Suppose $Y \sim BEZI(p, \mu, \phi)$. Then the probability density function of Y is

$$f_Y(y) = \begin{cases} p & \text{if } y = 0, \\ (1-p)\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} & \text{if } y \in (0,1), \end{cases}$$

where $\Gamma(\cdot)$ is the Gamma function; $0 < p < 1$, $0 < \mu < 1$ and $0 < \phi < \infty$. The mean and variance of Y are, respectively,

$$E(Y) = (1-p)\mu$$

and

$$Var(Y) = (1-p)\frac{\mu(1-\mu)}{\phi+1} + p(1-p)\mu^2,$$

where μ and ϕ are the mean and precision parameters of the Beta component. As is the Beta family of distributions, the BEZI family is quite flexible in shape.

In this paper, we use a HMM with BEZI emission densities to model zero-inflated proportions with serial correlation. We provide an Expectation-Maximization (EM) algorithm (Dempster et al., 1977) for parameter estimation and justify its applicability in

this setting. We compare the Viterbi algorithm (Viterbi, 1967) and posterior decoding (Durbin et al., 1998) for identifying the underlying state sequence. We perform two simulation studies to evaluate our algorithm, and we apply our methods to an oceanographic dataset. In both simulations and real data application, our model performs well. Not surprisingly, the initial values, the number of observations and the BEZI density shape all impact the performance of the EM algorithm. When decoding a hidden state chain, the Viterbi algorithm gives more accurate identification than does posterior decoding. However, the Viterbi algorithm is sensitive to BEZI density shapes. For both the EM algorithm and the Viterbi algorithm, asymmetric BEZI densities provide a lesser challenge than other density shapes. We provide suggestions about the selection of initial values for the EM algorithm.

In Section 2, we describe an example from marine science where interest lies in distinguishing two populations. The observations from there can be well represented by BEZI distributions, and because of the sampling method there is autocorrelation among observations. In Section 3, we develop a two-state HMM with BEZI as the emission distributions. We give a review of standard analytical methods for traditional HMM problems, and then provide the EM algorithm for our BEZI HMM along with its justification. In Section 4, we report results of two simulation studies used to evaluate the performance of our method. In Section 5, we apply our method to the marine science data. We discuss our findings, possible extensions of this work and several interesting issues about the EM algorithm in the HMM framework in Section 6.

3.3 Settlement of Onshore Barnacle Larvae

Tyburczy (2011) compares settlement distributions of onshore barnacle larvae with and without the occurrence of different oceanographic processes such as large-scale regional relaxation of upwelling, and some smaller scale processes such as localized diurnal upwelling driven by afternoon sea breezes. The data consist of observations on two types of barnacle larva (*Balanus* cf.*glandula* and *Chthamalus* spp.): daily or bi-daily settlement information and covariate information such as the occurrence of multiple physical processes (regional relaxation, front passage, and diurnal upwelling) in four study areas (Sandhill Bluff, Terrace Point, Bonny Doon Beach and Lighthouse Point) collected within northern Monterey Bay, CA in 2007 (May-Sept).

The larva settlement data are a combination of samples from onshore intertidal plate after normalized for hours of immersion based on tidal height of the plates, samples from larval traps deployed on and below the surface of multiple moorings with different depths and samples from sea water collected using pumps during several field work. The presence/absence of physical processes were also determined based on combinations of different physical conditions; for example, temperature, salinity, current direction and velocity, local wind force and nearshore pressure gradients. More details are provided in Tyburczy (2011). For *Balanus* cf.*glandula*, the distributions of larva settlement are compressed to the range between zero and one after some data manipulation; i.e., all observations are in the interval [0, 1), and more than half of them are zeros. Some examples of the *Balanus* cf. *glandula* larvea settlement data are shown in Figure 3.1. All panels are for the Sandhill Bluff site with and without the occurrence of different

ocean processes. Notice that in each panel, the two zero counts are also indicated.

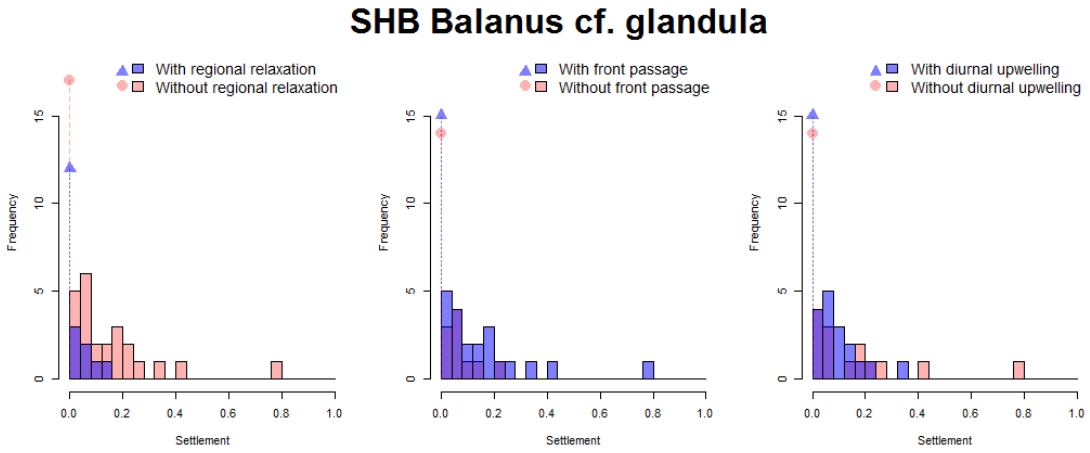


Figure 3.1: Histogram of settlement for *Balanus* cf.*glandula* in Sandhill Bluff (SHB) with and without physical processes (regional relaxation, front passage, and diurnal upwelling; left to right).

Tyburczy's ultimate goal was to build a new ecological model that describes the mechanism of nearshore barnacle transport and settlement procedure. To do this, the fundamental question is whether there is an association between the oceanographic processes and the larvae settlement outcomes. Evaluation of this association is complicated by non-normality and autocorrelation in the data, as well as uneven sampling intervals. In this paper, we use a HMM with BEZI emission densities to account for the non-normal distribution and the autocorrelation. In this work, we assume equal sampling intervals, which is approximately true for these data. We assume each time point corresponds to an unknown state from a Markov chain with two states – one for the population with occurrence of a particular physical process, which may or may not be observed (or correctly identified), and the other for the population without the occurrence of that pro-

cess. Given a particular state, a larva settlement outcome is generated from a population distribution in the BEZI family. We know the actual settlement sequence, but we assume that we do not know the state chain.

In Tyburczy's study, he determined the with- and without-processes time points based on values of other explanatory oceanographic features – arguably, a subjective determination. The appeal of the HMM approach is that it lets the data identify the two populations. And then one can attempt to align those determinations with values of the explanatory variables. When the determinations provided by the settlement data agree with some explanatory series, we argue that the binary formed explanatory series is the hidden selection process for switching between the states. Therefore, there is strong association between response and explanatory variable. Otherwise, the data provide no evidence of such an association. Using the HMM approach, we remove the subjectivity in determining the physical event. More importantly, the autocorrelations among observations are modeled by construction in HMM. It may happen that none of the pre-determined explanatory processes are related to the larva settlement, with the actual driving process being unmeasured. The decoded hidden state chain in this case could inspire further research into this driving process. Another advantage of HMM is that the states in the state chain are correlated because of the property of Markov process and in reality, the ecological driving processes are also likely to be correlated.

To use HMM to answer the association question, we first estimate the parameters of the HMM. Then, based on these estimated parameters, we decode the most likely hidden state sequence. Final, we compare the decoded sequence with the three pre-determined oceanographic process chains, which are coded as binary sequences, to see how well

they correspond. The more agreement between the decoded state chain and a particular process, the more likely the process is associated with larvae settlement.

3.4 Methodology

3.4.1 Traditional Hidden Markov Model

HMM are stochastic processes in which a latent process controls correlation between observations and a group of densities generate actual observations (Cappè et al., 2005). The latent process is a discrete Markov chain where the current state depends on the previous states only through the most recent one. At each time point, an observation depends only on the current state of the Markov process. We see the observations from emission distributions, but the state chain is hidden.

Transitions among states in a Markov process are characterized by the transition matrix, $\mathbf{A}$, where $[a_{ij}]$ represents the probability of transitioning from state i to state j in one step. An observation and the current state of the chain are related through the emission distributions. In the case of discrete emission densities, individual emission probabilities are summarized in a matrix, $\mathbf{B}$, where $[b_{ij}]$ represents the probability of emitting the j^{th} signal at the i^{th} state. When there are uncountable signal candidates, an observation at time t, say y_t, is consider to be a random draw from a probability density function (pdf), b_i, where i is determined by the current state of the chain.

The number of states in the Markov chain typically takes on a finite or countable number of possible values (Ross, 2010), $state1$, $state2$, $\cdots$. The initial probability of

being in the i^{th} state is denoted by π_i with $\sum_i \pi_i = 1$. For the purpose of our work, we consider HMM with two states and hence, $\pi = (\pi_1, \pi_2)$. Table 3.1 gives our notation for the HMM with a two-state Markov process and BEZI emission distributions. In this table $S_1, \cdots, S_T$ denote the state chain, and $y_1, \cdots, y_T$ denote the observed data. When **A**, **B** (or b_1 and b_2) and π are specified, we say the HMM is specified (Rabiner, 1989).

Table 3.1: Notation for a two-state zero-inflated Beta Hidden Markov Model, S_t denotes the state of the chain at time t, y_t denotes the signal chain at time t

Parameter	Notation	Comments
Transition Matrix	$\mathbf{A} = \begin{matrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{matrix}$	$a_{ij} = P(S_t = j \mid S_{t-1} = i)$ for $i, j = 1,2$ and all t $a_{11} + a_{12} = 1$ and $a_{21} + a_{22} = 1$ $\mathbf{a} = (a_{11}, a_{12}, a_{21}, a_{22})$
Emission Density	$\mathbf{B} = \begin{matrix} b_1(y) \\ b_2(y) \end{matrix}$	$b_i(y)$ is the pdf of $BEZI(p_i, \mu_i, \phi_i)$ for $i = 1,2$ $f_{Y_t \mid S_t = i}(y_t) = b_i(y_t)$ for all t $\mathbf{b} = (p_1, \mu_1, \phi_1, p_2, \mu_2, \phi_2)$
Initial Probability	$\pi' = \begin{matrix} \pi_1 \\ \pi_2 \end{matrix}$	$\pi_i = P(S_1 = i)$ for $i = 1,2$; $\pi_1 + \pi_2 = 1$

There are three fundamental problems in HMM: evaluation, decoding and estimation. In evaluation, a specific HMM (i.e., **A**, **B** and π) is assumed and one evaluates the probability of getting a particular sequence of observations under this model. The forward-backward algorithm (Baum and Eagon, 1967) was developed for solving this problem. Decoding involves determination of a best sequence of hidden states given a particular HMM and a sequence of observations. The Viterbi algorithm (Viterbi, 1967) is a common tool for this problem. The posterior decoding method (Durbin et al., 1998; Nilsson, 2005) is also sometimes adopted. Finally, in estimation, also called learning or training (Phil, 2004; Petrushin, 2000), given observed data, parameters of the HMM

are estimated to best account for these observations (Rabiner, 1989; Starner and Pentland, 1995). Several methods are developed for estimation, including a HMM version of the EM algorithm, the Baum-Welch algorithm (Baum et al., 1970; Baum, 1972), the Baldi-Chauvin algorithm (Baldi et al., 1994; Baldi and Chauvin, 1994) and Bayesian approach (Robert et al., 1993; Scott, 2002). These standard procedures for the common discrete and continuous HMM can be found in many HMM tutorials such as Rabiner (1989) and Petrushin (2000). However, none of this work deals with HMM whose emission densities are a mixture of discrete and continuous distributions, such as the BEZI. More recently, Turner (2008) gives a method that maximizes the likelihood of the HMM directly in the case of discrete emission distributions.

3.4.2 Model Specification

We illustrate the structure of a two-state HMM with the BEZI emission densities (BEZI-HMM) in Figure 3.2, for the first five time points. The solid arrows give the actual path of the Markov process. The dashed arrows give other possible transitions between the two states. The thick vertical arrows illustrate the connections between observations, y_1, $\cdots$, y_5, and their emission states. A similar illustration can be found in Churbanov and Winters-Hilt (2008), it is for a discrete density HMM.

$\mathbf{Y} = (Y_1, \cdots, Y_T)$ is the sequence of observations and $\mathbf{S} = (S_1, \cdots, S_T)$ is the latent sequence of Markov states. In the two-state BEZI-HMM, realizations of $\mathbf{Y}$ are values in the interval [0, 1), and $\mathbf{S}$ is a sequence of 1's and 2's.

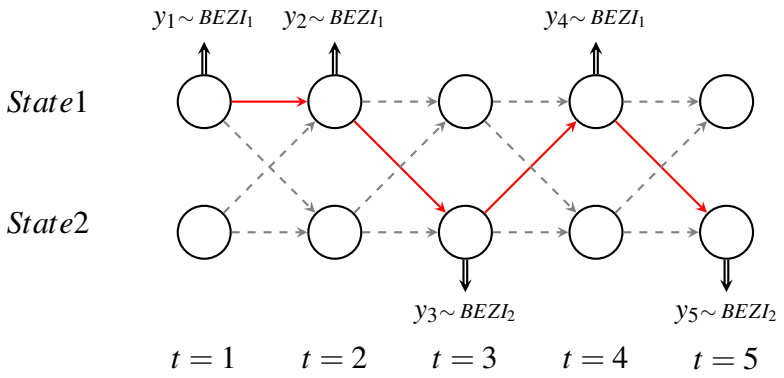


Figure 3.2: Illustration of two-state zero-inflated Beta Hidden Markov Model for five time points.

With $\theta = (\pi, \mathbf{a}, \mathbf{b})$, the complete likelihood function of the HMM is

$$L(\mathbf{Y}, \mathbf{S}; \theta) = L(\mathbf{Y}|\mathbf{S}; \theta) L(\mathbf{S}; \theta). \tag{3.1}$$

First consider $L(\mathbf{S}; \theta)$. Because $\mathbf{S}$ consists only of 1's or 2's, $L(\mathbf{S}; \theta)$ is the joint probability mass function (pmf) of $\mathbf{S}$, which we will write as $P_\theta(\mathbf{S})$, or simply $P(\mathbf{S})$ when there is no confusion. First

$$P(S_1) = \pi_1 I(S_1 = 1) + \pi_2 I(S_1 = 2) = \prod_{i=1}^{2} \pi_i^{I(S_1 = i)}.$$

Then, using the notation in Table 3.1, the conditional pmf of $S_t | S_{t-1}$ is

$$P(S_t|S_{t-1}) = a_{11} I(S_t = 1, S_{t-1} = 1) + a_{12} I(S_t = 2, S_{t-1} = 1) + a_{21} I(S_t = 1, S_{t-1} = 2)$$

$$+ a_{22} I(S_t = 2, S_{t-1} = 2) = \prod_{i,j=1}^{2} a_{ij}^{I(S_t = j, S_{t-1} = i)}.$$

Using conditional probability and the Markov property,

$$P(\mathbf{S}) = P(S_1, S_2, \cdots, S_T) = P(S_T|S_1, S_2, \cdots, S_{T-1})P(S_1, S_2, \cdots, S_{T-1})$$

$$= P(S_T|S_{T-1})P(S_1, S_2, \cdots, S_{T-1}) = \cdots = \prod_{t=2}^{T-1} P(S_{t+1}|S_t)P(S_1, S_2)$$

$$= \prod_{t=1}^{T-1} P(S_{t+1}|S_t) \cdot P(S_1)$$

$$= \left[\prod_{t=1}^{T-1}\prod_{i,j=1}^{2} a_{ij}^{I(S_{t+1}=j, S_t=i)}\right]\left[\prod_{i=1}^{2}\pi_i^{I(S_1=i)}\right]. \tag{3.2}$$

Equation (3.2) gives the expression for $L(\mathbf{S};\theta)$.

Next consider $L(\mathbf{Y}|\mathbf{S};\theta)$. As defined in Table 3.1, we have $f_{Y_t|S_t=i}(y_t) = b_i(y_t)$ for $t = 1, \cdots, T$ and $i = 1, 2$, in which

$$b_i(y) = p_i^{I(y=0)}(1-p_i)^{(1-I(y=0))} \quad \frac{\Gamma(\phi_i)}{\Gamma(\mu_i\phi_i)\Gamma((1-\mu_i)\phi_i)} y^{\mu_i\phi_i - 1}(1-y)^{(1-\mu_i)\phi_i - 1}{}^{(1-I(y=0))},$$

and

$$f_{Y_t|S_t}(y_t) = f_{Y_t|S_t=1}(y_t)I(S_t = 1) + f_{Y_t|S_t=2}(y_t)I(S_t = 2)$$

$$= b_1(y_t)I(S_t = 1) + b_2(y_t)I(S_t = 2) = \prod_{i=1}^{2} b_i(y_t)^{I(S_t=i)}.$$

Again, using conditional probability we have

$$L(\mathbf{Y}|\mathbf{S};\theta) = f_{\mathbf{Y}|\mathbf{S}}(y_1, \cdots, y_T;\theta) = f_{Y_T|Y_1,\cdots,Y_{T-1},\mathbf{S}}(y_T;\theta) f_{Y_1,\cdots,Y_{T-1}|\mathbf{S}}(y_1, \cdots, y_{T-1};\theta).$$

Based on the definition of HMM, for $t \in \{2, \cdots, T\}$,

$$f_{Y_t|Y_1,\cdots,Y_{t-1},S_1,\cdots,S_t}(y_t;\theta) = f_{Y_t|S_t}(y_t;\theta)$$

and

$$f_{Y_1,\cdots,Y_{t-1}|S_1,\cdots,S_{t-1},S_t}(y_1,\cdots,y_{t-1};\theta) = f_{Y_1,\cdots,Y_{t-1}|S_1,\cdots,S_{t-1}}(y_1,\cdots,y_{t-1};\theta)$$

since S_t comes after y_{t-1}. Consequently,

$$L(\mathbf{Y}|\mathbf{S};\theta) = f_{Y_T|S_T}(y_T;\theta) f_{Y_1,\cdots,Y_{T-1}|S_1,\cdots,S_{T-1}}(y_1,\cdots,y_{T-1};\theta)$$

$$= f_{Y_T|S_T}(y_T;\theta) f_{Y_{T-1}|S_{T-1}}(y_{T-1};\theta) f_{Y_1,\cdots,Y_{T-2}|S_1,\cdots,S_{T-2}}(y_1,\cdots,y_{T-2};\theta)$$

$$= \cdots = \prod_{t=1}^{T} f_{Y_t|S_t}(y_t;\theta)$$

$$= \prod_{t=1}^{T}\prod_{i=1}^{2} b_i(y_t)^{I(S_t=i)}. \tag{3.3}$$

By combining (3.2) and (3.3), Equation (3.1) can be written:

$$L(\mathbf{Y},\mathbf{S};\theta) = \prod_{t=1}^{T}\prod_{i=1}^{2} b_i(y_t)^{I(S_t=i)} \prod_{t=1}^{T-1}\prod_{i,j=1}^{2} a_{ij}^{I(S_{t+1}=j,S_t=i)} \prod_{i=1}^{2} \pi_i^{I(S_1=i)}. \tag{3.4}$$

The log-likelihood function, $l(\mathbf{Y},\mathbf{S};\theta) = log\,(L(\mathbf{Y},\mathbf{S};\theta))$, is then written as:

$$l(\mathbf{Y},\mathbf{S};\theta) = \sum_{t=1}^{T}\sum_{i=1}^{2} I(S_t = i) log\,(b_i(y_t))$$

$$+\sum_{t=1}^{T-1}\sum_{i,j=1}^{2} I(S_{t+1}=j,S_t=i)log(a_{ij})+\sum_{i=1}^{2} I(S_1=i)log(\pi_i). \tag{3.5}$$

Equation (3.5) is the sum of three components:

$$Q_{\mathbf{B}}(\mathbf{b})=\sum_{t=1}^{T}\sum_{i=1}^{2} I(S_t=i)log\left(b_i(y_t)\right),$$

$$Q_{\mathbf{A}}(\mathbf{a})=\sum_{t=1}^{T-1}\sum_{i,j=1}^{2} I(S_{t+1}=j,S_t=i)log(a_{ij})$$

and

$$Q_{\pi}(\pi)=\sum_{i=1}^{2} I(S_1=i)log(\pi_i).$$

$Q_{\mathbf{B}}(\mathbf{b})$ involves the BEZI densities, $Q_{\mathbf{A}}(\mathbf{a})$ involves the transition matrix and $Q_{\pi}(\pi)$ involves the initial state probabilities.

3.4.3 A Review of Standard Methods

3.4.3.1 Forward-Backward Algorithm

As mentioned in Section 3.4.1, the first problem in HMM is to evaluate the probability of getting a particular sequence of observations given the HMM parameters. The forward-backward algorithm (Baum and Eagon, 1967) provides a solution to this problem. To understand the algorithm, consider a two-state CDHMM that is comparable to our BEZI-HMM. Following Rabiner (1989), the "forward" variables are

$$\alpha_t(i)=f_{Y_1,Y_2,\ldots,Y_t,S_t=i}(y_1,\cdots,y_t;\theta), \text{ for } t=1,\cdots,T, i=1,2.$$

These are the probabilities of having a partially observed sequence, $y_1, y_2, \cdots, y_t$, and being in state i at time t, given θ. The "backward" variables are

$$\beta_t(i) = f_{Y_{t+1},\cdots,Y_T,S_t=i}(y_{t+1},\cdots,y_T;\theta), \text{ for } t = 1,\cdots,T, i = 1,2.$$

These are the probabilities of having a partially observed sequence, $y_{t+1},\cdots,y_T$, and being in state i at time t, given θ. We can calculate both sets of variables recursively.

To solve $\alpha_t(i)$, first initialize

$$\alpha_1(i) = \pi_i b_i(y_1) \text{ for } i = 1,2,$$

and then inductively,

$$\alpha_{t+1}(j) = \sum_{i=1}^{2} \alpha_t(i) a_{ij} \quad b_j(y_{t+1}) \text{ for } 1 \leq t \leq T-1, j = 1,2.$$

To solve $\beta_t(i)$, first initialize $\beta_T(i) = 1$ for $i = 1,2$, and then inductively,

$$\beta_t(i) = \sum_{j=1}^{2} a_{ij} b_j(y_{t+1}) \beta_{t+1}(j) \text{ for } t = T-1, T-2, \cdots, 1, i = 1,2.$$

The forward-backward solution for evaluation is:

$$f_{\mathbf{Y}}(y_1,\cdots,y_T;\theta) = f_{\mathbf{Y},S_T=1}(y_1,\cdots,y_T;\theta) + f_{\mathbf{Y},S_T=2}(y_1,\cdots,y_T;\theta)$$

$$= \sum_{i=1}^{2} \alpha_T(i) = \sum_{i=1}^{2} \beta_1(i)\alpha_1(i).$$

That is, the last two sums give the probability of any sequence of observations. Strictly speaking, we do not need to have backward variables to solve the evaluation problem. However, we introduce them here since they will be used in solving the third HMM problem, estimation. For the BEZI-HMM evaluation problem, one replaces the general form of the emission distribution, $b_i(y)$, by the corresponding BEZI density.

3.4.3.2 Viterbi Algorithm and Posterior Decoding

Consider the observed sequence, $\mathbf{Y}$, where the HMM parameters, θ, are known. The decoding question involves determining the optimal state sequence. Depending on how we define optimality, there are several solutions. The most common optimality criterion is to find the most likely state sequence that emits the observed sequence, $\mathbf{Y}$. More precisely, within all possible state sequences, we find the one with the highest probability of emitting the entire observed chain. Let $\mathbf{y}$ be a realization of $\mathbf{Y}$. Then mathematically, we need

$$\arg\max_{\mathbf{S}} L(\mathbf{S}|\mathbf{y};\theta).$$

The solution to this maximization problem is the Viterbi algorithm (Viterbi, 1967) where the most likely state chain is called a Viterbi path.

Because $L(\mathbf{S},\mathbf{y};\theta) = L(\mathbf{S}|\mathbf{y};\theta)L(\mathbf{y};\theta)$, when $\mathbf{y}$ and θ are known, maximizing $L(\mathbf{S}|\mathbf{y};\theta)$ is equivalent to maximizing $L(\mathbf{S},\mathbf{y};\theta)$. Following the notation in Rabiner (1989) and still considering a two-state HMM, we define

$$\delta_t(i) = \max_{S_1,\cdots,S_{t-1}} L(S_1,\cdots,S_{t-1},S_t = i,y_1,\cdots,y_t;\theta) \text{ for } i = 1,2.$$

Here $\delta_t(i)$ is the largest probability (likelihood) of being in state i at time t, after accounting for the first t observations along a single path. Furthermore,

$$\delta_{t+1}(j) = \max_{i=1,2}\{\delta_t(i)a_{ij}\}b_j(y_{t+1}) \text{ for } j = 1,2$$

For each $t = 1,\cdots,T$ and $i = 1,2$, we keep a record of the state path that maximizes $\delta_t(i)$ in two vectors, $\psi_t(1)$ and $\psi_t(2)$.

We now express the algorithm as follows:

Initialization : For i=1 and 2, $\delta_1(i) = \pi_i b_i(y_1)$ and $\psi_1(i) = 0$ because there is no history of state at $t = 1$

Recursion : For $2 \leq t \leq T$ and $j = 1,2$,

$$\delta_t(j) = \max_{i=1,2}\{\delta_{t-1}(i)a_{ij}\}b_j(y_t)$$

$$\psi_t(j) = \arg\max_{i=1,2}\{\delta_{t-1}(i)a_{ij}\}$$

Termination :

$$L^* = \max_{i=1,2}\{\delta_T(i)\}$$

and

$$S_T^* = \arg\max_{i=1,2}\{\delta_T(i)\}$$

Path backtracking : For $t = T-1, T-2, \cdots, 1$, $S_t^* = \psi_{t+1}(S_{t+1}^*)$. The output of the algorithm, $S_1^*,\cdots,S_T^*$, is the Viterbi path

Figure 3.3 gives an illustration of the Viterbi algorithm with five time points. The dashed lines give all possible paths, the solid black arrows indicate the $\psi_t(i)$'s, and the double arrows give the Viterbi path as $\mathbf{S}^* = (state2, state2, state2, state2, state2)$. A similar illustration for a three-state HMM can be found in Nilsson (2005), page 37.

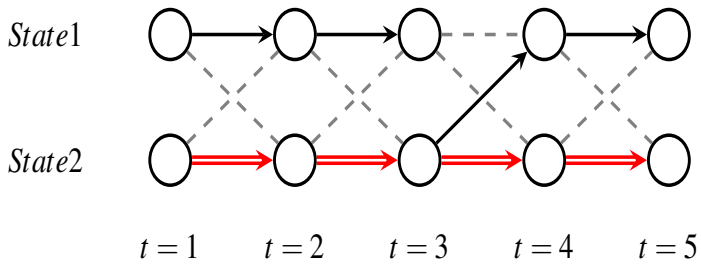


Figure 3.3: Illustration of the Viterbi algorithm: the total number of paths is $2^5 = 32$, the number of candidate paths in the Viterbi algorithm is 2.

Similar to the forward-backward algorithm, the justification for the Viterbi algorithm is the same for all emission densities. For BEZI-HMM, we can simply replace the $b_i(y)$'s by BEZI densities. Ross (2010) provides a justification for the applicability of the Viterbi algorithm.

An alternative to the Viterbi algorithm for the decoding problem is known as posterior decoding (Durbin et al., 1998). This method is based on a different optimality criterion. At each time point, the conditional probability of a particular state given observations and parameters, $P(S_t = i|\mathbf{y};\theta)$, is maximized. In other words, posterior decoding only uses marginal information. The solution chain may be invalid under certain transition matrices, (i.e., some transitions are impossible because of zero transition probability (Nilsson, 2005)). This decoding method is a sub-product of the EM algorithm, and as a result, we can get the decoded state chain when we solve the estimation

problem (Rabiner, 1989; Nilsson, 2005).

3.4.3.3 Baum-Welch Algorithm

The third general question in HMM is estimating the model parameters (θ) given an observed sequence, $\mathbf{Y}$. Because the state sequence, $\mathbf{S}$, is missing (or hidden), the EM algorithm (Dempster et al., 1977), in particular, the Baum-Welch algorithm (Baum et al., 1970; Baum, 1972) is often applied. The general Baum-Welch algorithm for strictly log-concave density functions, elliptically symmetric density functions, and mixtures and products of mixtures of strictly log-concave and/or elliptically symmetric densities can be found in Baum et al. (1970), Liporace (1982), Juang et al. (1986), and Rabiner et al. (1985), respectively. However, BEZI falls in none of these categories. Fortunately, we are able to prove that the EM algorithm update indeed increases the likelihood function and is able to achieve at least local maximum points for BEZI-HMM. We provide more details about the EM algorithm for BEZI-HMM along with the justification of its applicability in section 3.4.4.

3.4.4 BEZI-HMM Estimation

3.4.4.1 BEZI-HMM EM Algorithm

The utility of the EM algorithm (Dempster et al., 1977) is in finding maximum likelihood estimates given incomplete data. In our situation, the complete data are $(\mathbf{Y}, \mathbf{S})$, though we only observe $\mathbf{Y}$. We determine maximum likelihood estimates of θ by maximizing

the marginal likelihood function

$$L(\theta;\mathbf{y}) = \int L(\theta;\mathbf{y},\mathbf{S})d\mu(\mathbf{S}) \tag{3.6}$$

where, in the HMM framework, $\mu(\mathbf{S})$ is the discrete measure of the states $\mathbf{S}$. Equivalently, we maximize the log-likelihood $l(\theta;\mathbf{y}) = log(L(\theta;\mathbf{y}))$.

The EM algorithm maximizes $l(\theta;\mathbf{y})$ indirectly by iteratively maximizing the complete log-likelihood function $log(L(\theta;\mathbf{S},\mathbf{y}))$ (McLachlan and Krishnan, 2008). Because $\mathbf{S}$ are unobservable, we replace the components that are involved in $\mathbf{S}$ in $log(L(\theta;\mathbf{S},\mathbf{y}))$ by their conditional expectations using the current estimate of θ. More specifically, given the observations ($\mathbf{y}$) and current step of parameter estimates ($\theta^{(k)}$), define

$$Q(\theta;\theta^{(k)}) = \int log(L(\theta;\mathbf{S},\mathbf{y}))L(\theta^{(k)};\mathbf{S},\mathbf{y})d\mu(\mathbf{S}) = E_{\mathbf{S}}\left[log(L(\theta;\mathbf{S},\mathbf{y}))|\theta^{(k)}\right]$$

for all θ and $\theta^{(k)}$ in the parameter space (Baum et al., 1970; Jamshidian and Jennrich, 2000; Cappè et al., 2005). Then the algorithm maximizes $Q(\theta;\theta^{(k)})$ to obtain $\theta^{(k+1)}$. The process is repeated until a stopping rule is satisfied.

For BEZI-HMM, we give $log(L(\theta;\mathbf{y},\mathbf{S}))$ in Section 3.4.2 as Equation (3.5). Therefore

$$Q(\theta;\theta^{(k)}) = E_{\mathbf{S}}\left[log(L(\theta;\mathbf{S},\mathbf{y}))|\theta^{(k)}\right] = E_{\mathbf{S}}\left[Q_{\mathbf{B}}(\mathbf{b})|\theta^{(k)}\right] + E_{\mathbf{S}}\left[Q_{\mathbf{A}}(\mathbf{a})|\theta^{(k)}\right]$$

$$+E_{\mathbf{S}}\left[Q_{\pi}(\pi)|\theta^{(k)}\right].$$

The algorithm iterates between the E-step and the M-step to maximize these three components.

Expectation step (E-step) : Given $\mathbf{y}$ and current estimates $\theta^{(k)}$, we replace the indicators involved in the missing data ($\mathbf{S}$) in $Q_{\mathbf{B}}(\mathbf{b})$, $Q_{\mathbf{A}}(\mathbf{a})$ and $Q_{\pi}(\pi)$ by their expectations. That is, replace $I(S_t = i)$, $I(S_{t+1} = j, S_t = i)$ and $I(S_1 = i)$ by, respectively,

$$E(I(S_t = i)|\mathbf{y}, \theta^{(k)}) = P(S_t = i|\mathbf{y}, \theta^{(k)}), \tag{3.7}$$

$$E(I(S_{t+1} = j, S_t = i)|\mathbf{y}, \theta^{(k)}) = P(S_{t+1} = i, S_t = j|\mathbf{y}, \theta^{(k)}), \tag{3.8}$$

$$E(I(S_1 = i)|\mathbf{y}, \theta^{(k)}) = P(S_1 = i|\mathbf{y}, \theta^{(k)}). \tag{3.9}$$

We can obtain the probabilities on the right hand side of (3.7), (3.8) and (3.9) using the forward and backward variables defined in Section 3.4.3.1. To be more specific, the probability of being in state i at time t is

$$\gamma_t(i) = P(S_t = i|\mathbf{y}, \theta) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{l=1}^{2} \alpha_t(l)\beta_t(l)}.$$

The probability of being in state i at time t and state j at time $t+1$ is $\xi_t(i, j)$ for $t = 1, \cdots, T-1$, with

$$\xi_t(i, j) = P(S_{t+1} = j, S_t = i|\mathbf{y}, \theta) = \frac{\alpha_t(i)a_{ij}b_j(y_{t+1})\beta_{t+1}(j)}{\sum_{m=1}^{2}\sum_{l=1}^{2} \alpha_t(m)a_{ml}b_l(y_{t+1})\beta_{t+1}(l)}.$$

Therefore, we can replace the probabilities on the right hand sides of (3.7), (3.8), (3.9) by $\gamma_t(i)$, $\xi_t(i,j)$ and $\gamma_1(i)$ respectively.

Maximization step (M-step) : In M-step, given the expectations in (3.7), (3.8) and (3.9), we maximize $Q(\theta;\theta^{(k)})$ to obtain updated estimates.

We now summarize the entire algorithm as follows:

1. Obtain an initial value $\theta^{(0)}$
2. Given estimates, we calculate $\theta^{(k)}$ for $k = 0,1,2,\cdots$, (3.7), (3.8) and (3.9) by obtaining $\gamma_t(i)$, $\xi_t(i,j)$ and $\gamma_1(i)$. Because

$$\gamma_t(i) = \sum_{j=1}^{2} \xi_t(i,j),$$

$$\sum_{t=1}^{T-1} \gamma_t(i)^{(k)} = \text{ expected number of transitions from } i \text{ under current estimate } \theta^{(k)},$$

$$\sum_{t=1}^{T-1} \xi_t(i,j)^{(k)} = \text{ expected number of transitions from } i \text{ to } j \text{ under current estimate } \theta^{(k)}.$$

So that we update $\widehat{\pi_i}^{(k+1)}$ and $\widehat{a_{ij}}^{(k+1)}$ by

$$\gamma_1(i)^{(k)} = \text{ expected frequency in state } i \text{ at time 1 given } \theta^{(k)}$$

and

$$\frac{\sum_{t=1}^{T-1} \xi_t(i,j)^{(k)}}{\sum_{t=1}^{T-1} \gamma_t(i)^{(k)}} = \frac{\text{expected number of transitions from } i \text{ to } j \text{ given } \theta^{(k)}}{\text{expected number of transitions from } i \text{ given } \theta^{(k)}},$$

respectively.

Then we update the elements in $\mathbf{A}$, π_1 and π_2 by $\widehat{a_{ij}}^{(k+1)}$ and $\widehat{\pi}_i^{(k+1)}$ for $i, j = 1, 2$. This finishes the $(k+1)^{th}$ step of increasing $E_{\mathbf{S}}\ Q_{\mathbf{A}}(\mathbf{a})|\theta^{(k)}$ and $E_{\mathbf{S}}\ Q_{\pi}(\pi)|\theta^{(k)}|$. For $\mathbf{b}$, because there is no closed form expression, we use "Limited-memory quasi-Newton code for bound-constrained optimization method" (`L-BFGS-B`) (Byrd et al., 1995) to numerically update them by maximizing $E_{\mathbf{S}}\ Q_{\mathbf{B}}(\mathbf{b})|\theta^{(k)}$. For a comprehensive reference of relevant optimization algorithm, see Nocedal and Wright (1999)

3. Compare $\theta^{(k)}$ with $\theta^{(k+1)}$ (or $Q(\theta^{(k)}, \theta^{(k)})$ with $Q(\theta^{(k+1)}, \theta^{(k)})$). If they are close together (for example, their Euclidean distance is less than a pre-determined tolerance), stop the algorithm and report the estimates as $\theta^{(k+1)}$; otherwise return to step 2. Another way of terminating is when a pre-determined maximal number of iterations is exceeded.

3.4.4.2 Justification of the BEZI-HMM EM Algorithm

In this section, we argue that the EM algorithm in the proceeding section increases the marginal likelihood and converges to stationary points. Essentially, there are two steps to show this. First we must show that the marginal likelihood, $L(\theta; \mathbf{y})$, does not decrease after an EM iteration update. That is, $L(\theta^{(k+1)}; \mathbf{y}) \geq L(\theta^{(k)}; \mathbf{y})$ if $Q(\theta^{(k+1)}; \theta^{(k)}) \geq Q(\theta^{(k)}; \theta^{(k)})$. In other words, the EM algorithm is a monotone optimization algorithm. Second, we must show that the EM update sequence, $\{\theta^{(k)}\}_k$, converges to stationary points (θ^*) of the likelihood and hopefully one of them achieves the global maximum.

The general result for the first step has been proved by Dempster et al. (1977) (see Lemma 1 and Theorem 1 in that paper), and it can be applied to the BEZI-HMM directly[1]. The detailed steps can be found in McLachlan and Krishnan (2008) and Cappè et al. (2005) as well.

Several authors provide the second step for many common density families including the normal, Poisson, binomial and Gamma (Baum et al., 1970); however, there is no result for mixtures of discrete and continuous densities. We now develop the justification for this second step in the case of BEZI emission densities.

Although updating $Q(\theta;\theta^{(k)})$ involves simultaneously updating

$$E_{\mathbf{S}}\ \ Q_{\mathbf{B}}(\mathbf{b})|\theta^{(k)}\ \ , E_{\mathbf{S}}\ \ Q_{\mathbf{A}}(\mathbf{a})|\theta^{(k)}\ \ \text{and}\ E_{\mathbf{S}}\ \ Q_{\pi}(\pi)|\theta^{(k)}\ \ ,$$

given the current estimates, the parameters in $Q_{\mathbf{B}}(\mathbf{b})$, $Q_{\mathbf{A}}(\mathbf{a})$ and $Q_{\pi}(\pi)$ are disjoint. Increasing any of these does not depend on the others. Therefore, we can focus on $E_{\mathbf{S}}\ \ Q_{\mathbf{B}}(\mathbf{b})|\theta^{(k)}\ \ $ for our purposes since the other two parts are the same for the BEZI-HMM and for other types of HMM.

Recall that $\mathbf{b}$ represents the parameters in the BEZI emission densities. Let $\theta^{(k)}$ be the estimates of θ at step k. We need to maximize $E_{\mathbf{S}}\ \ Q_{\mathbf{B}}(\mathbf{b})|\theta^{(k)}\ \ $. Let

$$E_{\mathbf{S}}\ \ Q_{\mathbf{B}}(\mathbf{b})|\theta^{(k)}\ \ = Q(\mathbf{b},\theta^{(k)}) = \sum_{i=1}^{2}\sum_{t=1}^{T} P(S_t = i|\mathbf{y},\theta^{(k)}) log(b_i(y_t)) = \sum_{i=1}^{2} Q_i(p_i,\mu_i,\phi_i,\theta^{(k)})$$

[1]Dempster and the others give the result for regular exponential family, curve exponential family and a very general form of Q function. In the case of the last one, the requirement is (I use their notation) $Q(\phi^{'}|\phi) = E(logf(x|\phi^{'})|y,\phi)$ exist for all pairs of $(\phi^{'},\phi)$.

where $Q_i(p_i, \mu_i, \phi_i, \theta^{(k)}) = \sum_{t=1}^{T} P(S_t = i|\mathbf{y}, \theta^{(k)}) log(b_i(y_t)) = \sum_{t=1}^{T} \gamma_t^{(k)}(i) log(b_i(y_t))$ for $i = 1, 2$ with $\gamma_t^{(k)}(i)$ being calculated given $\mathbf{y}$, $\theta^{(k)}$.

If we can show $Q(\mathbf{b}, \theta^{(k)})$ is a strictly concave function of (p_i, μ_i, ϕ_i), then according to Baum et al. (1970), the EM algorithm update increases the likelihood function and will converge to a stationary point. Unfortunately, under the $BEZI(p, \mu, \phi)$ parameterization, $Q(\mathbf{b}, \theta^{(k)})$ is not strictly concave. To overcome this problem, we consider the standard parameterization for Beta distribution with two scale parameters, ν and ω. In this parameterization, the pdf of a BEZI random variable is

$$f_Y(y) = \begin{cases} p & \text{if } y = 0, \\ (1-p)\frac{\Gamma(\nu+\omega)}{\Gamma(\nu)\Gamma(\omega)} y^{\nu-1}(1-y)^{\omega-1} & \text{if } y \in (0,1), \end{cases}$$

for $\nu > 0$ and $\omega > 0$.

Given the new parameterization, for both $i = 1, 2$, we let $log(b_i(y))$'s denote the log-likelihood functions of the BEZI distribution. If we ignore the index

$$\begin{aligned} log(b(y)) &= I(y=0)log(p) + I(y=0)log(1-p) + I(y=0)log\ \frac{\Gamma(\nu+\omega)}{\Gamma(\nu)\Gamma(\omega)} \\ &+ I(y=0)(\nu-1)log(y) + I(y=0)(\omega-1)log(1-y). \end{aligned}$$

$Q(\mathbf{b}, \theta^{(k)})$ is the sum of Q_i, and each $Q_i(p_i, \nu_i, \omega_i, \theta^{(k)})$ is a linear combination of $log(b_i(y_t))$ with non-negative coefficients (as probabilities, $\gamma_t^{(k)}(i) \geq 0$). Convexity is closed under positive scaling and summation (Boyd and Lieven, 2004), and so is concavity. Therefore, if we can show that the $log(b_i(y_t))$'s are strictly concave functions of p_i, ν_i and ω_i, then we are done.

It is easy to check $log(p)$ and $log(1-p)$ are concave function of p as their second derivatives with respect to p are negative. Now the remaining question is, for fixed y, whether

$$l^*(y,\mu,\omega) = I(y=0)log\ \frac{\Gamma(\nu+\omega)}{\Gamma(\nu)\Gamma(\omega)}\ + I(y=0)(\nu-1)log(y) + I(y=0)(\omega-1)log(1-y)$$

is concave in ν and ω. Essentially, we must verify that the Hessian matrix of l^* with respect to ν and ω is negative definite. The first and second derivatives of l^* with respect to ν and ω are

$$\frac{\partial l^*}{\partial \nu} = I(y=0)\left[\varphi(\nu+\omega) - \varphi(\nu) + log(y)\right],$$

$$\frac{\partial l^*}{\partial \omega} = I(y=0)\left[\varphi(\nu+\omega) - \varphi(\omega) + log(1-y)\right].$$

$$\frac{\partial^2 l^*}{\partial \nu^2} = I(y=0)\ \ \varphi^{'}(\nu+\omega) - \varphi^{'}(\nu)\ \ ,$$

$$\frac{\partial^2 l^*}{\partial \nu \partial \omega} = \frac{\partial^2 l^*}{\partial \omega \partial \nu} = I(y=0)\varphi^{'}(\nu+\omega),$$

$$\frac{\partial^2 l^*}{\partial \omega^2} = I(y=0)\ \ \varphi^{'}(\nu+\omega) - \varphi^{'}(\omega)$$

where $\varphi(x) = \frac{dlog(\Gamma(x))}{dx} = \frac{\Gamma^{'}(x)}{\Gamma(x)}$ and $\varphi^{'}(x) = \frac{d\varphi(x)}{dx}$ are, respectively, the digamma and trigamma functions (Lehmann and Casella, 1998). We can show the Hessian matrix of l^* is negative definite as follows.

When k=1, using the trigamma expansion $\varphi^{'}(x) = \sum_{i=0}^{\infty} \frac{1}{(x+i)^2}$ (Artin, 1964), it is

easy to check

$$\frac{\partial^2 l^*}{\partial \nu^2} = \sum_{i=0}^{\infty} \frac{1}{(\nu+\omega+i)^2} - \sum_{i=0}^{\infty} \frac{1}{(\nu+i)^2} = \sum_{i=0}^{\infty} \frac{(2\nu+\omega+2i)(-\omega)}{(\nu+\omega+i)^2(\nu+i)^2} < 0.$$

When k=2,

$$\frac{\partial^2 l^*}{\partial \nu^2}\frac{\partial^2 l^*}{\partial \omega^2} - \left(\frac{\partial^2 l^*}{\partial \nu \partial \omega}\right)^2 = \varphi'(\nu)\varphi'(\omega) - \varphi'(\nu+\omega)\left(\varphi'(\nu)+\varphi'(\omega)\right). \tag{3.10}$$

We now need to show that (3.10) is strictly greater than 0. This cannot be done analytically. So we formulate this question as a linearly constrained optimization problem (Nocedal and Wright, 1999):

We minimize $f(\nu,\omega) = \varphi'(\nu)\varphi'(\omega) - \varphi'(\nu+\omega)\left(\varphi'(\nu)+\varphi'(\omega)\right)$, subject to the constraint $\nu > 0$ and $\omega > 0$. If the minimum of $f(\nu,\omega)$ is greater than 0, then (3.10) is strictly greater than 0 for all positive ν and ω. The R function `constrOptim()` is applied to obtain the minimum (R code is attached in appendix), and it is indeed positive. This means that we have stationary points for (p_i, ν_i, ω_i) for $i = 1,2$ from the EM updates.

As for the global maximality, there is no simple way to guarantee that. And as many authors point out (Turner, 2008; Aittokallio and Uusipaikka, 2000), the likelihood functions in HMM are very likely to have multiple local maxima. We can exclude saddle points by examining whether the observed information is positive definite (Aittokallio and Uusipaikka, 2000), however, which local maximum the algorithm converges to depends on the initial values. Uncertainty of achieving global maximality is not a unique problem for the EM algorithm. Multiple local maxima may complicate

any maximization technique, including the commonly used Newton algorithm or Fisher scoring method. At this point, we can only show that the maximality is located in the interior of the parameter space. It is easy to check that for fixed y, the likelihood function of BEZI ($b(y)$=$b(p,\nu,\omega;y)$) goes to 0 (so that $log(b(p,\nu,\omega;y))$ goes to $-\infty$) as p goes to 0 or 1; ν goes to 0 or $+\infty$; or ω goes to 0 or $+\infty$. This tells us the global maximality has to be achieved in the interior of the parameter space. The best possible way to increase the possibility of achieving global maximality is to start from multiple points in the parameter space and compare the maximum when the algorithm converges (Aittokallio and Uusipaikka, 2000; Turner, 2008).

Because the MLE is invariant to one-to-one transformations, the MLE of (p_i,ν_i,ω_i) gives the MLE of (p_i,μ_i,ϕ_i) for $i=1,2$. If the EM algorithm finds the MLE of (p_i,ν_i,ω_i), then it gives the MLE of (p_i,μ_i,ϕ_i) for $i=1,2$.

The justification for the applicability of the EM algorithm is complete.

There are two other ways to argue that the EM algorithm converges to stationary points for the BEZI family. Lehmann and Casella (1998) and McLachlan and Krishnan (2008) give the following theorem that provides an easily applicable condition to guarantee convergence to a stationary point.

> "If the expected complete data likelihood $Q(\theta|\theta_0,y)$ is continuous in both θ and θ_0, then all limit points of an EM sequence $\{\hat{\theta}_{(j)}\}$ are stationary points of $L(\theta|y)$, and $L(\hat{\theta}_{(j)}|y)$ converges monotonically to $L(\hat{\theta})$ for some stationary point $\hat{\theta}$."

In this contents, θ_0 is the current step estimate and θ is the free parameter to be updated. Using our notation, because we can obtain the first derivative of $Q(\theta;\theta^{(k)})$

and the existence of first derivatives guarantee continuity, the theorem may be applied directly. For both θ and θ_0 (our $\theta^{(k)}$), the continuity for $\mathbf{A}$ and π are trivial. Meanwhile, we show the first derivative for $\mathbf{b}$ in the proceeding argument; for $\mathbf{b}_0$'s (our $\mathbf{b}^{(k)}$), the derivatives of them are functions of $\gamma_t(i)$'s, $\alpha_t(i)$'s and $\beta_t(i)$'s, however, eventually they all come from the BEZI density so that the first derivatives are available.

The other useful result is with regard to regular exponential families, given by Little and Rubin (1987), page 137. It is:

> "If $f(Y|\theta)$ is a regular-exponential family and $l(\theta|y_{obs})$ is bounded, then $\theta^{(t)}$ converges to a stationary point θ^*."

Ospina and Ferrari (2010) claim that BEZI is a regular exponential family. We verify that $l(\theta|y_{obs})$ is indeed bounded from above in the Appendix.

3.4.4.3 Initial Value Selection

As mentioned in the previous discussion, the EM algorithm may converge to a local maximum. Without prior information, we suggest using a grid search method or using multiple randomly generated initial values (Aittokallio and Uusipaikka, 2000). In the first option, for every parameter in the model, start the algorithm with multiple points equally spaced within the parameter space; in the second option, start the algorithm with multiple initial values randomly generated from the parameter space. For both methods, we can obtain multiple stationary points, then the stationary point with the largest value of Q function may be viewed as "global maximum." For the BEZI parameters, we may also consider using method of moments or method of maximum likelihood to estimate

the starting values for the parameters in BEZI densities. For method of moments, we let $p_1 = p_2 = p$, the sample proportion of zeros, and solve μ and ϕ from the sample mean and sample variance of the non-zeros, letting $\mu_1 = \mu_2 = \mu$ and $\phi_1 = \phi_2 = \phi$. For both of these estimation methods, we assume there is only one BEZI population and that the samples are independent. In the case where some prior information is available (for example, a training set) more informative initial values could be obtained. One may apply global optimization technique as well. See Horst and Pardalos (1995) for a literature review of common global optimization methods.

In the next section, we describe simulation studies we preformed to evaluate our methods. But first, we would like to comment on an interesting scenario. Although $\pi_1 = \pi_2 = 0.5$ and $a_{11} = a_{12} = a_{21} = a_{22} = 0.5$ seems to be reasonable, naive candidates for the initial probabilities and the transition matrix for a two-state model when there is no prior information of the transitions, starting the EM algorithm with these probabilities is ineffective. The algorithm does not iterate at all or only a few times when we let the two emission densities have the same initial values. In other words, the algorithm is trapped in the initial point when we assume $p_1 = p_2$, $\mu_1 = \mu_2$, $\phi_1 = \phi_2$, $\pi_1 = \pi_2 = 0.5$ and a_{11}=a_{12}=a_{21}=$a_{22} = 0.5$. We suspect this awkward result is due to the fact that $\pi = (0.5, 0.5)$ is the limiting probability of $\mathbf{A} = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$. We suggest avoiding such initial values for π and $\mathbf{A}$.

3.5 Implementation and Simulation

To evaluate the performance of our BEZI-HMM model and its solutions for the standard HMM problems, we conduct two simulation studies. **Study 1** investigates the Viterbi algorithm and compares it to the posterior decoding method, and **Study 2** investigates the performance of the EM algorithm for the BEZI-HMM. The code for generating BEZI-HMM observations, estimating the parameters, and decoding the hidden state chain are all written in R.

There are four types of BEZI densities: symmetric, asymmetric, one-sided and bimodal (Cribari-Neto and Vasconcellos, 2002) as shown in Figure 3.4. For a two-state HMM, there are ten possible shape pair combinations that we can control using the parameters of the BEZI density.

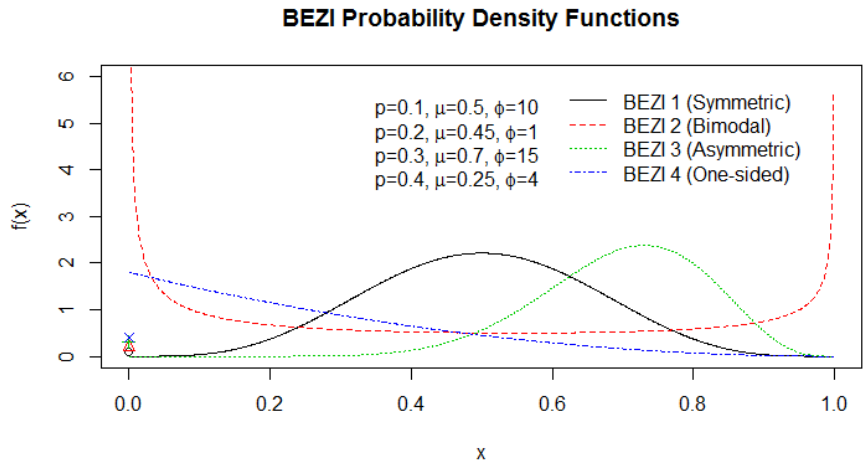


Figure 3.4: Four possible BEZI probability density plots

For both studies, we generate 1000 two-state BEZI-HMM sequences, each with 100

observations. More specifically, for each sequence, we first draw a random number from $uniform(0,1)$ and compare it to the initial probability of being in state 1, π_1, to determine whether the state is 1 or 2 at time point 1. Then the first observation is generated from the BEZI distribution that corresponds to the first state. For the succeeding time points, we generate the state at time t, S_t, by comparing a random number generated from $uniform(0,1)$ to the first element of a particular row of a two-row matrix; the row to compare to depends on the state at the previous time, S_{t-1}. When the random number is greater than that element, $S_t = 1$, otherwise $S_t = 2$[2]. The reminding observations are then generated from the BEZI densities that correspond to the state sequence.

3.5.1 Study 1

In this study, we consider how the values of the HMM parameters influence the performance of the Viterbi algorithm and posterior decoding methods. Specifically, we want to learn how the transition matrix (A) and the BEZI parameters ($\mathbf{b}$) affect our ability to decode the state chain. In these simulations, we ignore initial probabilities because these probabilities are used only once at the beginning of the process, and their effects, if any, are washed out.

We use the following performance metrics:

[2]For HMM having more than two-state, we compare the random number to a particular cumulative row sum of the transition matrix; the row to compare to depends on the state at the previous time point, S_{t-1}. When the random number is greater than the $(i-1)^{th}$ cumulative sum but less than the i^{th} cumulative sum, $S_t = i$.

Accuracy rate of individual state:

$$\text{Accuracy rate 1 } (AR_1) = \frac{\text{number of correctly decoded states}}{\text{number of total states}}$$

Accuracy rate of sequential state:

$$\text{Accuracy rate 2 } (AR_2) = \frac{\text{number of correctly decoded sequential states}}{\text{number of total sequential states}},$$

where sequential states are defined as pairs of $\{S_{t-1}, S_t\}$ for $t = 2, \cdots, T$

AR_1 evaluates the accuracy of identifying individual states, whereas AR_2 evaluates the accuracy of state identification while considering the Markov property in the state chain. To obtain a correct pair of states $\{S_{t-1}, S_t\}$, an algorithm has to have correctly decoded both S_{t-1} and S_t.

For both AR_1 and AR_2, we look at the average rates over 1000 trials, as well as their standard deviations, and 25%, 50% and 75% (Q_1, Q_2 and Q_3) quantiles.

In addition to the ten BEZI density shape pairs, we consider two transition matrices:

$$A_1 = \begin{bmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{bmatrix} \text{ and } A_2 = \begin{bmatrix} 0.2 & 0.8 \\ 0.7 & 0.3 \end{bmatrix}.$$

A HMM with transition matrix A_1 has a high probability of staying in the same state, whereas one with A_2 has large probability of moving between states. For all simulations, the initial probabilities are $\pi_1 = 0.4$ and $\pi_2 = 0.6$.

Tables 3.2 and 3.3 give the simulation results for the Viterbi algorithm and posterior

decoding method, respectively. We summarize the results as follows:

Table 3.2: Summary statistics (mean, standard deviation, 25%, 50% and 75% quantiles (Q_1, Q_2 and Q_3)) for the Viterbi algorithm accuracy rates (AR_1 and AR_2) under simulated data with BEZI-HMM($\mathbf{A}_i, p_1, \mu_1, \phi_1, p_2, \mu_2, \phi_2), i = 1, 2$; based on 1,000 trials. Sym, Asym, Bi and One stand for symmetric, asymmetric, bimodal and one-sided density shape, respectively.

Settings			AR_1 (%)					AR_2 (%)				
Density	$BEZI(p_i,\mu_i,\phi_i)$	A	Mean	SD	Q_1	Q_2	Q_3	Mean	SD	Q_1	Q_2	Q_3
Sym	(0.2, 0.5, 18)	A_1	79.55	5.33	76.00	80.00	83.00	67.71	7.44	62.63	67.68	72.73
Asym	(0.1, 0.3, 10)	A_2	79.96	5.21	77.00	80.00	83.00	68.21	7.34	63.64	68.69	72.73
Asym	(0.1, 0.3, 10)	A_1	82.20	4.88	79.00	82.00	86.00	71.40	7.01	66.67	71.72	76.77
Bi	(0.2, 0.5, 1)	A_2	79.70	5.45	76.00	80.00	83.00	68.30	7.50	63.64	68.69	73.74
Bi	(0.1, 0.3, 1)	A_1	65.89	7.39	61.00	66.00	71.00	53.25	8.64	47.47	53.54	58.59
Bi	(0.2, 0.5, 1.5)	A_2	66.49	7.06	62.00	67.00	71.00	53.08	8.24	47.47	53.54	58.59
One	(0.2, 0.25, 4)	A_1	69.41	6.94	65.00	69.50	74.00	56.64	8.39	50.51	56.57	62.63
Asym	(0.1, 0.3, 10)	A_2	65.50	7.25	61.00	66.00	70.00	52.07	8.53	46.46	52.53	57.58
One	(0.2, 0.25, 4)	A_1	81.05	5.25	77.00	81.00	85.00	69.92	7.34	64.65	69.70	74.75
Sym	(0.1, 0.5, 10)	A_2	78.98	5.57	75.00	79.00	83.00	67.14	7.72	62.63	67.68	72.73
One	(0.1, 0.25, 4)	A_1	72.30	6.83	68.00	73.00	77.00	59.74	8.47	54.55	60.61	65.66
Bi	(0.2, 0.5, 1)	A_2	73.00	6.40	69.00	73.00	78.00	60.13	8.03	54.55	60.61	65.66
Sym	(0.1, 0.5, 40)	A_1	61.99	8.19	57.00	62.00	68.00	49.52	9.22	43.43	49.49	55.56
Sym	(0.2, 0.5, 18)	A_2	61.52	7.39	57.00	61.00	66.25	48.05	8.17	43.43	47.47	53.54
Asym	(0.1, 0.3, 10)	A_1	95.01	2.40	93.00	95.00	97.00	90.67	4.30	87.88	90.91	93.94
Asym	(0.2, 0.8, 20)	A_2	95.47	2.29	94.00	96.00	97.00	91.57	4.08	88.89	91.92	93.94
One	(0.1, 0.1, 7)	A_1	71.86	6.47	68.00	72.00	76.00	59.25	7.92	53.54	58.59	64.65
One	(0.2, 0.25, 4)	A_2	70.80	6.62	66.00	71.00	76.00	57.52	8.18	51.52	57.58	63.64
Sym	(0.1, 0.5, 40)	A_1	89.62	3.35	88.00	90.00	92.00	81.64	5.56	77.78	81.82	85.86
Bi	(0.2, 0.25, 1)	A_2	89.77	3.41	88.00	90.00	92.00	81.72	5.63	78.54	81.82	85.86

- Comparing the Viterbi algorithm and posterior decoding:

  The Viterbi algorithm is overwhelmingly superior to posterior decoding. The averages of AR_1 and AR_2 for the Viterbi algorithm are 76.50% and 65.38%, respectively, whereas the averages of AR_1 and AR_2 for posterior decoding are 50.06% and 24.91%, respectively.

- Comparing AR_1 and AR_2:

  For both decoding methods, AR_1's are higher than AR_2's with smaller standard deviation in all settings. This is as expected because a correctly identified sequence

Table 3.3: Summary statistics (mean, standard deviation, 25%, 50% and 75% quantiles (Q_1, Q_2 and Q_3)) for posterior decoding accuracy rates (AR_1 and AR_2) under simulated data with BEZI-HMM($\mathbf{A}_i, p_1, \mu_1, \phi_1, p_2, \mu_2, \phi_2), i = 1, 2$; based on 1,000 trials. Sym, Asym, Bi and One stand for symmetric, asymmetric, bimodal and one-sided density shape, respectively.

Settings			AR_1 (%)					AR_2 (%)				
Density	$BEZI(p_i, \mu_i, \phi_i)$	A	Mean	SD	Q_1	Q_2	Q_3	Mean	SD	Q_1	Q_2	Q_3
Sym	(0.2, 0.5, 18)	A_1	49.74	4.20	47.00	50.00	53.00	20.03	4.84	17.17	20.20	23.23
Asym	(0.1, 0.3, 10)	A_2	50.50	6.34	46.00	51.00	55.00	29.97	6.77	25.25	30.30	34.34
Asym	(0.1, 0.3, 10)	A_1	49.95	4.06	47.00	50.00	52.00	19.87	4.57	16.92	20.20	23.23
Bi	(0.2, 0.5, 1)	A_2	50.48	6.28	46.00	51.00	55.00	30.03	6.71	25.25	30.30	34.34
Bi	(0.1, 0.3, 1)	A_1	49.68	4.02	47.00	50.00	52.00	19.86	4.56	16.16	19.19	23.23
Bi	(0.2, 0.5, 1.5)	A_2	50.80	6.04	47.00	51.00	55.00	30.30	6.45	26.26	30.30	34.34
One	(0.2, 0.25, 4)	A_1	49.64	4.17	47.00	50.00	53.00	19.86	4.69	16.16	19.70	23.23
Asym	(0.1, 0.3, 10)	A_2	50.02	6.22	46.00	50.00	54.25	29.53	6.55	25.25	29.29	34.34
One	(0.2, 0.25, 4)	A_1	49.57	4.09	47.00	50.00	52.00	19.68	4.53	17.17	19.19	22.22
Sym	(0.1, 0.5, 10)	A_2	50.42	6.01	46.00	50.00	54.00	29.96	6.44	25.25	29.29	34.34
One	(0.1, 0.25, 4)	A_1	49.90	4.07	47.00	50.00	53.00	20.05	4.62	17.17	20.20	23.23
Bi	(0.2, 0.5, 1)	A_2	50.22	6.17	46.00	50.00	55.00	29.81	6.62	25.25	29.29	34.34
Sym	(0.1, 0.5, 40)	A_1	49.62	3.09	47.00	50.00	52.00	19.87	4.36	17.17	20.20	23.23
Sym	(0.2, 0.5, 18)	A_2	50.35	6.12	46.00	50.00	55.00	29.89	6.56	25.25	29.29	34.34
Asym	(0.1, 0.3, 10)	A_1	49.70	3.95	47.00	50.00	52.00	19.94	4.48	17.17	20.20	23.23
Asym	(0.2, 0.8, 20)	A_2	50.35	6.04	46.00	50.00	54.00	29.89	6.45	25.25	29.29	34.34
One	(0.1, 0.1, 7)	A_1	49.76	3.89	47.00	50.00	52.00	19.98	4.43	17.17	20.20	23.23
One	(0.2, 0.25, 4)	A_2	50.29	6.20	46.00	50.00	55.00	29.80	6.60	25.25	29.29	34.34
Sym	(0.1, 0.5, 40)	A_1	49.63	4.04	47.00	50.00	52.25	19.85	4.65	16.16	19.19	22.22
Bi	(0.2, 0.25, 1)	A_2	50.52	6.30	46.00	51.00	55.00	30.13	7.62	25.25	30.30	34.34

of states requires two correctly identified consecutive individual states. When the accuracy rate for individual state is low, the accuracy rate for a subsequent state has to be low as well. The means and medians for AR_1 and AR_2 are close in both decoding methods.

- Comparing BEZI density shapes:

  For the Viterbi algorithm, among the ten two-density shape combinations, asymmetric vs asymmetric yields the highest accuracy rates (the average AR_1 is greater than 95% and the average AR_2 is greater than 90%[3], respectively); followed by

[3]We use similar correspondence for the following: the first number in parenthesis is the average of AR_1 and the second number is the average of AR_2.

symmetric vs bimodal ($> 89\%$ and $> 81\%$). Symmetric vs asymmetric, asymmetric vs bimodal and one-sided vs symmetric have similar rates (about 80% and 68%); one-sided vs bimodal and one-sided vs one-sided have similar rates (about 72% and 59%); bimodal vs bimodal and one-sided vs asymmetric have similar rates (about 66% and 53%). Symmetric vs symmetric is the worst case with the averaged AR_1 being less than 62% and the averaged AR_2 being less than 50%. For posterior decoding method, however, the differences are no longer striking among different density shape combinations.

- Comparing A_1 and A_2:

  On the one hand, for the Viterbi algorithm, as shown in Table 3.2, there is no clear distinction between the two transition matrices: Five out of ten BEZI parameter settings have higher averaged AR_1 under A_1, the other five have higher averages under A_2. Meanwhile, six out of ten settings have higher averaged AR_2 under A_1. On the other hand, for the posterior decoding, as shown in Table 3.3, A_2 has uniformly higher average for both AR_1 and AR_2, even though the differences in AR_1's are subtle between $\mathbf{A}_1$ and $\mathbf{A}_2$.

- Large sample size:

  Although we do not provide the results here, we had several 1000-observation chains and we found the rates for these are higher than those for the 100-observation chains within the same settings. We expect the rates to get better with the number of observations increasing.

3.5.2 Study 2

As in **Study 1**, we randomly generate artificial data from a BEZI-HMM with two states. We now evaluate the performance of the EM algorithm. This time we focus on comparing the effect of BEZI density shape combinations and initial value selection methods on the EM estimation results. For all the simulations, we use A_1 defined in **Study 1** as the transition matrix and $\pi_1 = 0.4$ and $\pi_2 = 0.6$ as the initial probabilities.

We consider three sets of initial values for the BEZI emission densities: the true parameters, method of moments estimates (MOM) and maximum likelihood estimates (MLE). In the calculation of the latter two, we assume there is only one BEZI population and all observations are independent. Initial values for transition probabilities and initial probabilities are taken to be the true values.

The estimation performance metrics we consider are bias, relative bias (bias/true), root mean squared error (RMSE) and CVRMSE (RMSE/mean). Among these, relative bias and CVRMSE are used to account for the different ranges of the parameter space: the parameter space for μ and p are bounded within 0 and 1 so they cannot vary as much as ϕ, whose parameter space is the non-negative real line. We only show the results for bimodal vs bimodal and asymmetric vs asymmetric density combinations. We summarize these in Table 3.4. We also look at several computing metrics, for example, number of iterations, computing time and the values of the maximized objective function ($Q(\theta, \theta^{(k)})$). We give these results in Table 3.5.

Again, we ignore initial probabilities because as Turner (2008) states

> "there is usually no hope of estimating the initial state distribution consistently, 'as there is only one random variable...(that is not even observed!)

> drawn from this (distribution)' ... It is interesting to note that if π is treated as a 'free' variable, the (inconsistent) maximum likelihood estimate will necessarily be obtained by setting one of the π_i equal to unity and all the others to zero. This index i to be chosen is one for which $L(y_1, \cdots, y_n|S_1 = i)$ is maximal."

This is exactly what we find in our simulations: in the output, estimates of π are either (0, 1) or (1, 0). The more common assumption is to use the limiting probabilities of the Markov chain as the initial probability (Cappè et al., 2005; Turner, 2008). For calculation of limiting probabilities of homogeneous Markov chain with two states, see Ross (2010), page 216.

We summarize our finding from **Study 2** in the following points.

- Comparing three initial values:

  MOM and MLE yield very similar results in all the estimation performance and computing metrics. This is because for a given simulation, MOM and MLE are not very far apart. When using the true population parameters as initial values, the bias and relative bias are usually smaller than when using MOM and MLE. The bias for ϕ is often large under all three initial values. For one-sided vs asymmetric, bimodal vs bimodal and symmetric vs symmetric density combinations, the true initial values sometimes have larger CVRMSE than the other two initial values. The numbers of iterations required by the EM algorithm to converge are smaller when we use the true parameters as initial values. The maximized objective function $Q(\theta, \theta^{(k)})$ usually has a larger value under true parameter initial values. The difference between MOM, MLE and the true value initials could be the a result of multiple local maxima.

Table 3.4: The EM algorithm accuracy metric summary. The length of chain is 100, the number of trials is 1000. True, MOM and MLE represent using true parameter values, method of moment estimates and maximum likelihood estimates as initial values for the algorithm, respectively. Asym and Bi stand for asymmetric and bimodal BEZI density shapes, respectively.

Density	Metric		a_{11}	a_{21}	a_{12}	a_{22}	p_1	μ_1	ϕ_1	p_2	μ_2	ϕ_2
	true	value	0.8	0.3	0.2	0.7	0.1	0.3	10	0.2	0.8	20
	bias	true	-0.004	0.009	0.004	-0.009	-0.002	0	0.860	-0.009	0	3.064
		MOM	-0.035	-0.030	0.035	0.030	0.033	0.191	5.061	-0.043	-0.201	-1.914
		MLE	-0.035	-0.030	0.035	0.030	0.033	0.190	5.127	-0.043	-0.200	-2.010
	rela-bias	true	-0.005	0.031	0.021	-0.013	-0.019	0.001	0.086	-0.043	-0.001	0.153
Asym		MOM	-0.043	-0.1	0.174	0.043	0.333	0.635	0.506	-0.215	-0.251	-0.096
vs		MLE	-0.043	-0.1	0.174	0.043	0.331	0.634	0.513	-0.217	-0.25	-0.101
Asym	RMSE	true	2.06	2.75	2.06	2.75	1.93	0.72	90.45	2.99	0.66	271.89
		MOM	2.76	3.13	2.76	3.13	2.78	7.59	277.86	3.43	7.68	302.03
		MLE	2.75	3.13	2.75	3.13	2.78	7.6	282.14	3.37	7.66	294.83
	CVRMSE	true	2.59	8.91	10.08	3.99	19.65	2.41	8.33	15.62	0.82	11.79
		MOM	3.61	11.59	11.78	4.28	20.89	15.48	18.45	21.83	12.81	16.7
		MLE	3.6	11.59	11.72	4.29	20.87	15.50	18.65	21.51	12.77	16.39
	true	value	0.8	0.3	0.2	0.7	0.1	0.3	1	0.2	0.5	1.5
	bias	true	-0.021	0.020	0.021	-0.020	0.002	-0.040	1.817	0.006	0.056	30.46
		MOM	0.022	0.017	-0.022	-0.017	0.046	0.042	0.628	-0.046	-0.073	7.846
		MLE	0.023	0.017	-0.023	-0.017	0.046	0.042	0.635	-0.047	-0.073	8.975
	rela-bias	true	-0.026	0.068	0.104	-0.029	0.019	-0.133	1.817	0.029	0.111	20.305
Bi		MOM	0.028	0.058	-0.112	-0.025	0.462	0.140	0.628	-0.232	-0.146	5.231
vs		MLE	0.028	0.058	-0.112	-0.025	0.464	0.139	0.635	-0.234	-0.146	5.983
Bi	RMSE	true	6.27	8.12	6.27	8.12	3.26	3.14	968.38	5.92	5.03	21162.25
		MOM	5.54	8.19	5.54	8.19	3.65	4.08	67.6	5.91	6.69	2342.63
		MLE	5.53	8.21	5.53	8.21	3.64	4.06	68.95	5.91	6.68	2784.22
	CVRMSE	true	8.04	25.33	28.39	11.94	32.01	12.06	343.79	28.75	9.06	662.19
		MOM	6.73	25.82	31.18	12.01	24.97	11.94	41.53	38.49	15.67	250.66
		MLE	6.73	25.86	31.18	12.02	24.89	11.9	42.18	38.55	15.64	265.80

- Comparing different BEZI density shapes:

  Among the ten BEZI density combinations, one-sided vs one-sided and bimodal vs bimodal have the largest relative bias, whereas asymmetric vs asymmetric and asymmetric vs symmetric have the smallest relative bias. Symmetric vs symmetric and one-sided vs symmetric have the third and fourth smallest relative bias. One-sided vs bimodal and bimodal vs bimodal have the largest CVRMSE. Possible reason for that is when a BEZI random variable, Y, has a one-sided or bimodal

Table 3.5: The EM algorithm computational metric summary. The length of chain is 100, the number of trials is 1000. True, MOM and MLE represent using true parameter values, method of moment estimates and maximum likelihood estimates as initial values for the algorithm, respectively. Sym, Asym, One and Bi stand for symmetric, asymmetric, one-sided and bimodal BEZI density shapes, respectively. The computing time is based on R user time in seconds, all simulations are conducted on the same PC with an Intel© Core^{TM} 2 Quad CPU Q6600 @ 2.40GHz 2.40GHz processor with 4.00 GB RAM.

Density shapes	Average iteration			Objective function			Computing time		
	true	MOM	MLE	true>MOM	true>MLE	MLE>MOM	true	MOM	MLE
Sym vs Asym	94	99	99	674	687	496	5985	6808	6819
Asym vs Bi	65	73	72	589	590	519	3714	4271	4243
Bi vs Bi	129	132	132	607	617	493	7814	7791	7782
One vs Asym	111	114	114	577	568	487	7497	7845	7828
One vs Sym	78	87	87	605	603	492	5138	5876	5885
One vs Bi	95	101	100	563	561	498	6109	6427	6403
Sym vs Sym	135	140	139	582	561	524	8103	8503	8491
Asym vs Asym	28	47	46	547	537	521	1768	3521	3519
One vs One	119	120	120	637	622	496	8548	8952	8883
Sym vs Bi	36	44	44	621	618	475	1900	2610	2620

density function, as its realization, y, approaches to 0 and/or 1, the corresponding $f(y)$ approaches to positive infinity.

The computing time and average number of iterations are large when we have symmetric vs symmetric, one-sided vs one-sided, bimodal vs bimodal and one-sided vs asymmetric. By contrast, asymmetric vs asymmetric has the smallest average number of iterations and the shortest computing time.

In summery, one-sided is the hardest shape for the EM algorithm, followed by bimodal and symmetric. Asymmetric is the easiest density shape for the algorithm to obtain convergence and obtain accurate results.

- Comparing individual parameters:

Among all the parameters we considered, the precision parameter, ϕ, has the largest bias and CVRMSE. We suspect the unbounded parameter space is the problem here. It is also interesting that ϕ_2 always has larger bias and CVRMSE than ϕ_1. We think the reason is that the transition matrix forces the HMM to have more observations from population 1 than from population 2 as the limiting probability for $\mathbf{A}_1$ is (0.6, 0.4). For the same reason, the estimates for p_1 and μ_1 usually have better properties than the estimates for p_2 and μ_2. The estimates for transition matrix have similar scale of bias and CVRMSE. Sometimes one of the off-diagonal elements (a_{21} and a_{12}) has relatively large bias. We suspect it is because the true values for the off-diagonal are smaller than those for the diagonals so that we have fewer transitions in the unobserved Markov chain, therefore, there are fewer data points to estimate a_{21} and a_{12} than there are to estimate a_{11} and a_{22}.

3.6 Application to the Barnacle Settlement Data

In this section, we apply the EM algorithm, the Viterbi algorithm (VA) and posterior decoding (PD) to the oceanographic dataset described in Section 3.3. We assume at each time point there is an unobserved state from a Markov chain with two states—one for the population with occurrence of certain physical process and the other for the population without the occurrence of that process. Given that state, a larva settlement outcome is generated from a BEZI population. We know the actual settlement sequence, but because all the BEZI distributions can generate zero-inflated proportions, we do not know the population chain. To use HMM to answer the ecological association question,

we first estimate the parameters of the HMM using the observed outcome sequence. Then, based on these estimates, we decode the most likely hidden state sequence. After that, we compare the decoded sequence with the three known oceanographic processes to see how consistent they are. The more agreement between the decoded state chain and a process, the more likely the process is associated with larvae settlement.

We assume there are only equally spaced observation intervals, and we first use the EM algorithm to estimate the BEZI-HMM parameters. Then we use these estimates as the known parameters in the decoding methods to decode the hidden state sequence.

We conduct a grid search of initial values to see the possible impact of local maxima on our results. To be more specific, we consider $\pi_1 = (0.1, 0.2, 0.3, 0.4, 0.5)$, $a_{11} = a_{22} = (0.05, 0.25, 0.45, 0.65, 0.85)$, $p_1 = p_2 = (0.2, 0.4, 0.6, 0.8)$, $\mu_1 = \mu_2 = (0.2, 0.4, 0.6, 0.8)$ and $\phi_1 = \phi_2 = (1, 5, 10, 20, 40)$. This grid search gives 10000 starting points, which consequently produce 10000 stationary points. We compare these stationary points and treat two stationary points as the same when the difference between their maximized Q functions is less than 0.001[4]. We pick the largest three Q values and count the frequency of achieving these stationary points. We give the frequency results in Table 3.6 and the corresponding parameter estimates in Table 3.7. Mostly, the estimated values of the top three searches are very close. The second and third searches at the BDB site seem different from the first. However, this is only because the second and third searches have labeled state 1 and 2 in a different direction than the first search. There are some differences among the top three searches for ϕ_1 at the TPT site and ϕ_2 at the LHP site. However, when ϕ is greater than 40, its influence on the density function of BEZI is

[4]We choose this number because in the EM update we use 0.0001 as stopping tolerance.

very small. In other words, the density function of BEZI with $\phi = 118$ and $\phi = 127$ are very close to each other, as those with $\phi = 145$ and $\phi = 136$. Therefore, we only use the results from the top search as if they are the known parameter values in the decoding procedure.

Table 3.6: Grid search results in the EM initialization. The grid we consider are π_1=(0.1, 0.2, 0.3, 0.4, 0.5), a_{11}=(0.05, 0.25, 0.45, 0.65, 0.85), a_{22}=(0.05, 0.25, 0.45, 0.65, 0.85), p_1=p_2=(0.2, 0.4, 0.6, 0.8), μ_1=μ_2=(0.2, 0.4, 0.6, 0.8) and ϕ_1=ϕ_2=(1, 5, 10, 20, 40), which gives 10000 searching points. SHB, TPT, BDB and LHP are the four study sites.

Study site	Q rank	max(Q)	Frequency	Study site	Q rank	max(Q)	Frequency
SHB	1	12.589	1	TPT	1	7.929	1
	2	12.588	1		2	7.912	2
	3	12.587	3		3	7.910	1
BDB	1	9.491	1	LHP	1	23.527	1
	2	9.478	2		2	23.285	1
	3	9.477	2		3	23.158	1

At all study sites, the decoded state chains using the stationary points with the highest Q value are compared with the three ecological processes (regional relation - relax, diurnal upwelling - dup and front passage - front) through 2×2 tables. The two groups (with and without the occurrence of a particular process) can be labeled in two ways. For example, in Table 3.8 , if we label state 1 as decoded "with process event," then the diagonal cells, x_{11} and x_{22} are the numbers of matched assignments, whereas if we label state 2 as decoded "with process event," the off-diagonal cells, x_{12} and x_{21} are the numbers of matched assignments. We use the one gives a larger number of total matches ($x_{11} + x_{22}$ or $x_{12} + x_{21}$) as the final label.

The process with a higher match is more likely to be associated with the ecological settlement outcome. Meanwhile, to account for the proportion of agreement by chance,

Table 3.7: The EM estimates from the top three searches in the grid search, rounded to the third decimal. We consider π_1=(0.1, 0.2, 0.3, 0.4, 0.5), a_{11}=(0.05, 0.25, 0.45, 0.65, 0.85), a_{22}=(0.05, 0.25, 0.45, 0.65, 0.85), p_1=p_2=(0.2, 0.4, 0.6, 0.8), μ_1=μ_2=(0.2, 0.4, 0.6, 0.8) and ϕ_1=ϕ_2=(1, 5, 10, 20, 40), which gives 10000 searching points. SHB, TPT, BDB and LHP are the four study sites.

Study site	Q rank	iter	a_{11}	a_{22}	p_1	μ_1	ϕ_1	p_2	μ_2	ϕ_2
SHB	1	98	0.766	0.643	0.219	0.095	18.009	0.868	0.482	4.348
	2	98	0.766	0.643	0.219	0.095	18.009	0.868	0.482	4.348
	3	96	0.766	0.643	0.219	0.095	18.010	0.868	0.482	4.347
TPT	1	130	0.000	0.763	0.366	0.030	118.263	0.921	0.092	87.780
	2	123	0.000	0.768	0.363	0.030	126.859	0.920	0.091	86.162
	3	123	0.000	0.768	0.363	0.030	126.859	0.920	0.091	86.162
BDB	1	426	0.940	0.251	0.999	0.082	24.022	0.008	0.051	100.909
	2	362	0.252	0.940	0.008	0.051	103.125	0.999	0.081	28.003
	3	412	0.252	0.940	0.009	0.051	100.876	0.999	0.082	24.012
LHP	1	93	0.767	0.925	0.001	0.085	22.026	0.853	0.033	148.219
	2	49	0.770	0.926	0.0001	0.084	23.049	0.853	0.033	136.708
	3	78	0.770	0.927	0.001	0.085	22.097	0.853	0.033	135.707

we also compute Cohen's Kappa statistics (denoted as κ) (Wood, 2007). As stated in Wood (2007), "Cohen's Kappa is an index of inter-rater reliability that is commonly used to measure the level of agreement between two sets of dichotomous rating or scores." The value of κ depends on the observed and expected agreement, for the details of calculation, see Wood (2007). A κ equals to 1 means prefect agreement, -1 means prefect disagreement, 0 means random level of agreement/disagreement, i.e., no relationship between ratings. The interpretation of κ is similar as the interpretation of p-value. That is, there is no clear cut-off value of significance. According to Viera and Garrett (2005), $\kappa < 0$ implies less than chance agreement, $0.01 \leq \kappa \leq 0.20$ implies slight agreement and $0.21 \leq \kappa \leq 0.4$ implies fair agreement.

We show the agreement results in Table 3.9. Three of the four study sites have

Table 3.8: Illustration of 2×2 table for group matching assignment.

	With process	Without process
state 1	x_{11}	x_{12}
state 2	x_{21}	x_{22}

the same coded results for the Viterbi algorithm and posterior decoding. For site TPT, posterior decoding gives more agreement for all three pre-determined processes. For each location, we boldface the process with the highest number of matches (i.e., the highest match rate). There are several cases where the match rate is higher than 0.6[5]. However, according to κ, half of the time, there is only chance agreement, the other half, there is slight agreement. None of the cases gives an indication of fair agreement. In other words, most of the time, the agreement between the decoded chain and a pre-determined process is not different from random agreement. For SHB, the front passage process may be associated with the larval settlement because it has the highest match and its κ statistic indicates a slight agreement. For the same reason, at BDB, diurnal upwelling may be associated with the larval settlement. Whereas for TPT and LHP, it is likely that none of the pre-determined processes is associated with the ecological outcome. Because we are assuming equal-spaced interval (not actually true) for this study, better results could emerge after accounting for the unequal spacing.

[5]The match rates have to be greater than 0.5 because of the way we defining agreement in the two by two table: we choose the larger one in the diagonal sum and the off-diagonal sum, and the sum of the two is one.

Table 3.9: Decoding results for the Viterbi algorithm (VA) and Posterior Decoding (PD) for the oceanographical data (species *Balanus* cf.*glandula*). HMM parameters are obtained from the EM algorithm using a grid search initial values. Relax, dup and front are regional relaxation, diurnal upwelling and front passage processes; SHB, TPT, BDB and LHP stand for study area Sandhill Bluff, Terrace Point, Bonny Doon Beach and Lighthouse Point. Match rates are in % scale. Boldfaced numbers are the largest ones among different processes within the same site, n gives the sample size. κ represents Cohen's Kappa statistic, italic numbers are greater than 0.01.

		VA			PD		
Site	Process	# match	match rate	κ	# match	match rate	κ
SHB	relax	34	56.67	*0.100*	34	56.67	*0.100*
(n=60)	dup	33	55.00	*0.094*	33	55.00	*0.094*
	front	**35**	**58.33**	*0.150*	**35**	**58.33**	*0.150*
TPT	relax	35	0.583	-0.140	36	0.600	-0.113
(n=60)	dup	36	0.600	-0.216	37	0.617	-0.194
	front	**39**	**0.650**	-0.198	**40**	**0.667**	-0.179
BDB	relax	35	0.636	-0.032	35	0.636	-0.032
(n=55)	dup	**48**	**0.873**	*0.154*	**48**	**0.873**	*0.154*
	front	30	0.545	-0.057	30	0.545	-0.057
LHP	relax	**36**	**0.600**	*0.020*	**36**	**0.600**	*0.020*
(n=60)	dup	30	0.500	*0.130*	30	0.500	*0.130*

3.7 Discussion

We have developed a Hidden Markov Model with zero-inflated Beta emission distributions to model zero-inflated proportions with serial correlation. A HMM version EM algorithm is provided with justification of its applicability to the mixture of a discrete and continuous distribution emission density. To simplify notation and simulations, we have considered a model with only two states; however, all the methods we developed are valid for more than two states. We have evaluated the model and methods have through simulations and applied them to an oceanographic data example.

According to our simulations, for the decoding problem, the Viterbi algorithm is

better than posterior decoding in terms of making more accurate identification of hidden states. While, the Viterbi algorithm seems to be more sensitive to the BEZI density shapes, posterior decoding is more sensitive to the transition matrix.

As for the estimation problem, we find:

1. The precision parameter of the Beta component, ϕ, is harder to estimate than the zero proportion, p, and the mean parameter of the beta component, μ, as ϕ usually has much larger bias and RMSE than the other two.

2. The accuracy of estimates depends on the number of observations. When the HMM have more observations from one state, the estimates of the BEZI parameters at that state have smaller bias and RMSE than the ones from the other state. For the same reason when more transitions happen in one direction, the corresponding element in the transition matrix $\mathbf{A}$ also has less bias and smaller RMSE.

3. The performance of the EM algorithm depends on the initial values. When the starting values are close to the truth, the algorithm usually provides more accurate estimates and takes fewer iterations to converge. There is no apparent distinctions between the method of moments and method of maximum likelihood for use as initial values.

4. The BEZI density shape has strong impact on the EM algorithm performance.

For both decoding and estimation problems, asymmetric BEZI density provides the least challenge to the algorithms. Symmetric density provides the most challenge to the Viterbi algorithm, whereas bimodal density provides the most challenge to the EM algo-

rithm. One-sided density provides slightly less challenge than bimodal and symmetric density do to the algorithms.

When applying the EM algorithm to BEZI-HMM parameter estimation in data analysis, we suggest starting with multiple initial values. Researchers could use randomly generated start points or conduct a grid search as we did in the oceanographic data example. Then when multiple stationary points are achieved, the one (or ones) with the maximum value of objective function can be viewed as final estimation.

Compared with traditional HMM with mixture of Gaussian emission distribution, BEZI-HMM can model zero-inflated proportions with relative fewer number of parameters. Compared with other time series models and mixed effect models, rather than assuming a linear dependence, BEZI-HMM considers the autocorrelation among observations in a more integral sense without assuming any distribution or structure of the correlation. Instead, it lets the data inform about its specific structure. This can remove the subjectiveness in determining the correlation structure as in other time series models, for example, when using standard time series model, one often eye-balls autocorrelation function and partial autocorrelation function to identify the orders of autoregressive and/or moving average structure (Montgomery et al., 2008), which may or may not be correct. Also the latent Markov process is more interpretable in practice since it can match certain assumptions or theory in applied fields.

The BEZI-HMM might be extended to multiple inflated Beta populations with careful justification of the validity of the EM algorithm. One could incorporate more complicated likelihood structure. For example, we could replace the simple BEZI density by the likelihood in generalized linear model settings where BEZI parameters are linked

to certain linear combination of covariates. This is particularly appealing in ecological data, where temporal correlation and zero inflation are common due to the nature of the field (Chiogna and Gaetan, 2007; Barry and Welsh, 2002) and extra information are available from the study process. Current statistical methods for ecological analysis often involve count data (Marin et al., 2005). Zero-inflated continuous data are mostly based on the log-normal (Tian, 2005), Gamma (Feuerverger, 1979) and exponential (Wu et al., 2012; Zhang et al., 2010) model. Often autocorrelation is ignored or considered in hard-shelled ways, for example, assumed as known and fixed, in which subjectiveness are common. Our study fills the gap in modeling correlated zero-inflated continuous data with a constrained support.

The EM algorithm is often criticized for its lack of an automatically produced standard error (McLachlan and Krishnan, 2008). There are two ways of calculating standard errors for MLE obtained from the EM algorithm (Baker, 1992): using the information matrix based on the likelihood model or using resampling techniques. For the information based methods, examples for the EM algorithm in general can be found in Baker (1992) and Jamshidian and Jennrich (2000). For HMM, Aittokallio and Uusipaikka (2000) discuss the situation for normal emission density HMM; Lystig and Hughes (2002) provide an idea for computing the observed information matrix for general HMM. The idea may be applicable to BEZI-HMM. However, the logic of using information matrix to obtain standard error is to use the asymptotic properties of MLE. Therefore, large sample size is a requirement for accurate standard error. As for the resampling technique, bootstrap approach can be implemented, however it is usually computationally expensive (Baker, 1992; Zucchini and MacDonald, 2009). Also for the HMM problem

because of the autocorrelation among observations, simply resampling individual observation may not be sufficient, we may have to use some technique like block bootstrap (Härdle et al., 2003).

One benefits of using the EM algorithm is that one may expand this framework to handle the unequally spaced interval problem in the oceanographic data example. The unequally spaced intervals could be a result of another form of missing data: the irregularly spaced observations come from a chain with regularly spaced observation with some observations missing. A hierarchical framework could be established by considering another structure that controls whether at any given time t the observation is actually observed. Therefore, when there are observations missing at some t, the sequence becomes unequally spaced. This is a topic for future work.

3.8 Appendix

3.8.1 A Justification for the Viterbi Algorithm

The objective of the Viterbi algorithm is to find the state sequence that maximizes $L(\mathbf{S}|\mathbf{y};\theta)$. Given $\mathbf{y}$ and θ,

$$\max L(\mathbf{S}|\mathbf{y};\theta) \Leftrightarrow \max L(\mathbf{S},\mathbf{y};\theta) \Leftrightarrow \max l(\mathbf{S},\mathbf{y};\theta) \text{ where } l(\cdot) = log(L(\cdot)).$$

Because $L(\mathbf{S},\mathbf{y};\theta) = \pi_{S_1} f_{S_1}(y_1) a_{S_1S_2} f_{S_2}(y_2) \cdots a_{S_{T-1}S_T} f_{S_T}(y_T)$, we have

$$l(\mathbf{S},\mathbf{y};\theta) = log(\pi_{S_1}) + log(f_{S_1}(y_1)) + log(a_{S_1S_2}) + log(f_{S_2}(y_2)) + \cdots$$

$$+log(a_{S_{T-1}S_T}) + log(f_{S_T}(y_T)).$$

Define $h_1(S_1) = log(\pi_{S_1}) + log(f_{S_1}(y_1))$ and $h_t(S_t, S_{t-1}) = log(a_{S_{t-1}S_t}) + log(f_{S_t}(y_t))$, then

$$l(\mathbf{S}, \mathbf{y}; \theta) = h_1(S_1) + \sum_{t=2}^{T} h_t(S_t, S_{t-1}).$$

We can always split $l(\mathbf{S}, \mathbf{y}; \theta)$ at time $t \geq 2$ into three terms:

$$l(\mathbf{S}, \mathbf{y}; \theta) = h_1(S_1) + \sum_{m=2}^{t} h_m(S_m, S_{m-1}) + \sum_{m=t+1}^{T} h_m(S_m, S_{m-1}),$$

where the first two terms only involve observations before t and the last term only involves observations after t.

For $k = 1, 2$, define

$$H_{t,k}(S_1, \cdots, S_{t-1}, S_t = k) = h_1(S_1) + \sum_{m=2}^{t-1} h_m(S_m, S_{m-1}) + h_t(k, S_{t-1}) \text{ for all } t.$$

To simplify notation, we abbreviate $H_{t,k}(S_1, \cdots, S_{t-1}, S_t = k)$ as $H_{t,k}$. According to this definition, the Viterbi algorithm aims to find $S_1, \cdots, S_{T-1}$ and k that maximize $H_{T,k}$. To do so, we can first find maximized $H_{T,k}$ for each k in the state space (in our two-state case, $k = 1, 2$), i.e., find $S_1, \cdots, S_{T-1}$ that maximize $H_{t,k}$ and then choose the k that has the largest $H_{T,k}$.

We show (in **Lemma 1**) that for $l = 1, 2$

$$H_{t,l}(S_1, \cdots, S_{t-2}, S_{t-1} = k, S_t = l) = H_{t-1,k}(S_1, \cdots, S_{t-2}, S_{t-1} = k) + h_t(l, k).$$

Because of this recursive relation, we can maximize $H_{t,k}$ for each t step by step and eventually find the maximized $H_{T,k}$ for each k.

Specifically, we use a "$*$" to stand for maximum. For $t = 1$, we have $H^*_{1,k} = h_1(k)$. For $t = 2$, the maximized $H_{2,k}$ is

$$H^*_{2,k} = \max_{l=1,2}\{H^*_{1,l} + h_2(k,l)\} = H^*_{1,l^*} + h_2(k,l^*) \text{ where } l^* = \arg\max_{l=1,2}\{H^*_{1,l}\}.$$

For $t = 3, \cdots, T$ we repeat the procedure successively as

$$H^*_{t,k} = \max_{l=1,2}\{H^*_{t-1,l} + h_t(k,l)\} = H^*_{t-1,l^*} + h_t(k,l^*),$$

where l^* gives the proceeding state that maximize $H^*_{t-1,l}$. Therefore $H^*_{t,k}$ is the sum of the log likelihoods of the best path up to time t that ends in state k. We keep a record of $\{l^*\}$ in $\psi_t(k)$ for $t = 1, \cdots, T$. By time T, two paths are formed and each path ends with different state. Now we select the path ends with k^*, where $k^* = \arg\max_{k=1,2}\{H^*_{T,k}\}$. This path is the Viterbi path.

Lemma 1 (Recursive relation)**.** *Show*

$$H_{t,l}(S_1, \cdots, S_{t-2}, S_{t-1} = k, S_t = l) = H_{t-1,k}(S_1, \cdots, S_{t-2}, S_{t-1} = k) + h_t(l,k).$$

Proof. Because

$$H_{t,l}(S_1, \cdots, S_{t-2}, S_{t-1} = k, S_t = l) = h_1(S_1) + h_2(S_2, S_1) + \cdots + h_{t-2}(S_{t-2}, S_{t-3})$$

$$+h_{t-1}(S_{t-1}=k,S_{t-2})+h_t(S_t=l,S_{t-1}=k)$$

and

$$H_{t-1,k}(S_1,\cdots,S_{t-2},S_{t-1}=k)=h_1(S_1)+h_2(S_2,S_1)+\cdots$$

$$+h_{t-2}(S_{t-2},S_{t-3})+h_{t-1}(S_{t-1}=k,S_{t-2}).$$

That is

$$H_{t,l}(S_1,\cdots,S_{t-2},S_{t-1}=k,S_t=l)=H_{t-1,k}(S_1,\cdots,S_{t-2},S_{t-1}=k)+h_t(S_t=l,S_{t-1}=k).$$

This gives the recursion relationship of $H_{t,k}$ □

3.8.2 A Sketch of Why $l(\theta|y_{obs})$ is Bounded

$L(\theta,\mathbf{y}_{obs})$ is our marginal log-likelihood function. According to (3.4), because $\mu(\mathbf{S})$ is the discrete measurement of the states $\mathbf{S}$, the marginal likelihood function is

$$L(\theta,\mathbf{y})= \quad L(\theta;\mathbf{y},\mathbf{S})d\mu(\mathbf{S})$$

$$=\sum_{\mathbf{S}}\prod_{t=1}^{T}\prod_{i=1}^{2}b_i(y_t)^{I(S_t=i)}\prod_{t=1}^{T-1}\prod_{i,j=1}^{2}a_{ij}^{I(S_{t+1}=j,S_t=i)}\prod_{i=1}^{2}\pi_i^{I(S_1=i)}.$$

Notice that for a HMM with T observations, there are 2^T possible state chains because S_t's consist of only 1's and/or 2's. For a particular state chain, say $S=(1,1,\cdots,1)$, the product part is just $\prod_{t=1}^{T}b_1(y_t)\prod_{t=1}^{T-1}a_{11}\pi_1$, which is a finite number given fixed values of $\mathbf{y}$'s. As a result, $L(\theta,\mathbf{y})$ is a finite sum of finite numbers, which has to be finite. Hence

the marginal log-likelihood is bounded from above.

4 Bayesian Analysis for Zero-inflated Beta Mixed Model with Autoregressive Random Effect

4.1 Abstract

In this paper, we propose a generalized linear mixed model with autoregressive random effect to model zero-inflated proportion data with serial correlation. The response distribution is zero-inflated Beta, conditional on realization of the random effect. Bayesian methodology is adopted for parameter estimation and statistical inference. We provide guidelines for monitoring convergence of our Markov Chain Monte Carlo used to simulate from the posterior distribution. We use simulations to evaluate the method. We find that fixed effect coefficients and the variance parameter in the autoregressive random effect are more likely to have convergence problems than other parameters. The autoregressive parameter, φ, is often not different from 0, despite the fact that its posterior interval almost always includes its true value. When the mean of the Beta component is small, the posterior interval of the zero proportion parameter often excludes its true value. Some intervals of fixed effect coefficients also exclude their true values. When the value of precision parameter of the Beta component, ϕ, increases, the performance of the method improves. We also apply the model to an oceanographic dataset.

Keywords: Zero-inflated Beta, Generalized Linear Mixed Model, Autoregressive Random Effect.

4.2 Introduction

Data that are proportions falling in the continuum (0, 1) are very common in practice. Examples include the percentage of conifer cover in a particular area, the proportion of household income spent on food and the volume of stroke lesion as a percentage of total brain volume. The family of Beta distributions provides broad flexibility for modeling proportions of some continuous measurement such as area, income and volume. It can happen, however, that an inflated number of zeros and/or ones in a sample of proportions can render the Beta distribution an unsuitable model since its support does not include 0 or 1. Ospina and Ferrari (2010) propose a mixed continuous-discrete distribution for data observed on [0, 1), (0, 1] or [0, 1]. The discrete component is defined by a degenerate (point mass) distribution that assigns non-zero probability to 0 and/or 1 depending on whether there is zero- and/or one-inflation. In particular, the zero-inflated Beta (BEZI), the primary focus for our study here, has a point mass at zero.

Suppose Y is a random variable following $BEZI(p,\mu,\phi)$. Then the probability density function of Y is

$$f_Y(y) = \begin{cases} p & \text{if } y = 0, \\ (1-p)\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} & \text{if } y \in (0,1), \end{cases}$$

where $\Gamma(\cdot)$ is the Gamma function; $0 < p < 1$, $0 < \mu < 1$ and $0 < \phi < \infty$. The mean and variance of Y are, respectively,

$$E(Y) = (1-p)\mu$$

and

$$Var(Y) = (1-p)\frac{\mu(1-\mu)}{\phi+1} + p(1-p)\mu^2, \tag{4.1}$$

where μ and ϕ are the mean and precision parameters of the Beta component. As is the Beta family of distributions, the BEZI family is quite flexible in shape.

Being a member of exponential family (Ospina and Ferrari, 2010), the BEZI family is a candidate for the response distribution in generalized linear models (Nelder and Wedderburn, 1972). For example, Ospina and Ferrari (2012) propose a regression model for a general class of zero-or-one inflated Beta distributions including BEZI. In this model, the authors link explanatory variables to all three parameters in the BEZI density through suitable link functions and they assume the response variables are independent after accounting for explanatory variables. However, the independence assumption may not always be valid. For example, when observations are gathered within a certain time period or spacial extent, the observations that are collected close together are likely to be related. In the presence of dependence, a generalized linear model is not appropriate.

As an extension of generalized linear model (McCulloch et al., 2008), the generalized linear mixed model (GLMM) can accommodate not only non-normal responses but also autocorrelation among those responses. GLMM can be viewed as an extension of linear mixed model as well. On the one hand, compared with the generalized linear model, the GLMM adds random effects to the linear predictor, and usually, random effects are assumed to be normally distributed. On the other hand, compared with the linear mixed model, conditioned on realizations of random effects, the GLMM has non-normally distributed response variable. For a brief history of GLMM, see Littell et al.

(2006), chapter 14.

As mentioned already, there are two keys in a GLMM: The first key is the response distribution. This usually takes the form of an exponential family and the conditional mean of a response is linked to the linear predictor through a link function. The second key is the random effects in the linear predictor. Random effects may be used to model the correlation of observations in clusters. Examples are plots within blocks in an experimental study or repeated measurements on subjects in a longitudinal study. Random effects can also be used to account for serial correlation when a sequence of observations are obtained through time. In this paper, we focus on this second type of correlation.

Given a GLMM, there are three general methods for parameter estimation (Montgomery et al., 2008):

(a) Linearize the conditional mean and then repeatedly apply linear mixed model techniques to the approximated model. Examples are pseudo-likelihood method (Wolfinger and O'Connell, 1993) and quasi-likelihood method (Breslow and Clayton, 1993)

(b) Use numerical methods to approximate the integrals involved in the marginal likelihood and develop a set of estimating equations based on this approximation. Examples are Laplace (Pinheiro and Bates, 1995) and quadrature approximation (Anderson and Aitkin, 1985; Pinheiro and Chao, 2006)

(c) Bayesian methods (Zeger and Karim, 1991; Hay and Pettitt, 2001; Robinson et al., 2009; Fong et al., 2010)

A major concern for (a) is the potential bias in estimating model parameters when there is presence of relatively large variance components (Pinheiro and Chao, 2006;

Fong et al., 2010). Lower order approximation in (b) may produce biased estimates as well (Pinheiro and Chao, 2006). The higher order approximation can be more accurate; however it increases the computational complexity (Pinheiro and Chao, 2006). Usually, likelihood-based approaches need large sample sizes to achieve the property of asymptotic sampling distribution of estimators (Fong et al., 2010). Meanwhile, because of nonlinearity (Zeger and Karim, 1991) and the uncertainty in estimating the variance components (Natarajan and Kass, 2000), variance expressions of the fixed effect estimates, typically denoted as $\widehat{Var}(\hat{\beta})$, are hard to derive. In non-standard distribution situation like BEZI, there are no existing tools or theoretical results in any of these likelihood-based methods.

Bayesian inference relies on the posterior distribution of parameters so it does not require approximate normality or asymptotic properties (Gelman et al., 2004). The posterior distribution provides Bayesian point estimates (posterior means or medians), posterior standard deviations, as well as interval estimates (posterior intervals). In addition, the interpretation of the interval estimates in Bayesian method is quite straightforward. Inferences about functions of parameters are also easy to obtain.

In this paper, we develop a generalized linear mixed model using a zero-inflated Beta distribution as the distribution of the response variables. Our model has an autoregressive random effect. To be more specific, we let $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_n)$ denote the vector of response variables. Each Y_t is a BEZI random variable with parameters (p, μ_t, ϕ) for $t = 1, 2, \cdots, n$. The mean of the Beta component, μ_t, is linked to fixed effects and an autoregressive random effect through a logit link function. More details of this model are given in Section 4.4.1.

There are some examples of generalized linear mixed model about zero-inflated data or having an autoregressive random effect in the literature. Lawrence (1982) proposes an autoregressive model with order 1 (the so-called AR(1) model) having a Gamma marginal distribution. Zeger (1988) proposes a quasi-likelihood based on time series model for counts by adding an unobservable process to the linear predictor in a log-linear model. McGilchrist (1995) gives a generalized linear mixed model with Binomial response and an AR(1) distributed random component that is applicable when there are multiple time series. Hall (2000) discusses zero-inflated Poisson and zero-inflated Binomial model with random intercept in the case of repeated measurements. Hay and Pettitt (2001) consider a time series model for counts in Bayesian framework. Robinson et al. (2009) propose a Bayesian analysis for split-plot experiments with Gamma response. Rocha and Cribari-Neto (2009) develop Beta autoregressive moving average models based on a regression model in Ferrari and Cribari-Neto (2004). Jazi et al. (2012) propose a first order integer valued autoregressive process with zero-inflated Poisson innovation. To the best of our knowledge, there are no studies concerning both zero-inflated proportions and serial correlation simultaneously. Our work fills the gap.

The organization of this paper is as follows. In Section 2, we describe an example from marine science where interest lies in distinguishing two populations. The observations from these populations are well-represented by BEZI distributions. Because of the sampling method, there is autocorrelation among the observations. In Section 3, we describe the generalized linear mixed model with BEZI distributed response in more detail and describe our Bayesian implementation. We discuss the choice of priors and how to monitor the MCMC convergence. In Section 4, we describe a simulation study

used to evaluate the performance of our model. In Section 5, we apply our model to the marine science data. We discuss our findings and possible extensions of this work in Section 6.

4.3 Settlement of Onshore Barnacle Larvae

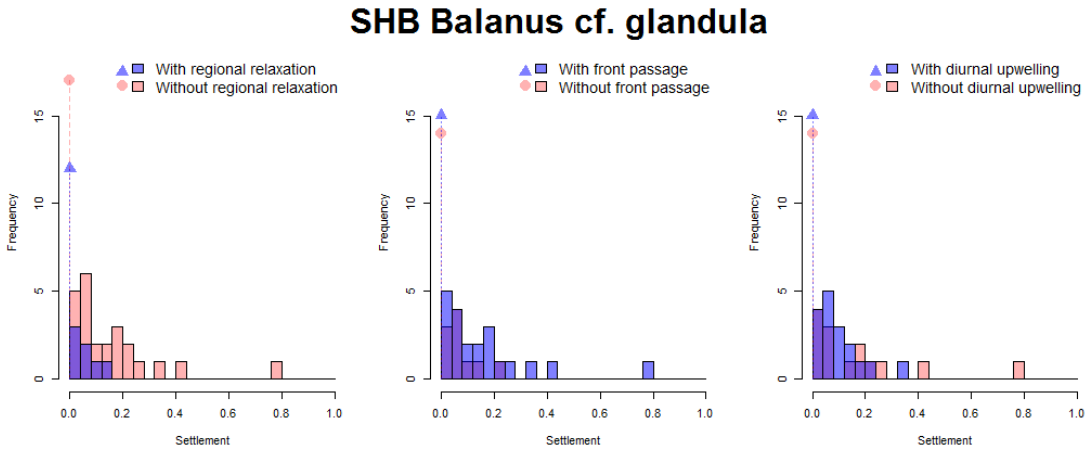


Figure 4.1: Histogram of settlement for *Balanus* cf.*glandula* at Sandhill Bluff (SHB) with and without physical processes (regional relaxation, front passage and diurnal upwelling; left to right).

Tyburczy (2011) compared settlement distributions of onshore barnacle larvae with and without the occurrence of different oceanographic processes such as large-scale regional relaxation of upwelling and smaller scale processes such as localized diurnal upwelling driven by afternoon sea breezes. The data consist of observations on two types of barnacle larva (*Balanus* cf.*glandula* and *Chthamalus* spp.): daily or bi-daily settlement information and covariate information such as the occurrence of multiple physical

processes (regional relaxation - relax, front passage - front, and diurnal upwelling - dup) in four study areas (Sandhill Bluff - SHB, Terrace Point - TPT, Bonny Doon Beach - BDB, and Lighthouse Point - LHP) within northern Monterey Bay, CA in 2007 (May-Sept). According to Tyburczy (2011), the larva settlement data are a combination of onshore intertidal plate samples that were normalized for hours of immersion based on tidal height of the plates, larva counts from larval traps deployed on and below the surface of multiple moorings with different depths and samples from water collected using pumps on several vessels. The presence/absence of physical processes were also determined based on combinations of different physical conditions, such as temperature, salinity, current direction and velocity, local wind forcing and nearshore pressure gradients. For *Balanus* cf.*glandula* (Bal in what follows), the distributions of larva settlement are compressed to the range between zero and one after the data manipulation; i.e., all observations are in the interval [0, 1), and more than half of them are zeros. Some examples of Bal larvea settlement data are shown in Figure 4.1. All panels are for the SHB site with and without different oceanographic processes. Notice that in each panel, the two zero counts are also indicated on the far left hand side.

Tyburczy's ultimate goal was to build an ecological model to describe the mechanism of nearshore barnacle transport and settlement. To do this, a fundamental question was whether there is an association between the larvae settlement outcomes and the oceanographic processes. Evaluation of this association is complicated by non-normality and autocorrelation in the data, as well as uneven sampling intervals. In this paper, we use a generalized linear mixed model with distribution from BEZI family as the response distribution to account for the non-normal distribution feature and the au-

tocorrelation simultaneously. In this work, we assume equal sampling intervals, which is approximately true for these data.

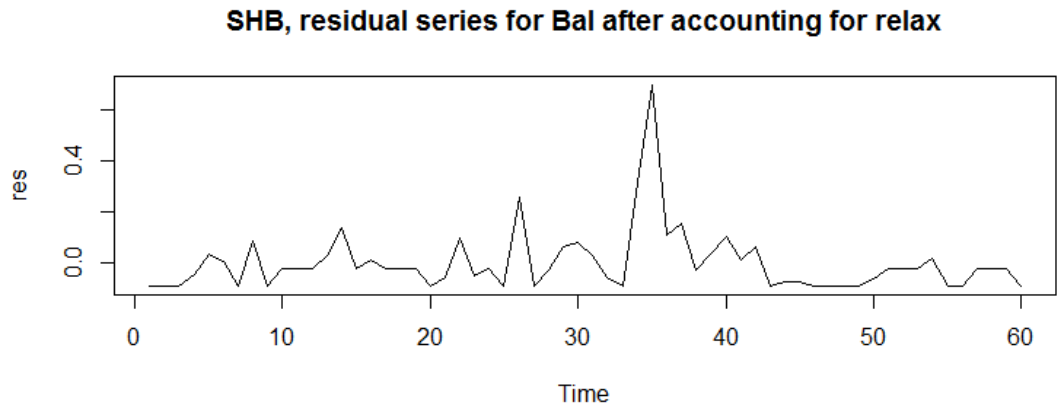


Figure 4.2: Residual series from a simple linear regression model: *Balanus* cf.*glandula* (bal) is regressed on regional relaxation (relax) process (p-value=0.054) for the SHB site.

The first benefit of using GLMM in setting like this is that it allows us to consider multiple processes simultaneously (with their interaction terms as well when necessary). More importantly, GLMM can account for the autocorrelation among observations. Figure 4.2 and 4.3 give time series plot and its corresponding autocorrelation function (ACF) and partial autocorrelation function (PACF) plots for residuals sequence from a simple linear regression model that has Bal settlement information as response and relax as explanatory variable. This residual series can reasonably be identified as an autoregressive process with order 1 because its PACF plot has a peak at lag 1 and its ACF plot decreases exponentially (Montgomery et al., 2008).

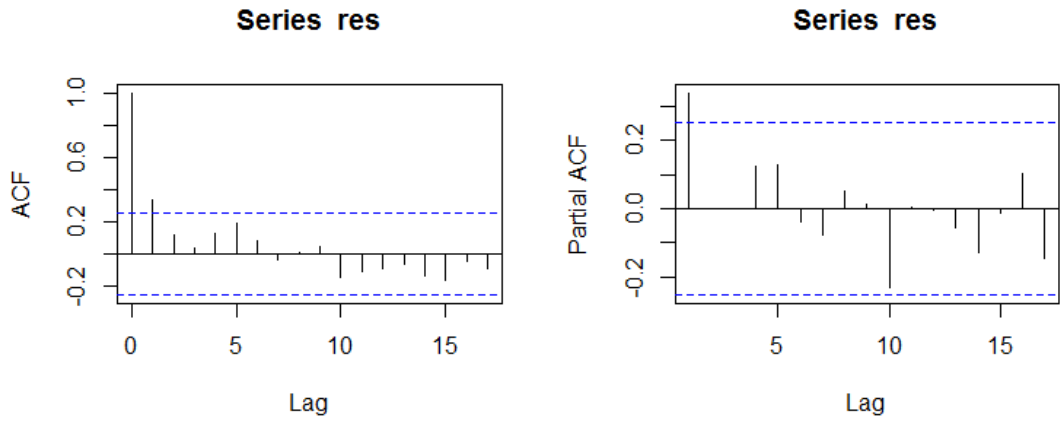


Figure 4.3: ACF and PACF plots of the residual series from a simple linear regression model: *Balanus* cf.*glandula* (bal) is regressed on regional relaxation (relax) process (p-value=0.054) for the SHB site. The series can be identified as an autoregressive process with order 1.

4.4 Methodology

In this section, we introduce the zero-inflated Beta generalized linear mixed model with autoregressive random effect, and we provide details of our Bayesian approach for parameter estimation.

4.4.1 Zero-inflated Beta Generalized Linear Mixed Model with AR(1) Random Effect

Let $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_n)$ denote the vector of response variables. Let

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1m} \\ 1 & X_{21} & X_{22} & \cdots & X_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nm} \end{bmatrix}$$

be the design matrix in which $\mathbf{X}_i = (X_{1i}, X_{2i}, \cdots, X_{ni})$ is the i^{th} explanatory variable for i=1,2, $\cdots, m$. In the oceangraphic dataset, because there are three processes, $m = 3$ and the $\mathbf{X}_i$'s are variables containing information about individual physical process. More specifically, the $\mathbf{X}_i$'s are binary vectors with 1 and 0 denoting presence and absence of each particular process. For $t = 1, 2, \cdots, n$, we denote $\mathbf{X}_t = (1, X_{t1}, X_{t2}, \cdots, X_{tm})^T$ and $\beta = (\beta_0, \beta_1, \cdots, \beta_m)$. Then

$$\rho_t = g(\mu_t) = \mathbf{X}_t\beta + \eta_t \tag{4.2}$$

is the linear predictor, in which $g(.)$ is the link function, $\mathbf{X}_t$ is the $1 \times (m+1)$ design matrix in the fixed effect for the t^{th} observation, β is the $(m+1) \times 1$ unknown fixed effect parameter vector and η_t is a random time effect. Because the mean parameter for the Beta component, μ_t, has support space $(0, 1)$, we use logit link function. That is

$g(\mu_t) = log\left(\frac{\mu_t}{1-\mu_t}\right)$. Let

$$h(\rho_t) = g^{-1}(\rho_t) = \frac{exp(\rho_t)}{1+exp(\rho_t)}, \tag{4.3}$$

$$\text{then } \mu_t = h(\rho_t) = h(\mathbf{X}_t\beta + \eta_t) = \frac{exp(\mathbf{X}_t\beta + \eta_t)}{1+exp(\mathbf{X}_t\beta + \eta_t)}.$$

For η_t, we use a stationary, invertible Gaussian autoregressive time series process with order 1 (the so-called AR(1) model (Montgomery et al., 2008)). We have

$$\eta_1 = \frac{\varepsilon_1}{\sqrt{1-\varphi^2}} \text{ and}$$

$$\eta_t = \varphi\eta_{t-1} + \varepsilon_t \text{ for } t \geq 2,$$

where $|\varphi| < 1$, $\{\varepsilon_t\}_t$ are Gaussian white noise (i.e., the ε_t's are independent random variables following $N(0,\sigma^2)$)[1]. The AR(1) process can also be expressed in a form of conditional distribution:

$$\eta_1 \sim N \quad 0, \frac{\sigma^2}{1-\varphi^2} \quad \text{and} \tag{4.4}$$

$$\eta_t|\eta_{t-1}, \cdots \eta_1 \sim N(\varphi\eta_{t-1}, \sigma^2) \text{ for } t = 2, \cdots n. \tag{4.5}$$

We will use (4.4) and (4.5) when we specify priors for $\eta_1, \eta_2, \cdots, \eta_n$ in Section 4.4.2.1.

We assume that given the values of the μ_t's, the Y_t's are independent random variables following $BEZI(p, \mu_t, \phi)$ for $t = 1, \cdots, n$. Let $\mathbf{y} = (y_1, y_2, \cdots, y_n)$ be realizations

[1]Notice that in general we could assume non-Gaussian white noise $\{\varepsilon_t\}$ so that $\eta = (\eta_1, \cdots, \eta_n)$ have means $\mathbf{0}$ and variance covariance matrix Σ, where Σ is a function of σ^2 and φ, or we could assume a Gaussian autoregressive moving average model (ARMA(p,q)) with η follows $N(\mathbf{0}, \Sigma)$, in which Σ is a function of p, q and other variance covariance parameters.

of $\mathbf{Y}$.

We show that the unconditional mean and variance of Y_t are

$$E(Y_t) = (1-p)\frac{exp(\mathbf{X}_t\beta)}{1+exp(\mathbf{X}_t\beta)}, \tag{4.6}$$

$$\begin{aligned} Var(Y_t) = {} & \frac{1-p}{\phi+1}\frac{exp(\mathbf{X}_t\beta)}{1+exp(\mathbf{X}_t\beta)} - (1-p)^2 \left[\frac{exp(\mathbf{X}_t\beta)}{1+exp(\mathbf{X}_t\beta)}\right]^2 \\ & + \frac{(1-p)\phi}{\phi+1}\frac{(exp(\mathbf{X}_t\beta))^2}{(1+exp(\mathbf{X}_t\beta))^2}\left[1+\frac{\sigma^2}{(1-\varphi^2)(1+exp(\mathbf{X}_t\beta))^2}\right]. \end{aligned} \tag{4.7}$$

Details are in Appendix.

According to (4.6), ϕ, σ^2 and φ do not impact the unconditional mean of the Y_t's. In (4.7), ϕ, the β's and p are the primary contributors to the unconditional variance. Consequently, p, the β's and ϕ are more influential than σ^2 and φ to the values of y_t's.

4.4.2 Estimation Method

Let $\theta = (\beta, p, \phi, \sigma^2, \varphi)$ be the unknown parameters. $\eta = (\eta_1, \cdots, \eta_n)$ are the unobserved random effects that can be viewed as parameters as well.

Given a sequence of μ_t's, the Y_t's are independent. Therefore the likelihood function of $\mathbf{Y}$ is a product of individual likelihoods. In other words,

$$L(\mathbf{y}|\theta,\eta) = \prod_{t=1}^{n} f_{Y_t}(y_t), \tag{4.8}$$

where $f_{Y_t}(y_t)$ is the probability density function of BEZI random variable given in Sec-

tion 4.2.

In Bayesian approaches, all inferences depend on the posterior distribution of model parameters. In our case, the posterior distribution of (θ, η) is

$$f(\theta, \eta|\mathbf{y}) = \frac{L(\mathbf{y}|\theta, \eta)f(\eta, \theta)}{f(\mathbf{y})} = \frac{L(\mathbf{y}|\theta, \eta)f(\eta, \theta)}{\int\int L(\mathbf{y}|\theta, \eta)f(\eta, \theta)d\theta d\eta}. \tag{4.9}$$

$L(\mathbf{y}|\theta, \eta)$ is the likelihood function in (4.8), $f(\eta, \theta)$ is the prior distribution of the parameters. We write $f(\eta, \theta)$ as

$$f(\eta, \theta) = f(\eta|\theta)f(\theta). \tag{4.10}$$

When the $\mathbf{y}$ is known, the denominator in (4.9) is a scaler (normalizing constant) (Dongen, 2006). Hence, combining (4.9) and (4.10), we have

$$f(\theta, \eta|\mathbf{y}) \propto L(\mathbf{y}|\theta, \eta)f(\eta, \theta) = L(\mathbf{y}|\theta, \eta)f(\eta|\theta)f(\theta).$$

That is, the posterior distribution is a combination of prior knowledge about the parameters ($f(\eta, \theta)$ or $f(\eta|\theta)f(\theta)$) and knowledge obtained in the data ($L(\mathbf{y}|\theta, \eta)$). The primary parameters of interest to us are the β's because they are the coefficients of the explanatory variables. The other parameters, (p, φ, ϕ, σ^2, η), can be viewed as nuisance parameters.

While the Bayesian approach has many advantages, it poses two important additional difficulties, namely, prior specification and issues related to convergence of the MCMC simulation (Dongen, 2006). In the next two sections, we discuss these explic-

itly.

4.4.2.1 Prior specification

We give thee joint prior distribution in (4.10). We now discuss the prior distributions for individual parameters.

Because of the AR(1) structure,

$$f(\eta|\theta) = \prod_{t=2}^{n} f(\eta_t|\eta_{t-1}, \cdots, \eta_1, \theta) f(\eta_1|\theta).$$

When we assume normally distributed white noise, according to (4.4) and (4.5), $f(\eta_t|\eta_{t-1}, \cdots, \eta_1, \theta)$ and $f(\eta_1|\theta)$ are normal densities whose explicit forms only depend on the values of φ and σ^2 and a sequence of realization of normal random variables. Therefore, the primary challenge here is to find reasonable prior distributions for θ.

We further assume information about each individual parameter in θ is independent of information about the others. In other words, we assume independent priors (Christensen et al., 2011):

$$f(\theta) = \prod_{i=0}^{m} f(\beta_i) f(p) f(\phi) f(\sigma^2) f(\varphi).$$

We use non-informative priors. The idea is to use a prior that is uniform over the range of interest so that all the possible values along that range are equally likely to be considered (Lambert et al., 2005). Ideally, given non-informative priors, Bayesian inference would be very close to likelihood inference when the sample size is large (the

Table 4.1: All priors investiaged for BEZI GLMM Bayesian method. Parameters in normal densities are means and variances.

Parameter	Prior	Parameter	Prior
β_0 and β_1	$N(0,1000000)$	ϕ	$Uniform(0,100)$
	$N(0,10000)$		$Uniform(0,40)$
			$Uniform(0,20)$
p	$Uniform(0,1)$		$Gamma(0.1,0.1)$
			$Gamma(0.01,0.01)$
σ	$Uniform(0,10)$		$\gamma=\frac{1}{\phi}\sim Gamma(0.001,0.001)$
	$Uniform(0,100)^{**}$		$\gamma=\frac{1}{\phi}\sim Gamma(0.01,0.01)$
	$\tau=\frac{1}{\sigma^2}\sim Gamma(0.001,0.001)^{*}$		$\gamma=\frac{1}{\sqrt{\phi}}\sim Gamma(0.01,0.01)$
			$\gamma=\sqrt{\phi}\sim Uniform(0,10)$
φ	$Uniform(-1,1)$		

$*$ Gamma distribution on the precision, a proper distribution (Spiegelhalter et al., 2003)
$**$ Uniform distribution on standard deviation, an improper prior (Gelman, 2006)

so-called "data to dominate" (Lambert et al., 2005)).

We investigate nine sets of priors on a single simulated dataset with one explanatory variable (so β_0 and β_1 are the only parameters for the fixed effects in the linear predictor). We enumerated these priors in Table 4.1. Each of these sets has an unique prior distribution for the precision parameter, ϕ. There are some discussions about priors for precision parameters. However, almost always, these are for the precision parameter in the normal density. For examples, Lambert et al. (2005) discuss several priors for variance components under hierarchical normal model (one-way random effect model). Kass and Wasserman (1996) and Gelman (2006) also have reviews about prior selection. ϕ is similar to the precision parameter in normal density, $\frac{1}{\sigma^2}$. Unlike $\frac{1}{\sigma^2}$ in the normal case, however, ϕ cannot uniquely specify the variance of a BEZI random variable because as shown in (4.1) p and μ also contribute to its variance. To the best of our knowledge, there is no discussion about prior specification for ϕ in literature. Because

Table 4.2: Final prior used for simulation and real data application.

Parameter	Prior
β_i for $i = 0, 1, \cdots, m$	$N(0, 10000)$
p	$Uniform(0, 1)$
ϕ	$Gamma(0.1, 0.1)$
σ	$Uniform(0, 10)$
φ	$Uniform(-1, 1)$

of this, we tried nine prior distributions for it, which gives our nine sets of priors.

For each prior set we used three chains started at different initial values from across the parameter space. We ran each chain for 70000 iterations. After discarding the first 1000 iterations as burn-in, we sampled every 150 iterations, leaving 1320 samples for inference. We assessed the convergence of MCMC by inspection of Gelman-Rubin statistics (Gelman and Rubin, 1992) and the trace plots of the samples.

Among all the parameters, the precision parameter of the Beta component, ϕ, is the most problematic. Quite often, when the marginal posterior distribution for the other parameters have small Monte Carlo errors (Spiegelhalter et al., 2003), ϕ has a quite large one. Sometimes β_0 and β_1 also have large Monte Carlo errors and wide posterior intervals. We suspect the uncertainty in ϕ gives the BEZI unstable variation, which consequently, influence the results for the β's.

The final priors we selected for both simulated and real data are given in Table 4.2. We choose uniform distributions as priors for p and φ because they have bounded parameter spaces and uniform distributions can provide flat density functions over those spaces. For the β's, as Gelman (2006) points out, when the predictor is binary and the regression is on the logit scale, for most applications the effect size will be less

than 10 and certainly less than 100, so we would expect small variation on these coefficients. In the selection of the prior for σ^2, we followed the suggestions in Gelman (2006) and Lambert et al. (2005), as our simulation provides consistent results to theirs: Inverse Gamma prior on σ^2 is quite sensitive to the choice of scale parameters (Gelman, 2006) and a uniform prior for σ has miscalibaration toward higher value. We use $Gamma(0.1, 0.1)$ as the prior for ϕ because the trace plots of samples under other priors have a few spikes for ϕ or for other parameters. All priors seem to provide convergent chains, according to the Gelman-Rubin statistic. We think it is less risky to use $Gamma(0.1, 0.1)$ as prior for ϕ because as we will discuss later, Gelman-Rubin statistics are not sufficient for monitoring convergence. We tested these final priors in Table 4.2 under different parameter settings, each with five simulated datasets. There is no indication of convergence problem.

4.4.2.2 MCMC Convergence

We use the Gelman-Rubin diagnostic statistic (Gelman and Rubin, 1992) to monitor chain convergence. We also use visual inspection of trace plots of all parameters (except the random effects). As described by Lambert et al. (2005), a Gelman-Rubin statistic is an estimated posterior variance of a particular parameter based on a mixture of several MCMC runs. The method requires multiple chains starting from different initial points. The statistic tries to distinguish the natural variability in a convergent chain from the typically larger variation in a pre-convergent chain (Lambert et al., 2005). Other tools for monitoring convergence include Geweke statistic (Geweke, 1992), Heidelberger and

Welch test (Heidelberger and Welch, 1981, 1983), Raftery-Lewis convergence diagnostic (Raftery and Lewis, 1992, 1996), a test for convergence based on Gelman-Rubin statistics (Brooks and Gelman, 1998) and the trace plots of parameters of interest (or other important nuisance parameters) (Brooks and Gelman, 1998). The WinBUGS manual (Spiegelhalter et al., 2003) gives a rule of thumb as well:

> "the simulation should be run until the Monte Carlo error for each parameter of interest is less than about 5% of the sample deviation."

However, as Ghosh et al. (2006) state:

> "These diagnostics should not be taken as a proof of convergence of the chains, however if there were any problems, usually the diagnostic factors point to some potential problem."

Based on the results of our simulation and application, we strongly recommend that instead of just looking at the Gelman-Rubin statistics (GR-stat from now on), it is essential to visually inspect the trace plots of the parameters as well. We found that even when the GR-stat is close to 1 for all parameters in θ, the history plots for some of the parameters can show obvious non-convergence as the multiple chains do not mix at all. We give an example of such situation in Section 4.6.

4.4.3 Implementation

We use the WinBUGS (Lunn et al., 2000) software to carry out our MCMC simulation to sample from the posterior distribution. Because BEZI is not a standard distribution, we specified its density function in WinBUGS using the "zero trick" (Spiegelhalter et al., 2003).

4.5 Simulation

We conduct a simulation study to evaluate the performance of our model and method. We look at the simplest case with a single explanatory variable ($\mathbf{X_1} = (X_{11}, X_{21}, \cdots, X_{n1})$), so we have β_0 and β_1 as coefficients for the fixed effects in the linear predictor. We consider several parameter settings (θ) as described in Section 4.5.3. For each θ, we simulate multiple datasets with 100 observations. Fifty of them have $x_{t1} = 1$, the rest have $x_{t1} = 0$. We run four chains on each dataset, with 100,000 iterations per chain (the first 50,000 iterations are discarded as burn-in, following the rule of thumb given in Brooks and Gelman (1998)). We sample every 150^{th} iteration, i.e., thin=150, to decrease the dependence among samples. The four chains give 1333 samples in total. Inferences on a particular dataset are based on these samples.

Convergence is more difficult to achieve under some parameter settings, so we use varying numbers of simulated datasets to get 100 results that converge. We use the `R2WinBUGS` package (Sturtz et al., 2005) to conduct the MCMC simulation and model fitting by calling WinBUGS from `R`.

4.5.1 Initial Values

As mentioned earlier, we ran four chains per dataset. Therefore four sets of initial values are needed. We use:

1. Parameter values used for simulating datasets ("the true" values)

2. Method of moments estimates 1 (MOM1)

3. Method of moments estimates 2 (MOM2)

4. Maximum likelihood estimates (MLE)

The difference among the last three sets of initial values essentially have to do with ϕ. For the other parameters we use naive estimates defined as follows:

- For p:
$$p^{ini} = \frac{\sum_{t=1}^{n} I(y_t = 0)}{n} \text{ (sample proportion)}$$

- For β_0 and β_1:
$$\beta_0^{ini} = log \quad \frac{\mu_0}{1-\mu_0}$$
$$\beta_1^{ini} = log \quad \frac{\mu_1}{1-\mu_1} \quad - log \quad \frac{\mu_0}{1-\mu_0}$$
where $\mu_0 = \quad \frac{\bar{y}_0}{1-p^{ini}} \quad$ and $\mu_1 = \quad \frac{\bar{y}_1}{1-p^{ini}} \quad$ with $\bar{y}_0$ and $\bar{y}_1$ the average of y_t's with $x_{t1} = 0$ and $x_{t1} = 1$, respectively

- For σ^2 and φ: We fit an AR(1) model on the residuals, $R_t = y_t - \hat{y}_t = y_t - (1 - p^{ini})\hat{\mu}_t$, to obtain estimates for σ^2 and φ, where $\hat{\mu}_t$=μ_0 or μ_1 depending on the value of x_{t1}

To get initial values for ϕ, we have:

MOM1 Suppose we get independent samples from one BEZI population (so that $\varphi = 0$) and there is no measurement error ($\sigma^2 = 0$), then we use the mean and variance expressions for BEZI random variable to solve ϕ as:

$$\phi^{ini\_MOM1} = \frac{(1-p^{ini})\mu(1-\mu)}{s^2 - p^{ini}(1-p^{ini})\mu^2} - 1 \text{ where } \mu = \frac{\bar{y}}{1-p^{ini}}$$

p^{ini}, $\bar{y}$ and s^2 are the sample proportion, mean and variance

MOM2 Consider only the nonzero observations and use the mean and variance expressions for Beta random variable to solve ϕ as:

$$\phi^{ini\_MOM2} = \frac{\bar{y}_+(1-\bar{y}_+))}{s_+^2} - 1$$

where $\bar{y}_+$ and s_+^2 are the sample mean and variance of nonzeros

MLE Use maximum likelihood method to obtain initial value for ϕ which is calculated by `R` function `gamlssML(.)` in package `gamlss` (Mikis Stasinopoulos and Akantziliotou, 2011).

4.5.2 Information Recorded

For each simulated dataset, we record the posterior mean, posterior standard deviation, posterior median, 2.5% and 97.5% empirical simulation quantiles for each parameter in θ. Then for each setting, we take the averages of these qualities over the 100 simulations that have converged MCMC. The 2.5% and 97.5% quantiles provide an equal-tail 95% posterior interval estimator for each parameter. Therefore, we also record whether the 95% posterior intervals include the true parameter values and whether they include zero.

4.5.3 Parameter Settings

Table 4.3 gives the parameter settings we consider for our simulations. When $\beta_0 = -1$ and $\beta_1 = -3$, the sample means of BEZI are about 0.14. When $\beta_0 = 1$ and $\beta_1 = 2$ and $\beta_0 = 0.1$ and $\beta_1 = 0.5$, the sample means are greater than 0.6 and about 0.45, respectively. We evaluate more cases under $\beta_0 = -1$ and $\beta_1 = -3$ because intuitively when the mean of the non-zero data is small, overall it seems more likely that there would be zero-inflation.

Table 4.3: Simulation parameter setup. Within each settings, there are five different values for ϕ: 1, 5, 10, 20 and 40.

Parameter	Setting 1	Setting 2	Setting 3	Setting 4	Setting 5	Setting 6	Setting 7	Setting 8
β_0	1	-1	0.1	-1	1	-1	-1	-1
β_1	2	-3	0.5	-3	2	-3	-3	-3
p	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
ϕ	1,5,10, 20,40	1,5,10, 20,40	1,5,10, 20,40	1,5,10, 20,40	1,5,10, 20,40	1,5,10, 20,40	1,5,10, 20,40	1,5,10, 20,40
σ^2	1	1	1	1	1	0.25	1	0.0625
φ	0.5	0.5	0.5	0.25	0.25	0.5	0.75	0.5

We have several ways of grouping the settings in Table 4.3 to understand how the different parameter values influence our simulation results. Specifically, to see the effect of changing the β's, we compare settings 1, 2 and 3 or settings 4 and 5; to see the effect of changing ϕ, we compare the five sets within each setting; to see the effect of changing σ^2, we compare settings 2, 6 and 8; to see the effect of changing φ, we compare settings 2, 4 and 7, or setting 1 and 5.

4.5.4 Simulation Result

4.5.4.1 Convergence

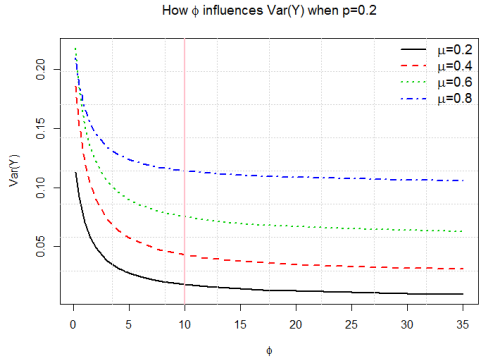
We use the GR-stat to monitor the convergence of the MCMC. According to Gelman et al. (2004), a GR-stat less than 1.1 typically means that the MCMC has converged. For fixed number of iteration, we found that marginally, σ^2 is the hardest parameter to get to converge. For all the settings, it almost always gets several GR-stat values that are greater than 1.1, with some value as large as 3 or 4, and occasionally even 6 or 7. β_0 also often has one or two GR-stat values that are greater than 1.1. Usually the large values are between 1.2 to 1.5 but occasionally as large as 1.8. By contrast, p never has any GR-stat value that is greater than 1.1 in our simulations. Sometimes ϕ and φ have one value around 1.3. β_1 also has value as big as 1.8, but usually its GR-stat values are smaller than 1.1.

When at least one of the parameters has a GR-stat greater than 1.1, we say the Markov chain does not converge to the posterior distribution. To fairly compare results we needed to run different numbers of simulations to get 100 converged MCMC for the different parameter settings. Table 4.4 gives the number of required simulations for each setting to have 100 convergent chains. Notice that for simulation purposes, we fixed the number of iterations, however, for a particular dataset, it is advisable to run a longer chain aiming to achieve convergence. Overall, settings 3 and 7 have the lowest numbers of required simulations, followed by settings 1 and 2. When ϕ=1, we almost always have one third of MCMC that do not converge. The general trend within the same setting is that when the value of ϕ increases, the number of required simulation decreases, and

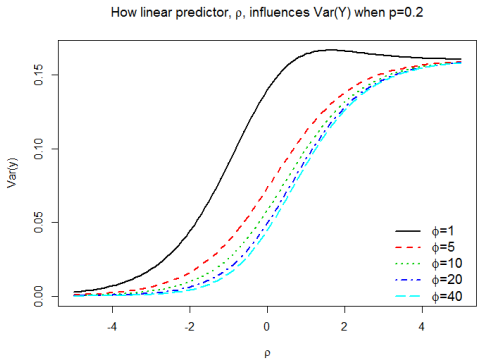
when ϕ is equal to or greater than 10, the numbers of required simulations are close. We think the reason for this is that as ϕ increases, the variance of Y decreases; however after ϕ is greater than 5, the variance does not change very much. See Figure 4.4b and Figure 4.4a.

Table 4.4: Number of simulations that are needed to achieve 100 convergent MCMC

ϕ	Setting 1	Setting 2	Setting 3	Setting 4	Setting 5	Setting 6	Setting 7	Setting 8
1	162	152	158	153	170	139	147	151
5	132	135	115	151	152	160	126	151
10	129	138	117	151	144	149	114	152
20	111	133	108	146	137	157	115	156
40	118	111	111	131	135	153	116	149



(a) How values of ϕ affect the variance: when ϕ as large as 10, the variance of y does not change much



(b) How values of linear predictor affect the variance: When $\phi = 1$, linear predictor has the biggest effect, the variance curve is quite different from others

Figure 4.4: How values of ϕ and linear predictor affect the variance of BEZI random variable

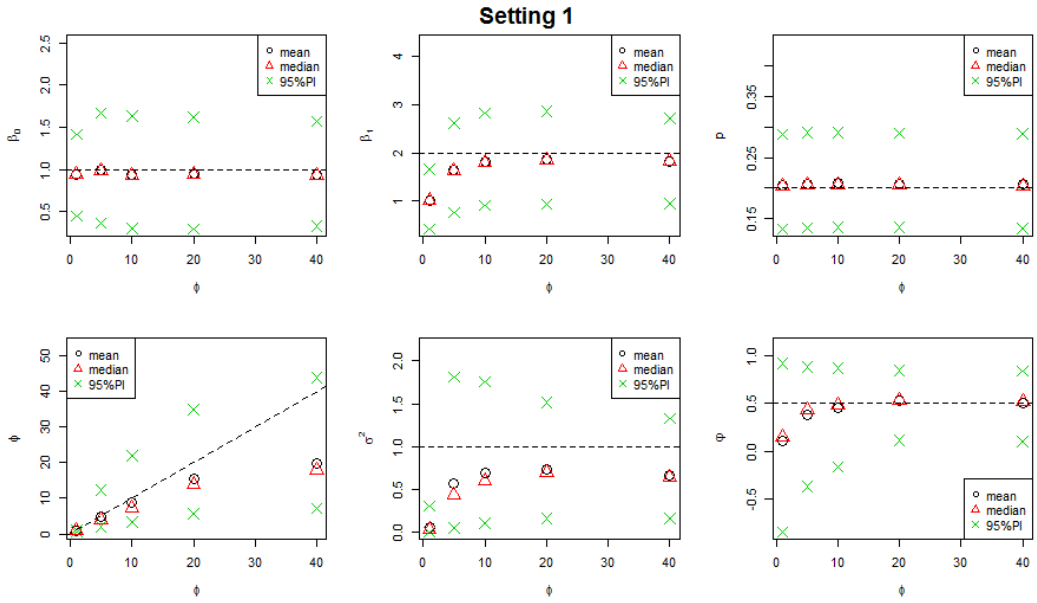


Figure 4.5: Averaged posterior quantities for setting 1 in Table 4.3. Reference lines give the true values. PI standards for posterior interval

4.5.4.2 Results

The posterior distributions for β_0, β_1 and p are quite symmetric, as their posterior means and medians are close to each other for all settings. For ϕ and σ^2, the posterior distributions have long tails on the right, and the posterior means are usually larger than the posterior medians. By contrast, the posterior distribution for φ is left skewed: the posterior medians are usually slightly greater than the posterior means. Most of the time, the posterior standard deviation of β_0, β_1 and p stay constant when ϕ increases. Occasionally, β_0 and β_1 have smaller standard deviation for ϕ=1 than for the other values of ϕ. The posterior standard deviation for ϕ increases when the value of ϕ increases. By contrast, the posterior standard deviations for σ^2 and φ usually drop when ϕ gets larger.

Some examples of the posterior quantities are given in Figure 4.5. For example, in the top left panel of Figure 4.5 the circles (means) and triangles (medians) are overlapped, and the distance between the pair of crosses (the length of the 95% posterior interval) has the smallest value for ϕ=1 and roughly remains constant for the other values of ϕ.

We first investigate how different values of β's affect the posterior interval coverage for themselves and other parameters. We first compare settings 1, 2 and 3. As we can see in Figure 4.6, there is no difference in the rates of covering the true values for β_0 and φ in the three settings. The rates are always greater than 0.9. For β_1 and p, setting 2 behaves differently from the other two: the rates of covering the true value are quite low except when $\phi = 40$ and they increase as the value of ϕ increases. The reason for this is likely that when $\beta_0 = -1$ and $\beta_1 = -3$, the mean of the Beta component (μ) is small; consequently, in simulations we get many observations that are close to zero, and by rounding, those values are treated as zeros, which creates problems for both p and ${\beta_1}^2$. Setting 1 also has low coverage (about 0.2) for β_1 when $\phi = 1$. In all three settings, the tendencies for ϕ and σ^2 are upward first then downward as the value of ϕ increases. The rate for σ^2 is extremely low when $\phi = 1$ in setting 1.

When we look at the rates of excluding zero (as shown in Figure 4.7), there is no difference in the three settings for p, ϕ and σ^2: all rates are 1. For β_0 and β_1, settings 1 and 2 are similar: the rates of excluding zero are about 0.9 for β_0 and 1 for β_1 except when ϕ=1. However, for setting 3, the rates are around 0.05 and 0.2 for β_0 and β_1, respectively. We suspect it is because in setting 3 the true values of the β's are not too

[2]Observations are used to estimate β_1 only when $x_{t1} = 1$, so there is less information for β_1 than for β_0. Also μ is smaller when $x_{t1} = 1$ than it is when $x_{t1} = 0$, which may give more close-to-zero observations in the first case.

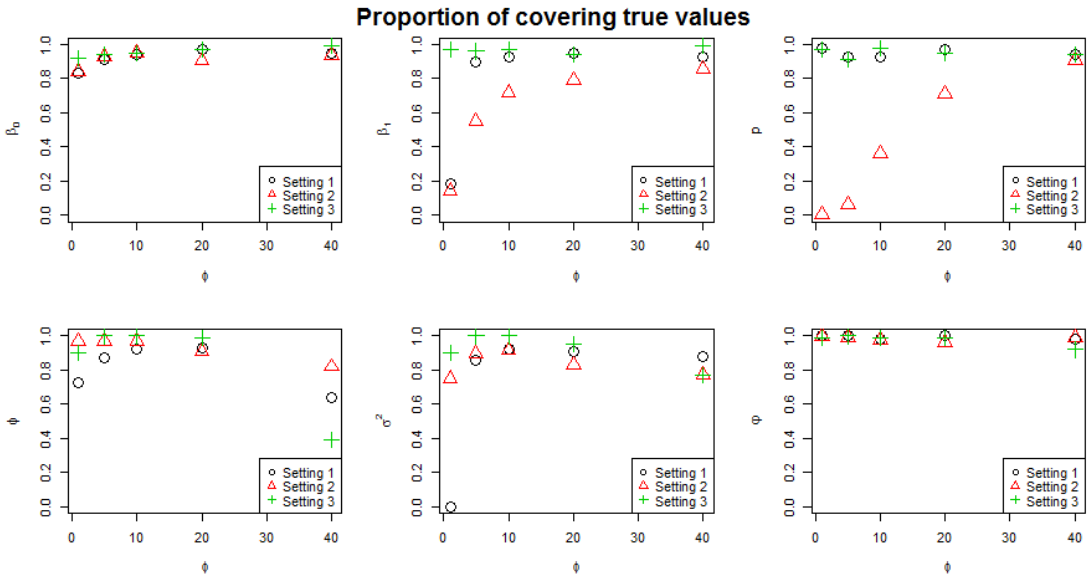


Figure 4.6: Rate of including the true value of parameter in the posterior interval; Settings 1, 2 and 3 have β_0 and β_1 in different values.

far from zero (as 0.1 for β_0 and 0.5 for β_1, respectively). In other words, it is possible the effect size is too small to detect. As ϕ increases, the rate of excluding zero for φ increases in all three settings; for a particular ϕ, setting 3 has the largest rate, setting 1 has the median one and setting 2 has the smallest one. The reason for this could be when the magnitudes of the β's are small, the signal in the AR correlation is enlarged so that it is easier to detect.

We now look at how values of other parameters alter the coverages of the posterior intervals for the β's. On the one hand, as illustrated in the top left panel of Figure 4.8, when the value of σ^2 changes (comparing settings 2, 6 and 8), the rate of covering the true value of β_0 does not vary much. Similarly, when the value of φ changes (bottom left panel of Figure 4.8, comparing settings 2, 4 and 7) the rate does not vary either. In

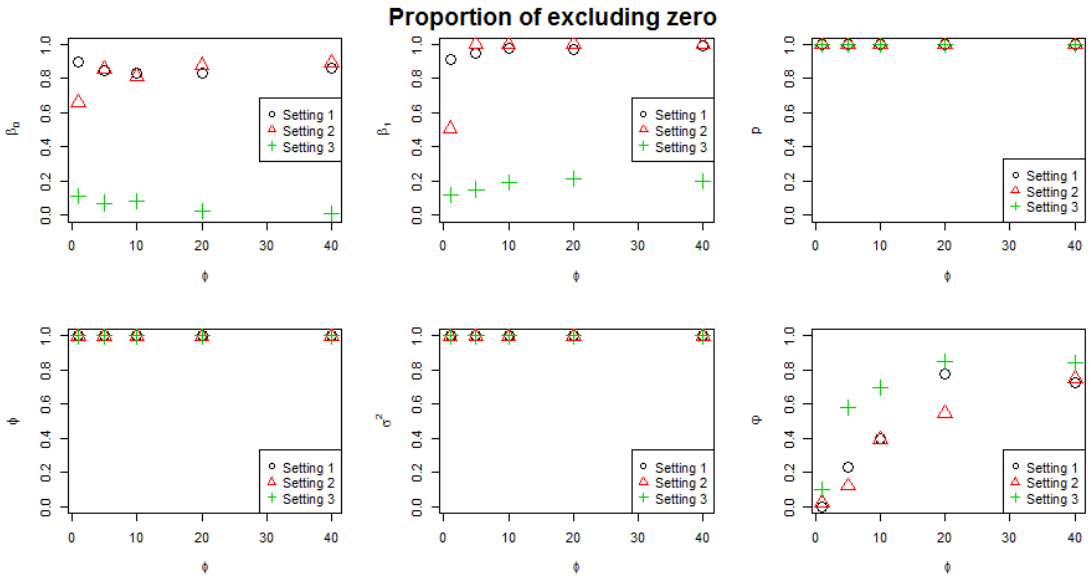


Figure 4.7: Rate of excluding zero in the posterior interval; Settings 1, 2 and 3 have β_0 and β_1 in different values.

both comparisons, increasing in ϕ does not affect the rates; they are almost always about 0.9. Whereas for β_1, setting 2 ($\sigma^2 = 1$) has a lower rate than settings 6 and 8, the later two have similar rates (top right panel of Figure 4.8). When the value of φ changes from 0.75 to 0.5 to 0.25, as shown in the bottom right panel of Figure 4.8, the rate of covering the true value for β_1 deceases. For all these groups, when the value of ϕ increases, on average the rates increase from 0.2 to 1.

On the other hand, when comparing the rate of excluding zero, see Figure 4.9, setting 2 has the lowest rate for β_0 among settings 2, 6 and 8 (the top left panel). The rates increase for all three settingss as ϕ increasing. For β_1, the rates are 1 in all cases except when $\phi = 1$. The rates for β_0 about 0.6 for all three settings (the top right panel). When the value of φ changes from 0.25 to 0.5 to 0.75, the rate of excluding zero for β_0

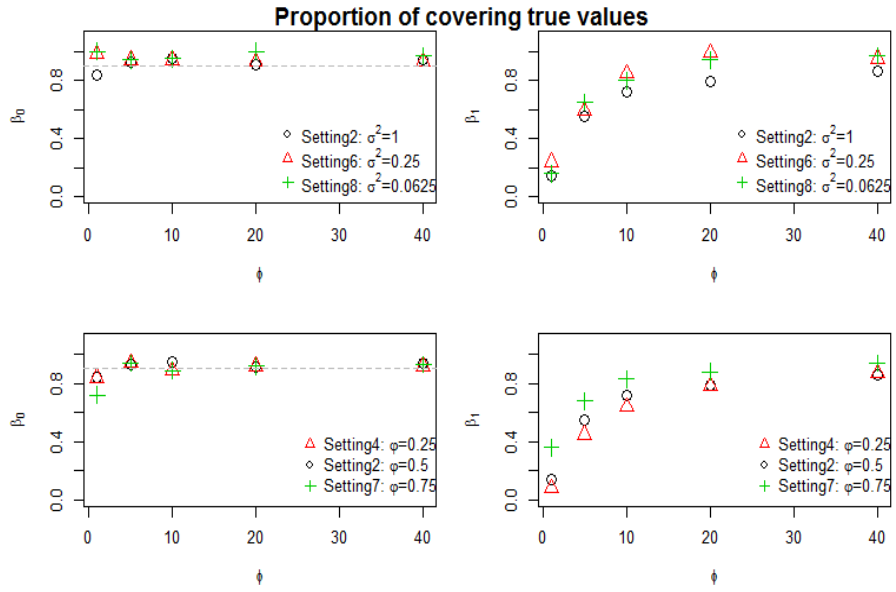


Figure 4.8: Rate of including the true value of β_0 and β_1 in the posterior intervals; true values of β_0 and β_1 are the same within each plot; the varied parameters are given in the legend.

decreases (the bottom left panel). $\varphi = 0.75$ also has the lower rate of excluding zero for β_1. However, there is no difference between $\varphi = 0.5$ and $\varphi = 0.25$ except when ϕ=1 (the bottom right panel). In this case, the rates in all three settings are less than 0.5.

Overall, ϕ=1 causes troubles in both coverage rates. We think this is related to the relatively large variation under such ϕ. In Figure 4.4b, the plot shows how the variance of a BEZI random variable changes against the value of the linear predictor. Notice that the curve for ϕ=1 is quite different from the other four.

In summary, the performance of our model seems quite sensitive to the true values of parameters. Under some values, it is difficult for the Markov chains to converge. For example, when $\phi = 1$, usually one third of the simulations do not converge. When $\phi \geq$

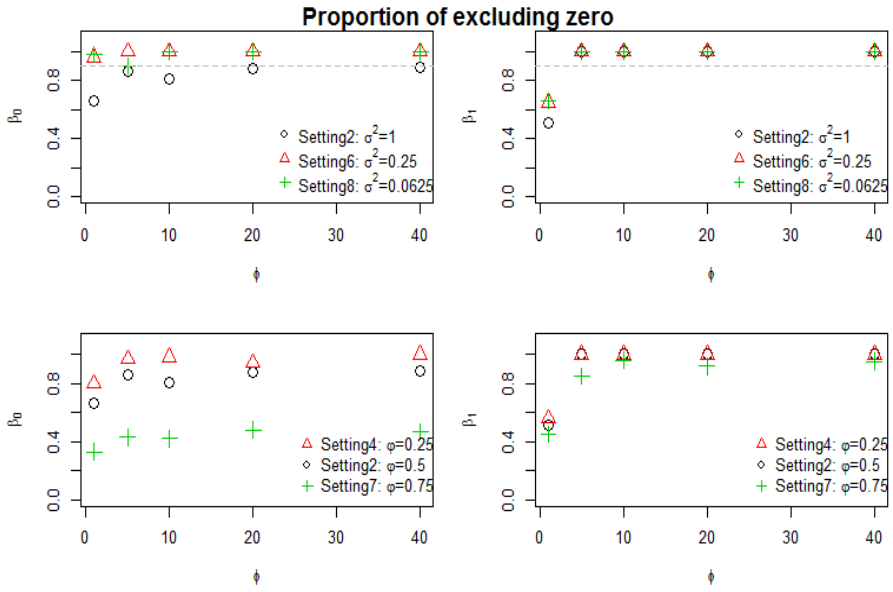


Figure 4.9: Rate of excluding zero for β_0 and β_1 in the posterior intervals; true values of β_0 and β_1 are the same within each plot; the varied parameters are given in the legend.

5 and the β's have large positive true values, the Bayesian estimates are accurate with reasonable precision. When the mean of the Beta component is small, caused by the small values of β_0 and β_1, the posterior intervals for β_1 and p exclude the true values quite often. When the β's are close to zero, the posterior intervals for them cover the true values, however, they cover zeros as well. In other words, the estimates are accurate but not very precise. When ϕ is small, the posterior interval for φ usually includes zero; it includes the true value as well. When the variance of the response decreases (ϕ increases), the performance improves.

4.6 Application to the Barnacle Settlement Data

In this section, we apply the generalized linear mixed model with autoregressive random effect to the oceanographic dataset described in Section 4.3. We consider the four study sites independently, as the researcher did. Because this is a real data example, we only have MOM1, MOM2 and MLE as initial values for our MCMC. For all applications, we start with three chains and run for 100000 iterations. The first 50000 are discarded as burn-in and we sample every 150^{th} iteration of the remaining. We run longer chains when it is necessary to achieve convergence. We provide the results for site SHB in Table 4.7.

We first consider the three physical processes one at a time as the explanatory variable. This is the situation we evaluated using simulation in Section 4.5. For the TPT site with diurnal upwelling process, the MCMC do not converge even for very long chains (400000 iterations with 200000 burn-in and 250 thin). The samples for β_0 and β_1 varied widely in the history plots. On a closer examination, we find that the settlement information consists of only zeros when there is no occurrence of diurnal upwelling, which likely explains the non-convergence. The result for TPT front looks suspicious because the posterior standard deviance is more than 100. On closer inspection, when there is occurrence of a front passage process, there are only zeros for settlement. So even though the MCMC indeed converged, the results are not very useful.

For all processes, the 95% posterior intervals for β_1 include zero, implying the settlement information is not different for time points with and without the occurrence of each physical process. In addition, the autoregressive component does not seem neces-

sary (95% posterior intervals include zero for φ).

Table 4.5: β's part in the linear predictor (ρ) when there are multiple explanatory variables (see equation (4.2)). The left panel has three explanatory variables and the right panel has two explanatory variables.

ρ_i	x_1	x_2	x_3	β's	ρ_i	x_1	x_2	β's
1	0	0	0	β_0	1	0	0	β_0
2	1	0	0	$\beta_0+\beta_1$	2	1	0	$\beta_0+\beta_1$
3	0	1	0	$\beta_0+\beta_2$	3	0	1	$\beta_0+\beta_2$
4	0	0	1	$\beta_0+\beta_3$	4	1	1	$\beta_0+\beta_1+\beta_2$
5	1	1	0	$\beta_0+\beta_1+\beta_2$				
6	1	0	1	$\beta_0+\beta_1+\beta_3$				
7	0	1	1	$\beta_0+\beta_2+\beta_3$				
8	1	1	1	$\beta_0+\beta_1+\beta_2+\beta_3$				

Table 4.6: Different ways of getting estimates of β's when there are multiple explanatory variables. Notations are consistent with Table 4.5.

β_i	Candidates when there are three explanatory variables				
β_0	ρ_1	$\rho_2+\rho_3-\rho_5$	$\rho_2+\rho_4-\rho_6$	$\rho_3+\rho_4-\rho_7$	$\rho_5+\rho_6+\rho_7-2\rho_8$
β_1	$\rho_2-\rho_1$	$\rho_5-\rho_3$	$\rho_6-\rho_4$	$\rho_8-\rho_7$	
β_2	$\rho_3-\rho_1$	$\rho_5-\rho_2$	$\rho_7-\rho_4$	$\rho_8-\rho_6$	
β_3	$\rho_4-\rho_1$	$\rho_6-\rho_2$	$\rho_7-\rho_3$	$\rho_8-\rho_5$	
β_i	Candidates when there are two explanatory variables				
β_0	ρ_1	$\rho_2+\rho_3-\rho_4$			
β_1	$\rho_2-\rho_1$	$\rho_4-\rho_3$			
β_2	$\rho_3-\rho_1$	$\rho_4-\rho_2$			

We also consider two and three explanatory variables together (for the LHP site, there were only regional relaxation and diurnal upwelling). These are situations we have not investigated by simulation. Notice that because we at most have three binary vectors, there are eight combinations for the fixed effect ($\mathbf{X_t}\beta$) in the linear predictor. If we just consider two of the three processes, there are four combinations. Because

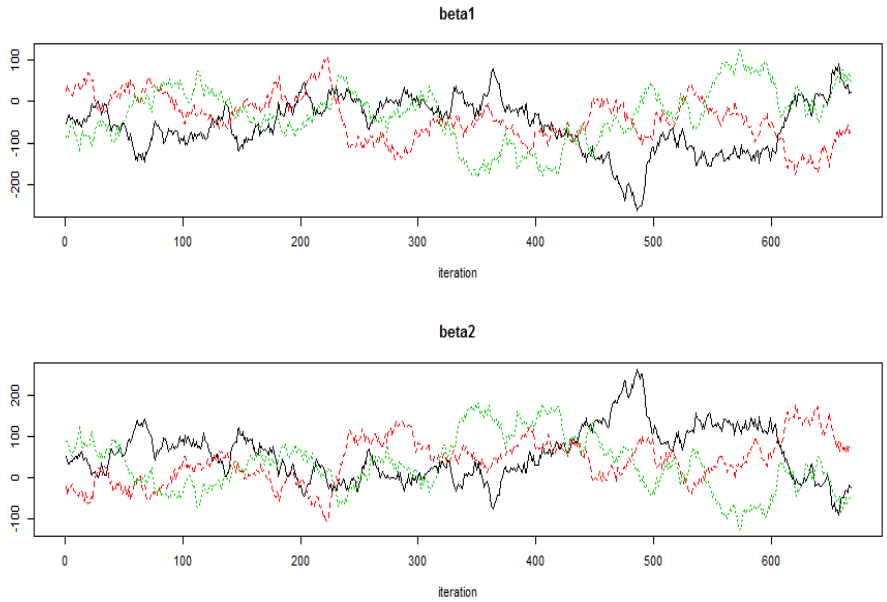


Figure 4.10: Trace plot for iterations of β_1 and β_2 in the BDB site. β_1 and β_2 are coefficients for relax and dup processes, respectively. GR-stat is less than 1.1 for all parameters. Trace plot shows non-convergence.

we assume an additive model, there are multiple ways of getting the initial estimates for the β's. For example, when there are two covariates, β_0 can be estimated by ρ_1 and $\rho_2 + \rho_3 - \rho_4$, (see the right panel of Table 4.5 and the bottom section of Table 4.6). To obtain initial values of β's for the MCMC, we first calculate all the candidates as indicted in Table 4.6, then we randomly take one candidate for each β as the initial value. The initial values for p, σ^2, φ and ϕ are calculated using the same method as in Section 4.5. For the TPT site and the BDB site, the MCMC with all three processes do not converge even for very long chains (8000000 iterations, 4000000 burn-in and 4000 thin). Hence we may consider two of the three processes at a time as it is in the LHP site. For TPT, the relax and dup combination and front and dup combination do

Table 4.7: Application result for the SHB site. For each model, we run three chains. Each chain has 100000 iterations (with 50000 burnin and 150 thin) unless specificed. GR-stat gives the average of Gelman and Rubin's diagnostic statistic (values are less than or close to 1 suggests convergence)

Parameter	Mean	SD	2.50%	25%	Median	75%	97.50%	GR
One process at a time								
Relax								
β_0	-1.967	0.387	-2.765	-2.172	-1.941	-1.729	-1.225	1.001
β_1	-0.640	0.506	-1.667	-0.970	-0.650	-0.296	0.390	1.002
p	0.484	0.064	0.360	0.442	0.482	0.529	0.606	0.999
ϕ	21.290	14.057	4.795	10.250	18.415	28.680	55.956	1.007
σ^2	0.625	0.452	0.006	0.280	0.601	0.875	1.665	1.027
φ	0.358	0.334	-0.502	0.184	0.394	0.592	0.911	1.006
Front								
β_0	-2.509	0.561	-3.627	-2.847	-2.476	-1.581	1.021	1.021
β_1	0.573	0.475	-0.356	0.248	0.558	0.889	1.528	1.005
p	0.483	0.0625	0.362	0.441	0.481	0.524	0.611	1.000
ϕ	20.287	13.238	4.670	10.030	17.980	26.867	54.836	1.001
σ^2	0.586	0.440	0.004	0.239	0.548	0.321	1.596	1.057
φ	0.380	0.363	-0.651	0.217	0.432	0.644	0.910	1.024
Dup (number of iterations = 120000)								
β_0	-1.877	0.431	-2.782	-2.118	-1.835	-1.587	-1.167	1.006
β_1	-0.289	0.415	-1.079	-0.5797	-0.293	-0.007	0.552	1.007
p	0.485	0.0626	0.362	0.444	0.485	0.528	0.601	1.000
ϕ	17.074	12.239	4.223	7.658	13.620	22.650	48.940	1.008
σ^2	0.529	0.462	0.002	0.113	0.463	0.802	1.646	1.047
φ	0.312	0.400	-0.723	0.125	0.383	0.597	0.897	1.019
Three processes together								
number of iterations =100000, burnin =50000, thin = 150								
β_0	-1.722	0.760	-3.287	-2.170	-1.682	-1.244	-0.310	1.008
β_1	-0.840	0.809	-2.352	-1.353	-0.859	-0.300	0.739	1.006
β_2	0.351	0.657	-0.905	-0.089	0.330	0.763	1.722	1.005
β_3	-0.807	0.470	-1.669	-1.136	-0.810	-0.492	0.108	1.005
p	0.484	0.064	0.362	0.440	0.485	0.527	0.609	1.003
ϕ	20.545	13.274	5.139	10.540	16.885	26.905	54.843	1.000
σ^2	0.560	0.417	0.005	0.215	0.515	0.833	1.488	1.010
φ	0.228	0.359	-0.598	0.009	0.267	0.475	0.852	1.004

not converge (400000 iterations, 200000 burn-in and 300 thin). β_0 and β_2 have very large GR-stats. We suspect zero-only settlement in the absence of diurnal upwelling is the reason. For the BDB relax and dup combination, the MCMC do not converge with 400000 iterations, 200000 burn-in and 300 thin. We found that despite the fact that the values of the GR-stat are all less than 1.1, the trace plots of β_1 and β_2 show non-convergence as the three chains do not mix at all (See Figure 4.10). At site BDB,

when dup=0 or relax=1 there is only zero settlement. This may be the reason for the non-convergence.

Again, none of the combination of processes seems important (95% posterior intervals include zero for all β_1, β_2 and β_3). There is no indication that the correlation part is needed either (95% posterior intervals include zero for φ).

Compared with the simulation study, in the real data application, obtaining convergent MCMC is more challenging. We suspect there are multiple reasons. First of all, the dataset has fewer number of observations (≤ 60) than the simulations. Besides, in this dataset there are lots of zeros as the response (proportion of zero is usually greater than 50%, sometimes greater than 80%). Those zeros can only provide knowledge about p. Even worse, those few non-zero settlement information are not evenly distributed in the fixed effect combinations. In other words, we are lack of information to update the knowledge about some parameters. The three facts together contribute to the non-convergence.

4.7 Discussion

We have developed a generalized linear mixed model with autoregressive random effect to model zero-inflated proportional data with serial correlation. Bayesian methodology is adopted for parameter estimation and statistical inferences. We discussed prior specification as well as monitoring the convergence of Monte Carlo Markov Chain that used to sample from posterior distribution. The model is implemented in WinBUGS; `R2WinBUGs` package is used to call WinBUGS from `R`. We have been evaluated our

model and method in simulations and applied them on an oceanographic data example.

According to the simulations, we found the model performance is sensitive to the true values of parameters. Marginally, σ^2 is the hardest parameter to get to converge, followed by the coefficients of the fixed effects (β's); whereas p is the easiest one to converge. In addition, the number of iterations for convergence seems to depend on the true values of parameters. The cases with small value of ϕ are usually hard to converge. Also, when φ is small, convergence needs a longer chain. In addition, with respect to the coverage of posterior intervals, φ is often not different from 0, though its posterior interval almost always covers the true value. When the mean of the Beta components is small, it is quite often that the posterior interval of p does not include the true value. Sometimes the intervals for β_1 and β_0 also exclude their true values and their coverage changes when other parameters vary. When the value of ϕ increases, the performance of the method usually improves. We expect the performance to improve with a larger sample size.

There are several directions for future work. The first is to investigate additional priors. Because most non-informative priors are not invariant to transformation, arbitrarily chosen priors may create extra problem for generalized linear mixed model. We started to investigate several opinions for ϕ, however, more systematic work should be done to further understand the impact of prior specification for this precision parameter. Meanwhile, in this work, we assume independent priors. However, we observed in our simulations that the marginal posterior distributions of p and the β's interact with each other since poor performed posterior intervals for p comes with small true value of the β's. A joint prior may have to be found to improve the method. Also, monitoring conver-

gence is essential to Bayesian approach so we may need more guidelines for evaluating convergence. Statistics like Gelman-Rubin diagnostic statistic provide simple numbers to look at, but, it is crucial to actually look at the history plots also. When possible, one should check the marginal convergence of all parameters, not just the ones of interest. It would also be interesting to develop classical methods for the same generalized linear mixed model and compare their performance to the Bayesian approach. Another possibility is to consider linking the explanatory variables to multiple parameters in the BEZI density. In this work we assume constant p and ϕ for all observations. However, these two parameters can be related to explanatory variables as well and the serial correlation could also affect them. One may consider linking p and/or ϕ with explanatory variables and serial correlation as more complex models. Again, Bayesian approach can be adopt for parameter estimation. Achieving convergence could be more challenge for MCMC as the posterior distribution is expected to be more complicated than the current model.

4.8 Appendix

4.8.1 Mean and Variance of Y_t

The unconditional mean and variance of η_t are $E(\eta_t)$=0, $Var(\eta_t) = \frac{\sigma^2}{1-\varphi^2}$, so that,

$$\eta_t \sim N(0, \frac{\sigma^2}{1-\varphi^2}) \text{ for } t = 1, \cdots, n.$$

As a result,

$$\rho_t \sim N(\mathbf{X}_t\beta, \frac{\sigma^2}{1-\varphi^2}).$$

Given $h(\rho_t)$ in 4.3 and using the delta method we have

$$h(\rho_t) - h(\mathbf{X}_t\beta) \sim N \quad 0, \frac{\sigma^2}{1-\varphi^2} \quad h^{(1)}(\mathbf{X}_t\beta) \quad ^2 \quad ,$$

where

$$h^{(1)}(\mathbf{X}_t\beta) = \frac{exp(\mathbf{X}_t\beta)}{(1+exp(\mathbf{X}_t\beta))^2}$$

is the first derivative of $h(\rho_t)$ with respect to ρ_t. Therefore,

$$\mu_t - h(\mathbf{X}_t\beta) \sim N \quad 0, \frac{\sigma^2}{1-\varphi^2}\frac{(exp(\mathbf{X}_t\beta))^2}{(1+exp(\mathbf{X}_t\beta))^4} \quad ,$$

that is

$$E(\mu_t) = \frac{exp(\mathbf{X}_t\beta)}{1+exp(\mathbf{X}_t\beta)},$$

$$Var(\mu_t) = \frac{\sigma^2}{1-\varphi^2}\frac{(exp(\mathbf{X}_t\beta))^2}{(1+exp(\mathbf{X}_t\beta))^4}.$$

$$E(Y_t) = E\left(E(Y_t|\mu_t)\right) = E((1-p)\mu_t) = (1-p)\frac{exp(\mathbf{x}_t\beta)}{1+exp(\mathbf{x}_t\beta)},$$

$$Var(Y_t) = E\left(Var(Y_t|\mu_t)\right) + Var\left(E(Y_t|\mu_t)\right) = \frac{1-p}{\phi+1}E(\mu_t(1-\mu_t)) + p(1-p)E(\mu_t^2)$$

$$+Var((1-p)\mu_t)$$

$$= \frac{1-p}{\phi+1}E(\mu_t - \mu_t^2) + p(1-p)E(\mu_t^2) + (1-p)^2\left(E(\mu_t^2) - E^2(\mu_t)\right)$$

$$= \frac{1-p}{\phi+1}E(\mu_t) - (1-p)^2E^2(\mu_t) + \ -\frac{1-p}{\phi+1} + p(1-p) + (1-p)^2 \ \ E(\mu_t^2)$$

$$= \frac{1-p}{\phi+1}\frac{exp(\mathbf{x}_t\beta)}{1+exp(\mathbf{x}_t\beta)} - (1-p)^2 \ \ \frac{exp(\mathbf{x}_t\beta)}{1+exp(\mathbf{x}_t\beta)}^{\ 2}$$

$$+\frac{(1-p)\phi}{\phi+1}\frac{(exp(\mathbf{x}_t\beta))^2}{(1+exp(\mathbf{x}_t\beta))^2} \ \ 1+\frac{\sigma^2}{(1-\varphi^2)(1+exp(\mathbf{x}_t\beta))^2} \ \ .$$

5 Discussion

My dissertation is a collection of methods for analyzing serially correlated zero-inflated proportion data. It is motivated by a marine science example in which the original question of interest is addressed by comparing two samples. However, because of the sampling procedure, the observations within and across samples are serially correlated and there are some uncertainties in the actual determination of two samples. Due to data manipulation, the majority of observations are zero-inflated proportions. In addition, the observations were collected in unequally spaced sampling intervals. The researchers who collected these data used standard nonparametric tests that focus on location parameters to compare two samples. However, the analysis of these data is complicated by non-normality and autocorrelation, as well as uneven sampling intervals. This is not typically a case that a permutation t-test or a Wilcoxon rank sum test can answer adequately. In this work, we develop three more sophisticated approaches for analyzing data of this nature.

5.1 Summary

Chapter 2 focuses on simultaneously comparing multiple features of two populations of zero-inflated proportions. We look at a two-sample comparison problem assuming the two samples are correctly identified and the observations are independent. Using zero-inflated Beta as underlying distribution family, we first propose multiple parametric

test statistics that not only compare location parameters but also scale parameters and zero-inflation proportions. We adopt Fisher's method to combine multiple individual test statistics, each emphasizing the comparison between one pair of parameters. We consider both likelihood ratio test statistics and score test statistics. We then propose two non-parametric and two semi-parametric tests as alternatives. These are based on permutation (relabeling observations) and can be expended to multiple types of data inflation problems. Our methods are easy to implement, computationally efficient and can be expanded to more than two populations and to other multiple parameter families. Simulations showed that the likelihood-based tests perform well for large sample sizes and that the statistics based on combining likelihood ratio test statistics outperforms the ones based on combining score test statistics. The permutation-based tests have overall better performance in terms of both power and type I error rate.

Chapter 3 concerns both the non-normality and the autocorrelation among observations with the two samples. We use a Hidden Markov Model with zero-inflated Beta emission densities to model zero-inflated proportion data with serial correlation. We show that the standard EM algorithm for Hidden Markov Model parameter estimation can be applied in this case with emission distributions that are mixtures of discrete and continuous parts. Our simulations show the effectiveness of this approach. We find that the initial values, the number of observations and the BEZI density shape all impact the performance of the EM algorithm. We provide some suggestions about the choice of initial values. We compare the Viterbi algorithm and posterior decoding method for decoding the unobserved states. When decoding a hidden state chain, the Viterbi algorithm gives more accurate identifications than does posterior decoding. However, the

Viterbi algorithm is sensitive to BEZI density shapes. For both the EM algorithm and the Viterbi algorithm, asymmetric BEZI densities provide a lesser challenge than other density shapes.

Chapter 4 is also concerned with the non-normality and autocorrelation of our samples. We address the question of interest using a generalized linear mixed model with Bayesian approach for parameter estimation and statistical inference. The response distribution is zero-inflated Beta, conditional on realizations of autoregressive random effect. We provide guidelines of monitoring the convergence of our MCMC used to simulate from the posterior distribution. We use simulations to evaluate our method. We find that fixed effect coefficients and the variance parameter in the autoregressive random effect are more likely to have convergence problems than the other parameters. Also, the accuracy and precision of the method is sensitive to the true values of parameters. The autoregressive parameter, φ, is often not different from 0, despite its posterior interval almost always covers its true value. When the mean of the Beta component is small, the posterior interval of zero proportion often excludes its true value. Sometimes intervals of fixed effect coefficients also exclude their true values. When the value of precision parameter of the Beta component, ϕ, increases, the performance of the method improves. In addition, several times we find that it is possible that after many iterations of the MC, all the Gelman-Rubin statistics are close to 1 but the trace plots are not well mixed. We strongly recommend that researchers use visual inspection to assist diagnostic statistics in monitoring convergence of the MCMC.

5.2 A Comparison between Hidden Markov Models and Generalized Linear Mixed Models

As discussed by Cox (1981), time series analysis for dependent data can be categorized as observational-driven or parameter-driven. Parameter-driven models introduce the autocorrelation through a latent process, whereas observation-driven models on the other hand define the autocorrelation in the observations directly (i.e., the distribution of a variable, Y_t, is a function of previous observations, $Y_1, \cdots, Y_{t-1}$) Hidden Markov Model (Rabiner, 1989) is an example of parameter-driven model and generalized linear mixed model is an example of observation-driven model.

On the one hand, the first advantage of using Hidden Markov Models in a setting like the motivating example is that Hidden Markov Models let the data identify the two populations themselves and the autocorrelation among observations are considered through the hidden stochastic process by default. In addition, the decoded hidden state chain could inspire further research when none of the pre-determined explanatory processes are related to the ecological outcomes. Most importantly, the states in the state chain are correlated because of the property of Markov process. Therefore, it may provide more reasonable explanations for the natural phenomena.

On the other hand, in generalized linear mixed models, multiple predictors can be considered simultaneously, as well as their interaction terms when necessary. Continuous valued predictors are also valid in the generalized linear mixed models. By contrast, in Hidden Markov Models, because the autocorrelation is introduced through a latent process, it is more convenient to just consider one or two predictors with discrete sup-

port so that the underlying hidden process has a tractable number of states. Another advantage of generalized linear mixed models is that it is possible to extend them to cluster data situation in which correlations are generated because of repeated measurements or grouping.

Hidden Markov Models are more useful when researchers are not only interested in the autocorrelation between response but also predictors. Generalized linear mixed models are more useful when researchers want to make inference for particular predictors. Both types of models have their own assumptions. Generalized linear mixed models usually assume a particular form of random effect, for example, normally distributed. Hidden Markov Models have no distribution assumption about the autocorrelation; however they do assume certain properties in their transition matrix, such as a one step stationary transition matrix. Both models have their place in practice. Researchers should choose models that match their data collecting methods and questions of interest.

5.3 Future Work

Some interesting extensions of this work include:

- Expand Hidden Markov Model to unequally spaced situation

  An extra layer could be added to the zero-inflated Beta Hidden Markov Model to account for unequal spaces between observations by thinking of the irregularly spaced observations as coming from a chain with regularly spaced observations with some observations missing. Between the population states and the observational sequence, we could consider another structure that controls whether at any

given time, t, the observation is actually observed. Therefore, when there are observations missing at some t, the sequence becomes unequally spaced. Because the EM algorithm is developed for situations with missing data, it is a natural extension to the case of unequally spaced intervals.

- Expand generalized linear mixed model to clustered random effect

  In chapter 4, we considered a generalized linear mixed model with autoregressive random effects. Another way of getting correlations is through clusters. Examples are blocks in experimental design study or subjects in longitudinal study. In the first case, plots in the same block are usually correlated, whereas in the second example, measurements taken on the same subject are typically correlated. It would be interesting to develop methods for zero-inflated Beta generalized linear mixed model with clustered random effect.

- Develop estimation methods for generalized linear mixed models in classical statistics framework

  We developed Bayesian approach for a zero-inflated Beta generalized linear mixed model. It would be interesting to develop estimating methods under classical statistic framework (linearizing the conditional mean or numerical approximating integrals in the marginal likelihood) and compare the performance of the different methods.

Bibliography

Aittokallio, T. and Uusipaikka, E. (2000). Computation of standard errors for maximum-likelihood estimates in Hidden Markov Models. *Turku Centre for Computer Science, Technical Report No 379, University of Turku*.

Anderson, D. A. and Aitkin, M. (1985). Variance component models with binary response: Interviewer variability. *Journal of the Royal Statistical Society. Series B (Methodological)*, 47(2):203–210.

Artin, E. (1964). *The Gamma Function*. Holt, Rinehart and Winston, Inc., New York.

Baker, J. K. (1975). The dragon system - An overview. *IEEE Transations on Acoustics, Speech, and Signal Processing*, ASSP-23(1):24–29.

Baker, S. G. (1992). A simple method for computing the observed information matrix when using the EM algorithm with categorical data. *Journal of Computational and Graphical Statistics*, 1(1):63–76.

Baldi, P. and Chauvin, Y. (1994). Hidden Markov Models of biological primary sequence information. *Neural Computation*, 6:307–318.

Baldi, P., Chauvin, Y., Hunkapiller, T., and McClure, M. A. (1994). Hidden Markov Models of biological primary sequence information. *Proceedings of the National Academy of Sciences, USA*, 91(3):1059–1063.

Barry, S. C. and Welsh, A. H. (2002). Generalized additive modelling and zero inflated count data. *Environmentrics*, 157:179–188.

Baum, L. E. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities III*, pages 1–8.

Baum, L. E. and Eagon, J. A. (1967). An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bulletin American Mathematical Society*, 73:360–363.

Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *The Annuals of Mathematical Statistics*, 37:1554–1563.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annuals of Mathematical Statistics*, 41(1):164–171.

Bishop, M. and Thompson, E. (1986). Maximum likelihood alignment of DNA sequences. *Jounal of Molecular Biology*, 190(2):159–165.

Boyd, S. and Lieven, V. (2004). *Convex Optimization*. Cambridge University Press, The Edinburgh Building, Cambridge, CB2 8RU, UK, 1 edition.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208.

Cappè, O., Moulines, E., and Ryden, T. (2005). *Inference in Hidden Markov Models*. Springer Science and Business Media, Inc., New York, NY, 1 edition.

Chiogna, M. and Gaetan, C. (2007). Semiparametric zero-inflated Poisson models with application to animal abundance studies. *Environmentrics*, 18:303–314.

Chree, C. (1913). Some phenomena of sunspots and of terrestrial magnetism at Kew Observatory. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 212:75–116.

Chree, C. (1914). Some phenomena of sunspots and of terrestrial magnetism. Part II. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 213:245–277.

Christensen, R., Johnson, W., Branscum, A., and Hanson, T. E. (2011). *Bayesian Ideas and Data Analysis, an Introduction for Scientists and Statisticians*. CRC Press, Florida, USA, 1 edition.

Churbanov, A. and Winters-Hilt, S. (2008). Implementing EM and Viterbi algorithms for Hidden Markov Model in linear memory. *BMC Bioinformatics*, 9(224).

Conniffe, D. (2001). Score tests when a nuisance parameter is unidentified under the null hypothesis. *Journal of Statistical Planning and Inference*, 97:67–83.

Cox, D. R. (1981). Statistical analysis of time series: Some recent developments. *Scandinavian Journal of Statistics*, 8(2):93–115.

Cramer, H. (1946). *Mathematical Methods of Statistics*. Princeton Univ Press.

Cribari-Neto, F. and Vasconcellos, K. L. P. (2002). Nearly unbiased maximum likelihood estimation for the Beta distribution. *Journal of Statistical Computation and Simulation*, 72(2):107–118.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Dongen, S. V. (2006). Prior specification in Bayesian statistics: Three cautionary tales. *Journal of Theoretical Biology*, 242:90–100.

Durairajan, T. M. (1985). Bahadur efficient test for the parameters on inverse Gaussian distribution. *Journal of the Indian Society of Agricultural Statistics*, 37(2):192–197.

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge and New York.

Ferrari, S. L. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815.

Feuerverger, A. (1979). On some methods of analysis for weather experiments. *Biometrika*, 66(3):655–658.

Fisher, R. A. (1922). On the interpretation of χ^2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, 85(1):87–94.

Fisher, R. A. (1950). *Statistical Methods for Research Workers*. Oliver and Boyd; 11th edition, London.

Fong, Y., Rue, H., and Wakefield, J. (2010). Bayesian inference for generalized linear mixed models. *Biostatistics*, 11(3):412–397.

Freedman, D. A. (2007). How can the score test be inconsistent? *The American Statistician*, 61(4):291–295.

Gelman, A. (2006). Prior distribution for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–533.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman and Hall Ltd, Cambridge, UK, 2 edition.

Gelman, A., Hill, J., and Yajima, M. (2012). Why we usually don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, 5:189–211.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–511.

Geweke, J. (1992). *Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments*. Clarendon Press, Oxford,UK.

Ghosh, S. K., Mukhopadhyay, P., and Lu, J. C. (2006). Bayesian analysis of zero-inflated regression models. *Journal of Statistical Planning and Inference*, 136:1360–1375.

Hall, D. B. (2000). Zero-inflated Poisson and binomial regression with random effect: a case study. *Biometrics*, 56:1030–1039.

Härdle, W., Horowitz, J., and Kreiss, J. P. (2003). Bootstrap methods for time series. *International Statistical Review*, 71:435–459.

Hay, J. L. and Pettitt, A. N. (2001). Bayesian analysis of a time series of counts with covariates: an application to the control of an infectious disease. *Biostatistics*, 2(4):433–444.

Heidelberger, P. and Welch, P. D. (1981). A spectral method for confidence interval generation and run length control in simulations. *Communications of the ACM - Special Issues on Simulation Modeling and Statistical Computing*, 24(4):233–245.

Heidelberger, P. and Welch, P. D. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, 31(6):1109–1144.

Horst, R. and Pardalos, P. (1995). *Handbook of Global Optimization*. Kluwer Academic Publishers, Dordrecht, 1 edition.

Jamshidian, M. and Jennrich, R. I. (2000). Standard errors for EM estimation. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 62(2):257–270.

Jazi, M. A., Jones, G., and Lai, C. D. (2012). First-order integer valued AR processes with zero inflated Poisson innovations. *Journal of Time Series Analysis*, 33:954–963.

Juang, B. H., Levinson, S. E., and Sondhi, M. M. (1986). Maximum likelihood estimation for multivariate mixture observations of Markov chains. *IEEE Transations on Information Theory*, 32(2):307–309.

Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91:1343–1370.

Krogh, A., Brown, M., Mian, I. S., Sjolander, K., and Haussler, D. (1994). Hidden Markov Models in computational biology: Applications to protein modeling. *Journal of Molecular Biology*, 235(5):1501–1531.

Lambert, P. C., Sutton, A. J., Burton, P. R., Abrams, K. R., and Jones, D. R. (2005). How vague is vague? a simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine*, 24:2401–2428.

Lawrence, A. (1982). The innovation distribution of a Gamma distributed autoregressive process. *Scandinavian Journal of Statistics*, 9(4):234–236.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*. Springer & Verlag, New York, 2 edition.

Liporace, L. A. (1982). Maximum likelihood estimation for multivariate observations of Markov sources. *IEEE Transactions on Information Theory*, 28(5):729–734.

Littell, R. C. and Folks, J. L. (1971). Asymptotic optimality of Fisher's method of combining independent tests. *Journal of American Statistical Association*, 66(336):802–806.

Littell, R. C. and Folks, J. L. (1973). Asymptotic optimality of Fisher's method of combining independent tests II. *Journal of American Statistical Association*, 68(341):193–194.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., and Schabenberger, O. (2006). *SAS for Mixed Models*. SAS Institute Inc., Cary, NC, 2 edition.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Liu, H. and Chan, K. S. (2008). Constrained generalized additive modelling with zero inflated data. *Technical Report, The University of Iowa, Department of Statistics and Actuarial Science*, 388.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10:325–337.

Lystig, T. C. and Hughes, J. P. (2002). Exact computation of the observed information matrix for hidden markov models. *Journal of Computational and Graphical Statistics*, 11(3):678–689.

Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2(1):49–55.

Mamon, R. S. and Elliott, R. J., editors (2007). *Hidden Markov Models in Finance*. International Series in Operations Research & Management Science. Springer, 1 edition.

Marin, T. G., Wintle, B. A., Rhodes, J. R., Kuhnert, P. M., Field, S. A., Low-Choy, S. J., Tyre, A. J., and Possingham, H. P. (2005). Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. *Ecology Letters*, 8:1235–1246.

McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. (2008). *Generalized, Linear, and Mixed Models*. John Wiley & Sons, Inc, New Jersey, 2 edition.

McGilchrist, C. (1995). The derivation of BLUP, ML, REML estimation methods for generalised linear mixed models. *Communications in Statistics - Theory and Methods*, 24(12):2963–2980.

McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2 edition.

Mikis Stasinopoulos, B. R. and Akantziliotou, C. (2011). *Generalized Additive Models for Location Scale and Shape*. R package version 4.1-1.

Montgomery, D. C., Jennings, C. L., and Kulahci, M. (2008). *Introduction to Time Series Analysis and Forecasting*. John Wiley & Sons, Inc., New Jersey, USA, 1 edition.

Morgan, B. J. T., Palmer, K. J., and Ribout, M. S. (2007). Negative score test statistics. *The American Statistician*, 61(4):285–288.

Mukhopadhyay, N. (2000). *Probability and Statistical Inference*. Marcel Dekker, Inc., New York, NY.

Natarajan, R. and Kass, R. E. (2000). Reference Bayesian methods for generalized linear mixed models. *Journal of the American Statistical Association*, 95(449):227–237.

Nelder, J. and Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384.

Nilsson, M. (2005). *First Order Hidden Markov Model [Elektronisk resurs] : Theory and Implementation Issues*. Research Report. Blelange Institute Technology, Sweden.

Nocedal, J. and Wright, S. J. (1999). *Numerical Optimization*. Springer - Verlag, New York.

Ospina, R. and Ferrari, S. L. P. (2010). Inflated Beta distribution. *Statistical Papers*, 51(1):111–126.

Ospina, R. and Ferrari, S. L. P. (2012). A general class of zero-or-one inflated Beta regression models. *Computational Statistics and Data Analysis*, 56:1609–1623.

Pace, L. and Salvan, A. (1997). *Principles of Statistical Inference from a Neo-Fisherian Perspective*. World Scientific Publishing Company; 1st edition, Singapore.

Pardo, B. and Birmingham, W. (2005). Modeling form for on-line following of musical performances. *Proceedings of the Twentieth National Conference on Artificial Intelligence*. Pittsburgh, Pennsylvania.

Paul, S. R. and Jiang, X. (2005). Testing the homogeneity of several two-parameter populations. *The Canadian Journal of Statistics*, 33(1):131–143.

Petrushin, V. A. (2000). Hidden Markov Models: Fundamentals and applications. *Online Symposium for Electronics Engineer*.

Phil, B. (2004). Hidden Markov Models. *Department of Computer Science and Software Engineering, The University of Melbourne*.

Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, 4(1):12–35.

Pinheiro, J. C. and Chao, E. C. (2006). Efficient Laplacian and adaptive Gaussian quadrature algorithm for multilevel generalized linear mixed models. *Journal of Computational and Graphical Statistics*, 15(1):58–81.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rabiner, L. R. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Rabiner, L. R. and Juang, B. H. (1993). *Fundamentals of Speech Recognition*. PTR Prentice Hall.

Rabiner, L. R., Juang, B. H., Levinson, S. E., and Sondhi, M. M. (1985). Some properties of continuous Hidden Markov Model representations. *AT&T Technical Journal*, 64(6):1251–1270.

Raftery, A. E. and Lewis, S. M. (1992). One long run with diagnostics: Implementation strategies for Markov Chain Monte Carlo. *Statistical Science*, 7(4):493–497.

Raftery, A. E. and Lewis, S. M. (1996). *The Number of Iterations, Convergence Diagnostics and Generic Metropolis Algorithms*. Chapman & Hall Press, London,UK.

Ramsey, F. L. and Schafer, D. W. (2002). *The Statistical Sleuth, A Course in Methods of Data Analysis*. DUXBURY, Pacific Grove, CA, 2 edition.

Ridout, M., Demetrio, C. G. B., and Hinde, J. (1998). Models for count data with many zeros. *In Proceedings of the XIXth International Biometric Conference*, pages 179–192.

Rigoll, G., Kosmala, A., Rottland, J., and Neukirchen, C. (1996). A comparison between continuous and discrete density Hidden Markov Models for cursive handwriting recognition. *Pattern Recognition, Proceedings of the 13th International Conference on Pattern Recoginition*, 2:205–209.

Robert, C. P., Ceneux, G., and Diebolt, J. (1993). Bayesian estimation of hidden Markov chains: A stochastic implementation. *Statistics and Probability Letters*, 16:77–83.

Robinson, T. J., Anderson-Cook, C. M., and Hamada, M. S. (2009). Bayesian analysis of split-plot experiments with nonnormal responses for evaluating nonstandard performance criteria. *Technometrics*, 51(1):56–65.

Rocha, A. V. and Cribari-Neto, F. (2009). Beta autoregressive moving average models. *Test*, 18:529–545.

Ross, S. M. (2010). *Introduction to Probability Models*. Academic Press, Oxford, UK, 10 edition.

Satish, L. and Gururaj, B. I. (1993). Use of Hidden Markov Models for partial discharge pattern classification. *IEEE Transactions on Electrical Insulation*, 28(2):172–182.

Scott, S. L. (2002). Bayesian methods for Hidden Markov Models: Recursive computing in the 21th century. *Journal of the American Statistical Association*, 97(457):3i7–351.

Singh, N. (1986). A simple and asymptotically optimal test for the equality of normal population: A pragmatic approach to one-way classification. *Journal of the American Statistical Association*, 81(395):703–704.

Sorenson, H. W. and Alspach, D. L. (1971). Recursive Bayesian estimation using Gaussian sums. *Automatica*, 7(4):465–479.

Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. (2003). WinBUGS user manual, version 1.4. *Technical Report*.

Starner, T. and Pentland, A. (1995). Real-time American sign language visual recognition from video using Hidden Markov Models. Master's thesis, MIT. Program in Median Arts.

Sturtz, S., Ligges, U., and Gelman, A. (2005). R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software*, 12(3):1–16.

Thiagarajah, K. (2012). Combined tests for the homogeneity of Weibull (or extreme value) populations with censored data. *Journal of Statistical Theory and Practice*, 6(4):783–792.

Tian, L. (2005). Inferences on the mean of zero-inflated lognormal data: the generalized variable approach. *Statistics in Medicine*, 24:3223–3232.

Turner, R. (2008). Direct maximization of the likelihood of a hidden markov model. *Computational Statistics and Data Analysis*, 52:4147–4160.

Tyburczy, J. A. (2011). *Nearshore Distribution of Barnacle and Mussel Larvae and Oceanographic Mechanisms of Onshore Transport and Delivery*. PhD thesis, Oregon State University.

Verbeke, G. and Molenberghs, G. (2007). What can go wrong with the score test? *The American Statistician*, 61(4):289–290.

Viera, A. J. and Garrett, J. M. (2005). Understanding interobserver agreement: The Kappa statistic. *Family Medicine*, 37(5):360–363.

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269.

Wolfinger, R. and O'Connell, M. (1993). Generalized linear mixed models: A pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, 48:233–243.

Wood, J. M. (2007). Understanding and computing Cohen's Kappa: A tutorial. *WebPsychEmpiricist*.

Wu, J., Zhang, L., and Johnson, W. D. (2012). The permutation test as an ancillary procedure for comparing zero-inflated continuous distributions. *Open Journal of Statistics*, 2:274–280.

Yamato, J., Ohya, J., and Ishii, K. (1992). Recognizing human action in time-sequential images using Hidden Markov Model. *Proceedings 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 379–385.

Zeger, S. L. (1988). A regression model for time series of counts. *Biometrika*, 75:621–629.

Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects: A gibbs sampling approach. *Journal of the American Statistical Association*, 86(413):79–86.

Zeileis, A., Kleiber, C., and Jackman, S. (2008). Regression models for count data in R. *Journal of Statistical Software*, 7(2):1–38.

Zhang, L., Wu, J., and Johnson, W. D. (2010). Empirical study of six tests for equality of populations with zero-inflated continuous distribution. *Communications in Statistics - Simulation and Computation*, 39:1196–1211.

Zucchini, W. and MacDonald, I. L. (2009). *Hidden Markov Models for Time Series An Introduction Using R*. Chapman & Hall/CRC Press, Boca Raton, FL, 1 edition.

APPENDIX

A R Code for Zero-inflated Beta Hidden Markov Model

```
##############      R code for two state BEZI-HMM     ##############
library(gamlss) #BEZI density function is given in this library

######## Sample a HMM from a hidden markov chain
## in:  pai = initial distribution as vector of size 2
##      A = transition matrix of size 2
##      p1, mu1, phi1, p2, mu2, phi2 = parameters for BEZIs
##      n = positive integer, length of the chain
## out: (x,y) = sample trajectory of size n of a HMM defined by
##      (pi,A,p1,mu1,phi1,p2,mu2,phi2):
##      x = sample trajectory of size n of a Markov Chain with
##      initial distribution pi and transition matrix A
##      y = observations such that the conditionnal distribution of
##      y[k] given x[k] is BEZI(x[k], :)
BEZIHMMsample <- function(pai, A, theta, n){
  p1<-theta[1];
  mu1<-theta[2];
  phi1<-theta[3];
  p2<-theta[4];
  mu2<-theta[5];
  phi2<-theta[6];
  cA <- t(apply(A, 1, cumsum));
  x <- array(0, n);
  y <- array(0, n);
  x[1] <- 1+sum(as.numeric(runif(1)>cumsum(pai)));
  y[1] <- ifelse(x[1]==1,rBEZI(1,mu=mu1,sigma=phi1,nu=p1),
                         rBEZI(1,mu=mu2,sigma=phi2,nu=p2));
  for (t in 2:n){
    x[t] <- 1+sum(as.numeric(runif(1)>cA[x[t-1],]));
    y[t] <- ifelse(x[t]==1,rBEZI(1,mu=mu1,sigma=phi1,nu=p1),
                           rBEZI(1,mu=mu2,sigma=phi2,nu=p2));
```

```
  }
  list(x=x,y=y);
}

#######                  Viterbi algorithm                 ########
## obtain HMM state decoding, given hmm parameters
## in:  pai = initial distribution as vector of size 2
##      A = transition matrix of size 2
##      p1, mu1, phi1, p2, mu2, phi2 = parameters for BEZIs
##      y = observation sequence
##      nstate = number of state
## out: delta = delta[i,t] is the largest probability of partial
##      observation sequence Y_{1:t} and state sequence S_{1:t} till
##      time t with S_t=i given theta log scale.
##      psi = psi[i,t] is the proceeding state (S_{t-1}) that maximize
##      delta[i,t]
##      S_star = viterbi path
Viterbi<- function(y, pai, A,  theta){
  p1<-theta[1];
  mu1<-theta[2];
  phi1<-theta[3];
  p2<-theta[4];
  mu2<-theta[5];
  phi2<-theta[6];
  n <- length(y);
  dims <- dim(A);
  #To save the largest prob at each t
  delta <- matrix(nrow = dims[1], ncol = n);
  # To save the largest argument
  psi <- matrix(nrow = dims[1], ncol = n);
  S_star<-rep(NA,n);
  #Initializtion
  delta[,1] <- log(pai*c(dBEZI(y[1],mu=mu1,sigma=phi1,nu=p1),
                         dBEZI(y[1],mu=mu2,sigma=phi2,nu=p2)));
  psi[,1]<-c(0,0)
  #Recursion
  for (t in 2:n){
    delta[,t]<- apply(t(delta[,t-1]+log(A)),1,max)+
```

```
     log(c(dBEZI(y[t],mu=mu1,sigma=phi1,nu=p1),
      dBEZI(y[t],mu=mu2,sigma=phi2,nu=p2)))
    psi[,t]<-apply(t(delta[,t-1]+log(A)),1,which.max)
  }
  #Termination
  L_star<-max(delta[,n]) #log L max
  S_star[n]<-which.max(delta[,n])
  #Backtracking
  for (t in (n-1):1)
  {
    S_star[t]=psi[S_star[t+1],t+1]
  }
  S_star;
}

####### Posterior decoding: obtain HMM state given hmm parameters
## in:  pai = initial distribution as vector of size 2
##      A = transition matrix of size 2
##      p1, mu1, phi1, p2, mu2, phi2 = parameters for BEZIs
## out: alpha = forward variable, the probability of partial
##      observation sequence
##      Y_{1:t} till time t and state i at time t given theta
##      beta = backward variable, the probability of partial
##      observation sequence
##      from t+1 to the end, given state i at time t and theta
##      gamma = the probability of being state i at time t, given
##      Y_{1:T} and theta
PosteriorDecoding<- function(y, pai, A,  theta){
  p1<-theta[1];
  mu1<-theta[2];
  phi1<-theta[3];
  p2<-theta[4];
  mu2<-theta[5];
  phi2<-theta[6];
  n <- length(y);
  dims <- dim(A);
  res <- list();
  alpha <- matrix(nrow = dims[1], ncol = n); #forward variable
```

```
  beta <- matrix(nrow = dims[1], ncol = n); #backward variable
  alpha[,1] <- pai*c(dBEZI(y[1],mu=mu1,sigma=phi1,nu=p1),
                     dBEZI(y[1],mu=mu2,sigma=phi2,nu=p2));
  beta[,n]<-c(1,1)
  for (t in 2:n){
   alpha[,t]<- (alpha[,t-1] %*% A) * c(dBEZI(y[t],mu=mu1,sigma=phi1,nu=p1),
                                       dBEZI(y[t],mu=mu2,sigma=phi2,nu=p2))
   beta[,n-t+1]<-(beta[,n-t+2]*c(dBEZI(y[n-t+2],mu=mu1,sigma=phi1,nu=p1),
  dBEZI(y[n-t+2],mu=mu2,sigma=phi2,nu=p2)))%*% t(A)
  }
 gamma<-matrix(nrow = dims[1], ncol = n); # prob of being in i at time t
 gamma<-alpha*beta/apply( alpha*beta,2,sum)
 Posterior<-apply(gamma,2, which.max) # The max step.
 Posterior;
}

########             Forward backward function            ########
###### obtain HMM alpha, beta, gamma, xi, given hmm parameters
## in:  pai = initial distribution as vector of size 2
##      A = transition matrix of size 2
##      p1, mu1, phi1, p2, mu2, phi2 = parameters for BEZIs
## out: alpha = forward variable, the probability of partial
##      observation sequence Y_{1:t} till time t and state i at
##      time t given theta
##      beta = backward variable, the probability of partial
##      observation sequence from t+1 to the end, given state i
##      at time t and theta
##      gamma = the probability of being state i at time t,
##      given Y_{1:T} and theta
##      xi = the probability of being in state i at time t and state j
##      at time t+1, for t=1,...T-1
## Notes: Rescaling part follows suggestion in Rabiner 1989.
##         -->avoid underflow
forwardbackward <- function(y, pai, A,  theta){
  p1<-theta[1];
  mu1<-theta[2];
  phi1<-theta[3];
  p2<-theta[4];
```

```
mu2<-theta[5];
phi2<-theta[6];
n <- length(y);
dims <- dim(A);
res <- list();
res$alpha <- matrix(nrow = dims[1], ncol = n); #forward variable
c<-rep(NA,n) #scaling factor
a <- pai*c(dBEZI(y[1],mu=mu1,sigma=phi1,nu=p1),
           dBEZI(y[1],mu=mu2,sigma=phi2,nu=p2));
c[1]<-sum(a)
res$alpha[,1] <- a/c[1];
for (t in 2:n){
 a<-(res$alpha[,t-1] %*% A) * c(dBEZI(y[t],mu=mu1,sigma=phi1,nu=p1),
 dBEZI(y[t],mu=mu2,sigma=phi2,nu=p2))
 c[t]<-sum(a)
 res$alpha[,t]<-  a/c[t];
}
res$beta <- matrix(nrow = dims[1], ncol = n); #backward variable
res$beta[,n]<-c(1,1)/c[n]
for (t in seq(n-1, 1, -1)){
  res$beta[,t] = A %*% (c(dBEZI(y[t+1],mu=mu1,sigma=phi1,nu=p1),
  dBEZI(y[t+1],mu=mu2,sigma=phi2,nu=p2))*res$beta[, t+1])/ c[t];
}
# prob of being in i at time t
res$gamma<-matrix(nrow = dims[1], ncol = n);
for (k in 1:n)
{res$gamma[,k]<-(res$alpha*res$beta)[,k]/
                        apply( res$alpha*res$beta,2,sum)[k]  }
#apply( res$alpha*res$beta,2,sum) is the marginal likelihood
# prob of being in i at time t and j at time t+1
res$xi<-matrix(nrow = dims[1]*dims[2], ncol = n-1);
tempxi<-matrix(nrow = dims[1]*dims[2], ncol = n-1)
for (l in 1:(n-1))
{
  tempxi[,l]=c(t(res$alpha[,l]%*%t(res$beta[,(l+1)]))*
  c(dBEZI(y[l+1],mu=mu1,sigma=phi1,nu=p1),
  dBEZI(y[l+1],mu=mu2,sigma=phi2,nu=p2)))*c(t(A))
}
```

```
  # The rows are xi_{t}(1,1), xi_{t}(1,2), xi_{t}(2,1), xi_{t}(2,2)
  res$xi<-tempxi/apply(tempxi,2,sum)
  res;
}
# forwardbackward(y, pai, A, theta)

##  log-likelihood function for single observation
##  in:  BEZI parameters and observation value
##  out: loglikelihood value
log.f<-function(theta,obs)
{
  p<-theta[1]
  mu<-theta[2]
  phi<-theta[3]
  ind<-ifelse(obs==0,1,0)
  logf<-ind*log(p)+(1-ind)*log(1-p)+(1-ind)*(lfactorial(phi-1)-
  lfactorial(mu*phi-1)- lfactorial((1-mu)*phi-1))
   +(1-ind)*(mu*phi-1)*ifelse(log(obs)==-Inf,0,log(obs))+
  (1-ind)*log(1-obs)*((1-mu)*phi-1)
  return(logf)
}

######### likelihood function section for the BEZI part
##  in: gamma = gamma variables from forwardbackward
##      eta = unknow BEZI parameters
##  out: negative likelihood function of the BEZI part,
##       given the gamma's estimated at current stage.
##      ->use negative because the optimization minimze objective
##       function.
neg.Q.BEZI<-function(eta, gamma, y){
  p1<-eta[1]
  mu1<-eta[2]
  phi1<-eta[3]
  p2<-eta[4]
  mu2<-eta[5]
  phi2<-eta[6]
  Q<-sum(gamma[1,]*log.f(c(p1,mu1,phi1),y),gamma[2,]*
  log.f(c(p2,mu2,phi2),y))
```

```
  return(-Q)
}

#####        Baum-Welch algorithm for two-state BEZI-HMM
##   Need to run function forwardbackward, log.f, and neg.Q.BEZI first
## in:  pai = initial distribution as vector of size 2 ->initial value
##      A = transition matrix of size 2 ->inital value
##      p1, mu1, phi1, p2, mu2, phi2 = parameters for BEZIs ->inital value
##      alpha = forward variable, the probability of partial observation
##      sequence Y_{1:t} till time t and state i at time t given theta
##      beta = backward variable, the probability of partial observation
##      sequence from t+1 to the end, given state i at time t and theta
##      gamma = the probability of being state i at time t, given Y_{1:T}
##      and theta
##      xi = the probability of being in state i at time t and state j at
##      time t+1, for t=1,...T-1
## note: alpha,beta,gamma,xi are obtained from forwardbackward function.
## out:  pai.new, A.new, p1.new, mu1.new, phi1.new, p2.new,
##       mu2.new,phi2.new as HMM estimation.
BWupdate<-function(y, pai, A, theta, tol=1e-4, maxIt=200){
  n <- length(y);
  dims <- dim(A);
  pai.ij<-array(0,dims[1]*dims[2])
  it<-0;
  oldpai<-pai-tol;
  oldA<-A-tol;
  oldtheta<-theta-tol;
  while(((sum(abs((oldA-A))) +
  sum(abs(oldtheta-theta))+sum(abs(oldpai-pai)))> tol) &
   (it<maxIt)){
    oldpai<-pai
    oldA<-A
    oldtheta<-theta
    #E-step
    fbvar<-forwardbackward(y, oldpai, oldA,  oldtheta)
    #fbvar are initial fitting for forwardbackward variables
    pai<-fbvar$gamma[,1] #<-Updated initial prob
    pai.ij<-apply(fbvar$xi,1,sum)/c(apply(fbvar$gamma[,-n],1,sum)[1],
```

```
    apply(fbvar$gamma[,-n],1,sum)[1],apply(fbvar$gamma[,-n],1,sum)[2],
    apply(fbvar$gamma[,-n],1,sum)[2])
    A<-matrix(pai.ij,nrow=2,byrow=TRUE) #<-updated transition prob matrix
    #M-step
    omega0<-optim(theta, neg.Q.BEZI, y=y, gamma=fbvar$gamma,
    method="L-BFGS-B", control=list(maxit=500,pgtol=1e-2),
    lower=c(1e-4,1e-4,1e-4,1e-4,1e-4,1e-4),
    upper=c(1-1e-4,1-1e-4,Inf,1-1e-4,1-1e-4,Inf))
    # lower=c(1e-4,1e-4,1e-4,1e-4,1e-4,1e-4),
    # upper=c(1-1e-4,1-1e-4,100000,1-1e-4,1-1e-4,100000))
    theta<-omega0$par #<-updated BEZI parameters
    it<-it+1;
  }
  est <- list();
  est$pai<-pai;
  est$A<-A;
  est$theta<-theta;
  est$it<-it;
  est$Q<--neg.Q.BEZI(theta,fbvar$gamma,y)
  est;
}

############    Linearly constrained optimzation problem   ############
##### Verify the negative definited Hessian matrix of l^{\ast}  #####
objfun<-function(x)
{
  a<-x[1]
  b<-x[2]
  trigamma(a)*trigamma(b)-trigamma(a+b)*(trigamma(a)+trigamma(b))
}

grfun<-function(x)
{
  a<-x[1]
  b<-x[2]
  c(psigamma(a,2)*trigamma(b)-psigamma(a+b,2)*(trigamma(a)+trigamma(b))
  -trigamma(a+b)*psigamma(a,2) ,
  psigamma(b,2)*trigamma(a)-psigamma(a+b,2)*(trigamma(a)+trigamma(b))
```

```
  -trigamma(a+b)*psigamma(b,2))
}

constrOptim(c(0.5,0.5),objfun,grfun,ui=rbind(c(1,0),c(0,1)),ci=c(0,0),
outer.iterations = 1000, outer.eps = 1e-09)

# $par
# [1] 16.44413 16.44413
# $value
# [1] 5.972717e-05  <- postive
# $counts
# function gradient
# 63       42
# $convergence
# [1] 0
# $message
# NULL
# $outer.iterations
# [1] 8
# $barrier.value
# [1] -0.005919779


##################       BEZI-HMM toy example code      #############
#initial prob
pai <- c(0.4, 0.6);
#transition matrix
A <- matrix(c(0.8, 0.2, 0.3, 0.7), byrow=TRUE, nrow=2);
#parameters for the two BEZI population
p1<-0.1
mu1<-0.3
phi1<-1
p2<-0.2
mu2<-0.5
phi2<-1.5

theta<-c(p1,mu1,phi1,p2,mu2,phi2)
```

```
#BEZIsample
real<-list()
real$pai<-pai
real$A<-A
real$theta<-theta
real$eta<-eta
real

##Look at the pdf plot
x<-seq(0,0.999,0.001)
fx1<-dBEZI(x,mu=mu1,sigma=phi1,nu=p1)
fx2<-dBEZI(x,mu=mu2,sigma=phi2,nu=p2)
plot(x[-1],fx1[-1],type='l',lty=1,xlab='x', ylab='f(x)',ylim=c(0,13),
main='Probability Density Function')
lines(x[-1],fx2[-1],type='l',col=2,lty=2)
points(x[1],fx1[1])
points(x[1],fx2[1],col=2,pch=2)
legend(x=0.6,y=13.5,bty='n',c('BEZI 1','BEZI 2'),col=c(1,2), lty=c(1,2))
mtext(text=substitute(paste(alpha,"=",k, ", ", mu,"=",m,", ", phi,"=",p ),
list(k=p1, m=mu1, p=phi1)),side=3,line=-5.8, outer=TRUE)
mtext(text=substitute(paste(alpha,"=",k, ", ", mu,"=",m,", ", phi,"=",p ),
list(k=p2, m=mu2, p=phi2)),side=3,line=-6.8, outer=TRUE)

#number of observations
n=10
#sample
BEZIsample<-BEZIHMMsample(pai,A,theta,n)
#y is observation chain, x is the hidden state chain
y<-BEZIsample$y
x<-BEZIsample$x

########         Viterbi algorithm toy example           ############
Vresult<-Viterbi(y,pai,A,theta)
count<-ifelse(x==Vresult,1,0)
sum(count)/length(count)

########         Posterior decoding toy example          ############
PD<-PosteriorDecoding(y,pai,A,theta)
```

```
########            BW algorithm toy example              ############

#Initials
thetait<-c(0.2,0.5,2,0.2,0.5,3)
etait<-c(0.2,0.25,0.25,0.5,2,3)

#Use the mu and phi parameterization for the Beta part.
fit<-BWupdate(y, pai, A, theta, maxIt=1000)
fit
#bias:
cbind(c('a11','a21','a12','a22','p1','mu1','phi1','p2','mu2','phi2'),
c(as.vector(fit$A-A,),as.vector(fit$theta-theta)))

fit.it<-BWupdate(y, pai, A, thetait, maxIt=1000)
fit.it
#bias:
cbind(c('a11','a21','a12','a22','p1','mu1','phi1','p2','mu2','phi2'),
c(as.vector(fit.it$A-A,),as.vector(fit.it$theta-theta)))
# to initial
cbind(c('p1','mu1','phi1','p2','mu2','phi2'),
as.vector(fit.it$theta-thetait))

#######  Grid search for the EM algorithm, real data application ######
## Search grids are defined as follows:
## pi_1=c(0.1,0.2,0.3,0.4,0.5)
## a_11=c(0.05,0.25,0.45,0.65,0.85)
## a_22=c(0.05,0.25,0.45,0.65,0.85)
## p_1=p_2=p0=c(0.2,0.4,0.6,0.8)
## mu_1=mu_2=mu=c(0.2,0.4,0.6,0.8)
## phi_1=phi_2=phi=c(1,5,10,20,40)
##########################################################################
# total number of searching point
ncomb=10000 #=5*5*5*4*4*5
# search index
i=0
# save Q, iter, theta fitted, A fitted and pai fited
Q<-rep(NA,ncomb)
```

```
iter<-rep(NA,ncomb)
thetafit<-matrix(NA,nrow=ncomb,ncol=6)
Afit<-matrix(NA,nrow=ncomb,ncol=4)
paifit<-matrix(NA,nrow=ncomb,ncol=2)

# Notice that we need to read the data first:
# The code here assuming y is the observation sequence

# Start grid search
for (k00 in c(1,2,3,4,5))
{pai<-c(0.1*k00,1-0.1*k00)
for (k11 in c(0,1,2,3,4))
{ a11<-0.05+0.2*k11
  for (k22 in c(0,1,2,3,4))
  {a22<-0.05+0.2*k22
    for (k33 in c(1,2,3,4))
    {p0<-0.2*k33
      for ( k44 in c(1,2,3,4))
      {mu0<-0.2*k44
        for (k55 in c(1,5,10,20,40) )
        { phi0<-1*k55
          A<-matrix(c(a11, 1-a11, 1-a22, a22), byrow=TRUE, nrow=2);
          thetagrid<-c(p0,mu0,phi0,p0,mu0,phi0)
          i=i+1
          fit<-BWupdate(y, pai, A, thetagrid, maxIt=1000)
          Q[i]<-fit$Q
          iter[i]<-fit$it
          thetafit[i,]<-fit$theta
          Afit[i,]<-as.vector(fit$A)
          paifit[i,]<-fit$pai
        }
      }
    }
   }
  }
 }

# Round Q values for stationary points into the forth digit
```

```
round4Q<-round(Q,4)
# number of unique Q after rounding
lthuniq4Q<-length(unique(round4Q))
#find the top three values
top3<-c(max(unique(round4Q)),sort(unique(round4Q),
partial=lthuniq4Q-1)[lthuniq4Q-1],
sort(unique(round4Q),partial=lthuniq4Q-2)[lthuniq4Q-2])

#See how many Q's are equal to the top 3
match.seq<-match(round4Q,top3)

#give the counts
table(match.seq[is.na(match.seq)==F])

#give the top 3 index
Qmax
top3
which(round4Q==top3[1])
which(round4Q==top3[2])
which(round4Q==top3[3])

#outputs for the max Q
Afit[which(Q==max(Q)),]
thetafit[which(Q==max(Q)),]
iter[which(Q==max(Q))]
paifit[which(Q==max(Q)),]

#second best
Afit[which(round4Q==top3[2]),]
thetafit[which(round4Q==top3[2]),]
iter[which(round4Q==top3[2])]
paifit[which(round4Q==top3[2]),]

#third best
Afit[which(round4Q==top3[3]),]
thetafit[which(round4Q==top3[3]),]
iter[which(round4Q==top3[3])]
paifit[which(round4Q==top3[3]),]
```

```
#### Use the estimated result as input for decoding methods
# Give the fitted values (obtained in the EM algorithm output)
pai_fit<-c(0,1)
A_fit<-matrix(c(0.000, 1, 0.237, 0.763), byrow=TRUE, nrow=2);
theta_fit<-c(0.366,0.030,118.263,0.921,0.092,87.780)

# Decoding process
VA<-Viterbi(y,pai_fit,A_fit,theta_fit)
PD<-PosteriorDecoding(y,pai_fit,A_fit,theta_fit)
# print the decoded chains
VA
PD
# See how the decoded chains are consistant with the processes
trVA<-table(VA,relax)
tdVA<-table(VA,dup)
tfVA<-table(VA,front)
cbind(max(trVA[1,1]+trVA[2,2],trVA[2,1]+trVA[1,2]),
round(max(trVA[1,1]+trVA[2,2],trVA[2,1]+trVA[1,2])/sum(trVA),4))
cbind(max(tdVA[1,1]+tdVA[2,2],tdVA[2,1]+tdVA[1,2]),
round(max(tdVA[1,1]+tdVA[2,2],tdVA[2,1]+tdVA[1,2])/sum(tdVA),4))
cbind(max(tfVA[1,2]+tfVA[2,3],tfVA[1,3]+tfVA[2,2]),
round(max(tfVA[1,2]+tfVA[2,3],tfVA[1,3]+tfVA[2,2])/sum(tfVA),4))
trVA
tdVA
tfVA
trPD<-table(PD,relax)
tdPD<-table(PD,dup)
tfPD<-table(PD,front)
cbind(max(trPD[1,1]+trPD[2,2],trPD[2,1]+trPD[1,2]),
round(max(trPD[1,1]+trPD[2,2],trPD[2,1]+trPD[1,2])/sum(trPD),4))
cbind(max(tdPD[1,1]+tdPD[2,2],tdPD[2,1]+tdPD[1,2]),
round(max(tdPD[1,1]+tdPD[2,2],tdPD[2,1]+tdPD[1,2])/sum(tdPD),4))
cbind(max(tfPD[1,2]+tfPD[2,3],tfPD[1,3]+tfPD[2,2]),
round(max(tfPD[1,2]+tfPD[2,3],tfPD[1,3]+tfPD[2,2])/sum(tfPD),4))
trPD
tdPD
tfPD
```

B R Code for Zero-inflated Beta Generalized Linear Mixed Model with Autoregressive Random Effect

```
#####################################################################
## Function for simulation BEZI-AR1 data:
## Input:  n number of observations
##         p proportion of zero in BEZI distribution
##         beta0 intercept in linear predictor, logit link on mu
##         beta0 coefficient for x, logit link on mu
##         phi  phi parameter in BEZI distribution
##         sigma standard deviation of Normal white noise, logit
##         link on mu varphi  autoregression coefficient for AR1,
##         logit link on mu
## Output: x: simulated x predictor, the first half are 1, the
##         second half are 0, can be modified to more complex
##         case, for example, continuous variable
##         y: simulated BEZI response, notice that each y has its
##         unique mu because of the AR1 part on logit link.
##         mu_bar: average of mu parameters for all simulated
##         data, can give idea about the general density of y.
## Note: mu is the mean of Beta component
#####################################################################
BEZIAR1sim<-function(n,p,beta0,beta1,phi,sigma,varphi)
{
  sigmasq<-sigma^2
  x<-c(rep(1,n/2),rep(0,n/2)) #binary, n/2 are 1, rests are 0
  # simulate WN
  eta<-rnorm(n,0,sigmasq)
  # AR 1
  epsilon<-rep(0,n)
  epsilon[1]<-eta[1]/(sqrt(1-varphi^2))
  for (t in 2:n)
  {
    epsilon[t]=varphi*epsilon[t-1]+eta[t]
  }
  # mu_t
```

```
  mu_t<-rep(0,n)
  mu_t<-exp(beta0+beta1*x+epsilon)/(1+exp(beta0+beta1*x+epsilon))
  mu_bar<- mean(mu_t)
  # simulate y
  y<-rep(0,n)
  for (i in 1:n)
  {
    y[i]<-rBEZI(1,mu=mu_t[i],sigma=phi,nu=p)
  }
  list(y=y,x=x,mu_bar=mu_bar)
}

#####################################################################
## Call WinBUGs from R, use pacake R2WinBUGS
## Turn on all the packages and add options, give random seeds
library(gamlss)
library(R2WinBUGS)
library(xtable)
set.seed(37)

#####################################################################
## Give simulation configuration:
nsim<-50 #number of simulation
n<-100     # number of observation
beta0.true=-1 # intercept for mean function,  \beta_0
beta1.true=-3   # coefficient for covariate,  \beta_1
p.true=0.2       # zero proportion in BEZI, p
phi.true=1       # precision parameter in BEZI, \phi
sigma.true=1   # standard error of WN,  \sigma
varphi.true=0.5 # correlation coefficent for AR 1,
sigma2.true<-sigma.true^2
true<-c(beta0.true,beta1.true,p.true,phi.true,sigma2.true,
        varphi.true)

#####################################################################
# Simulate data
# y matrix[,i] is response in i th simulated dataset
ymatrix<-matrix(0,n,nsim)
# x matrix[,i] is predictor in i th simulated dataset
xmatrix<-matrix(0,n,nsim)
```

```
mu_bar<-rep(0,nsim)
for (i in 1:nsim)
{
  BEZIsample<-BEZIAR1sim(n,p.true,beta0.true,beta1.true,phi.true,
  sigma.true,varphi.true)
  #y is observation chain, x is the hidden state chain
  ymatrix[,i]<-BEZIsample$y
  xmatrix[,i]<-BEZIsample$x
  mu_bar[i]<-BEZIsample$mu_bar
}
mean_y<-apply(ymatrix,2,mean)
var_y<-apply(ymatrix,2,var)
median_y<-apply(ymatrix,2,median)
C<-matrix(0,nsim,66)
for (i in 1:nsim)
{
  y=ifelse(round(ymatrix[,i],6)==1,round(ymatrix[,i],6)-0.000001,
  round(ymatrix[,i],6))
  x=xmatrix[,i]
  data<-list("n","y","x")

  inits1<-list(beta0=beta0.true, beta1=beta1.true,p=p.true,phi=phi.true,
  sigma=sigma.true,varphi=varphi.true, eta0=0,
  eta=as.vector(arima.sim(list(order=c(1,0,0),ar=varphi.true),n=n)))
  #can define other intials as in inits1 above
  inits2<-initial.1(y,x)
  inits3<-initial.2(y,x)
  inits4<-initial.3(y,x)
  inits<-list(inits1,inits2,inits3, inits4)
  parameters<-c("beta0","beta1","p","phi","sigmasquare","varphi")
  sim <- bugs(data, inits, model.file = "Z://WB sim//sp2prior7_Jan14.odc",
              parameters, n.chains = 4, n.iter =100000, n.burnin =50000,
              n.thin = 150, bugs.directory = "Z://winbugs14//WinBUGS14")
  # A save sim result, in column order: mean sd 2.5% 25% 50% 75% 97.5%
  #                    in row order: beta0,beta1,p,phi,sigmasquare,varphi
  A<-matrix(as.numeric(sim$summary[1:6,]),nrow=6)
  #whether 95% posterior interval covers true
  cover<-ifelse(true>=A[,3]&true<=A[,7],1,0)
  # whether 95% posteror interval covers 0: 1 means cover, so not significant.
  sign<-ifelse(0>=A[,3]&0<=A[,7],1,0)
```

```
  # B save sim result, in column order: mean sd 2.5% 25% 50% 75% 97.5%,
  #                                     Cover true, cover 0
  #                     in row order: beta0,beta1,p,phi,sigmasquare,varphi
  B<-as.matrix(cbind(A,cover,sign))
  # Save result for one sim in a vector.
  C[i,]<-as.numeric(c(B[1,],B[2,],B[3,],B[4,],B[5,],B[6,]))
}
#Save the result
Final<-cbind(mu_bar,mean_y,var_y,median_y,C)
write.csv(Final,file="Z://WB sim//Final4.0.csv")
#Print the result
result<-apply(Final,2,mean)
basic.sec<-result[1:4]
beta0.sec<-result[5:15]
beta1.sec<-result[16:26]
p.sec<-result[27:37]
phi.sec<-result[38:48]
sigmasquare.sec<-result[49:59]
varphi.sec<-result[60:70]
basic.sec
beta0.sec
beta1.sec
p.sec
phi.sec
sigmasquare.sec
varphi.sec
```

C WinBUGS Code for Zero-inflated Beta Generalized Linear Mixed Model with Autoregressive Random Effect

```
model
{
  #Define the AR1 part on mu
```

```
# White Noise
m[1]<-varphi*eta0
eta[1]~dnorm(m[1],tau)
for (t in 2:n)
{
  m[t]<-varphi*eta[t-1]
  eta[t]~dnorm(m[t],tau)
}
#likelihood part
for ( t in 1:n)
{
  #define 0,1 indicator
  z[t]<-step(y[t]-0.00000001)  #if y[t]=0 then z[t]=0, if y[t]>0, then z[t]=1
  #link function part
  #SINGLE PREDICTOR
  logit(mu[t])<-beta0+beta1*x[t]+eta[t]  #  g(mu)=xbeta+eta
  #TWO PREDICTORS
  #logit(mu[t])<-beta0+beta1*x1[t]+beta2*x2[t]+eta[t]  #g(mu)=xbeta+eta
  #THREE PREDICTORS
  #logit(mu[t])<-beta0+beta1*x1[t]+beta2*x2[t]+beta3*x3[t]+eta[t]
  #g(mu)=xbeta+eta
  #ll[t] is log likelihood
  ll[t]<-log(p*equals(y[t],0)+z[t]*(1-p)*betapdf[t])
  # when y[t]=0 log(0)=-infty, log(1)=0
  betapdf[t]<-exp( loggam(phi)-loggam(mu[t]*phi)-loggam((1-mu[t])*phi)
  +(mu[t]*phi-1)*log(y[t])+((1-mu[t])*phi-1)*log(1-y[t]) )
  # This is the zero trick part. Specify a new sampling distribution.
  zeros[t]<-0
  #need to have a large number to ensure m>0 as poisson mean
  muPoisson[t]<- -ll[t]+100000
  zeros[t]~dpois(muPoisson[t])
  inv.like[t]<-1/exp(ll[t])
}
LL<-2*sum(ll[]) ##Deviance, may use as model selection, as AIC
#nuisance parameters
sigmasquare<-sigma*sigma
tau<-1/sigmasquare
tau0<-(1-varphi*varphi)*tau
#priors
beta0~dnorm(0,1.0E-4)
```

```
  beta1~dnorm(0,1.0E-4)
  #ADD beta 2 and 3 when there are multiple predictors
  #beta2~dnorm(0,1.0E-4)
  #beta3~dnorm(0,1.0E-4)
  p~dunif(0,1)
  phi~dgamma(0.1,0.1)
  sigma~dunif(0,10)
  varphi~dunif(-1,1)
  eta0~dnorm(0,tau0)
}
```