PH.D. THESIS

Logistic Mixtures of Generalized Linear

Model Time Series

by: *Neal Overton Jeffries*
Advisor: *Benjamin Kedem*

Ph.D. 98-3

ISR develops, applies and teaches advanced methodologies of design and analysis to solve complex, hierarchical, heterogeneous and dynamic problems of engineering technology and systems for industry and government.

ISR is a permanent institute of the University of Maryland, within the Glenn L. Martin Institute of Technology/A. James Clark School of Engineering. It is a National Science Foundation Engineering Research Center.

Web site http://www.isr.umd.edu

Abstract

Title of Dissertation: LOGISTIC MIXTURES OF GENERALIZED LINEAR MODEL TIMES SERIES

Neal Overton Jeffries, Doctor of Philosophy,

Dissertation directed by: Professor Benjamin Kedem
Mathematics Department
Statistics Program

In this dissertation we propose a class of time series models for mixture data. We call these logistic mixtures. In such models the mixture's component densities have a generalized linear model (GLM) form. The regime probabilities are allowed to change over time and are modeled with a logistic regression structure. The regressors of both the component GLM distributions and the logistic probabilities may include covariates as well as past values of the process. We develop an EM algorithm for estimation, give conditions for consistency and asymptotic normality, examine the model through simulations, and apply it to rain rate data. Finally, we consider a likelihood ratio based test for determining if the data arise from a logistic mixture versus the null hypothesis of the data coming from a single distribution (i.e. no mixture). Because the mixture probabilities are not constant we are able to develop a test that avoids some of the problems associated with likelihood ratio tests of mixtures.

LOGISTIC MIXTURES OF GENERALIZED LINEAR MODEL TIMES SERIES

by

Neal Overton Jeffries

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland at College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

Advisory Committee:

Professor Benjamin Kedem, Chairman
Professor Eric Slud
Professor Paul Smith
Assistant Professor Jing Qin
Professor Prakash Narayan

Dedication

To my parents and Amy.

Acknowledgements

My advisor, Professor Ben Kedem, has been a friend and mentor for several years and provided invaluable knowledge and support during this process. On many occasions, Professor Eric Slud offered timely suggestions that allowed me to surmount many problems. Ruth Pfeiffer acted as a valuable sounding board and suggested we apply these models toward rain rate data. I have been blessed by the patience and support of Dr James Dambrosia at the National Institutes of Health and Dr. Johnetta Davis in the Office of Graduate Minority Education. My parents and Amy Kincaid have provided unwavering love and support throughout my studies. This work would not have been possible without these people. Thank you.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In this dissertation we introduce a broad class of time series models that are applicable to data arising from mixtures of parametric distributions. The class of models we propose are formed by combining two time series following generalized linear model (GLM) distributions and modeling the probability of which distribution is applicable through a logistic regression structure. We call these logistic mixture (LM) models. These models are motivated by the realization that time series data may have parameters that are themselves changing over time. This is a contrast to the situation expressed by a simple ARMA(k, l) process:

$$Y_t - \xi_1 Y_{t-1} - ... - \xi_k Y_{t-k} = \epsilon_t + \beta_1 \epsilon_{t-1} + ... + \beta_l \epsilon_{t-l}$$

where ϵ_t are i.i.d. white noise and $\xi_1, ..., \xi_k, \beta_1, ..., \beta_l$ are unknown constants such that the roots of $1 - \xi_1 Z - ... - \xi_k Z^k$ and $1 + \beta_1 Z + ... + \beta_l Z^l$ lie outside the unit circle in the complex plane. While ARMA models have been popular and successful for describing several observed time series, they are not flexible enough to account for situations when the investigator believes the parameters are not constant.

Mixture models were developed as a way of allowing data to arise from a combination of two or more distinct data generation processes. See McLachlan and Basford (1988), Titterington, et. al. (1985), or Everitt and Hand (1981) for good introductions to mixtures, primarily in the context of i.i.d. random variables. As a basis for later comparison to our model and to introduce notation we present a simple parametric mixture model: let

- Y_t denote i.i.d. observed univariate data, $t \in \{1, 2, ..., T\}$,

- $f(y_t; \alpha), \alpha \in A$ denote a class of probability densities with respect to a common sigma-finite measure where A is a subset of $\mathbb{R}^q$ for some $q \in \mathbb{N}$,

- I_t is an i.i.d <u>unobserved</u> state variable that determines the conditional distribution of Y_t. By this we mean $I_t \in \{1, 2, ..., r\}$ for some r and

$$Y_t \text{ is distributed as } \begin{cases} f(y_t; \alpha_1) \text{ if } I_t = 1, \\ f(y_t; \alpha_2) \text{ if } I_t = 2, \\ \vdots \\ f(y_t; \alpha_r) \text{ if } I_t = r \end{cases}$$

  with $\alpha_i \neq \alpha_j$ if $i \neq j$. A standard mixture model would have $p_1 = \mathbb{P}[I_t = 1] > 0, p_2 = \mathbb{P}[I_t = 2] > 0, ..., p_r = \mathbb{P}[I_t = r] > 0$, and $\sum_1^r p_i = 1$ so the density for Y_t is given by $g(y_t; \alpha_1, \ldots, \alpha_r, p_1, \ldots, p_r) = \sum_1^p f(y_t; \alpha_i) p_i$.

Throughout this work we use $f(\cdot)$ to denote the conditional (also called component) densities of a mixture density and $g(\cdot)$ will denote the mixture density composed of the $f(\cdot)$s.

This model has the important property that the parameters governing the data generation change among the set $\{\alpha_1, \alpha_2, ..., \alpha_r\}$ according to the random

values of the I_t's. One can estimate the parameters in the model by performing maximum likelihood estimation. If the I_t's were observable we would write the joint likelihood (joint in Y_t and I_t) as

$$\prod_{t=1}^{T}\prod_{j=1}^{r}\left(f(y_t;\alpha_i)\mathbb{P}\left[I_t=j\right]\right)^{\chi_{[I_t=j]}}$$

where $\chi_{[S]}$ is an indicator function of the event S. As we do not observe the I_t we write the marginal likelihood of the Y_t's as

$$\prod_{t=1}^{T}\sum_{j=1}^{r} f(y_t;\alpha_j)p_j.$$

and perform maximum likelihood to estimate $\alpha_1, p_1, \alpha_2, p_2, ..., \alpha_r, p_r$ using some maximization procedure and applying the usual asymptotic theory for testing purposes.

The logistic mixture models we consider in this dissertation are characterized by the following modifications to the standard mixture model:

- We assume $r = 2$, i.e. there are only two states which we label as state (or regime) 1 and state 0. We do not anticipate there being much difficulty in extending our analysis to $r > 2$.

- Let W_t denote a vector of observable auxiliary, or exogenous, information.

- We write $f(y_t \mid W_t, Y_{t-1}, W_{t-1}, ..., Y_{t-p}, W_{t-p}; \alpha_i)$ instead of $f(y_t; \alpha_i)$. We now interpret $f(y_t \mid W_t, Y_{t-1}, W_{t-1}, ..., Y_{t-p}, W_{t-p}; \alpha_i)$ to be the conditional distribution of Y_t given $W_t, Y_{t-1}, W_{t-1}, ..., Y_{t-p}, W_{t-p}$ and $I_t = i$ for $i \in \{0, 1\}$.

- $f(y_t \mid W_t, Y_{t-1}, W_{t-1}, ..., Y_{t-p}, W_{t-p}; \alpha_i)$ has the form of a canonical GLM with parameters given by known functions of α_i and $W_t, Y_{t-1}, W_{t-1}, ..., Y_{t-p}$, W_{t-p}.

- The state probabilities are not constant but given by a logistic regression model:

$$\mathbb{P}\left[I_t = 1 \mid W_t, Y_{t-1}, W_{t-1}, ..., Y_{t-p}, W_{t-p}\right] = \frac{\exp(Z_t'\gamma)}{1 + \exp(Z_t'\gamma)} \tag{1.1}$$

  where Z_t is a vector made up of known functions of $W_t, Y_{t-1}, W_{t-1}, ..., Y_{t-p}$, W_{t-p} and γ is an unknown set of logistic regression coefficients.

With these modifications we have explicitly introduced temporal dependence in making the conditional distributions a function of the past. From this model data is generated in the following manner:

- Given γ and Z_t, compute $\mathbb{P}\left[I_t = 1 \mid W_t, Y_{t-1}, W_{t-1}, ..., Y_{t-p}, W_{t-p}\right]$.

- Generate a Bernoulli random variable, I_t, with mean given by the probability above.

- If $I_t = 1$ generate Y_t from $f(y_t \mid W_t, Y_{t-1}, W_{t-1}, ..., Y_{t-p}, W_{t-p}; \alpha_1)$; if $I_t = 0$ generate Y_t from $f(y_t \mid W_t, Y_{t-1}, W_{t-1}, ..., Y_{t-p}, W_{t-p}; \alpha_0)$.

DeSarbo and Wedel (1995) provided a general EM algorithm approach to estimating parameters in finite mixtures of independent GLM data with covariates and constant transition probabilities. The logistic modeling we add allows us to incorporate non-constant probabilities and model the important effects covariates may have on the regime probability. This is an important and useful extension of existing mixture models. Adding the flexibility of non-constant probabilities should improve the model's discriminatory power – thus allowing better estimation of the component distributions' parameters. Furthermore, the parameters of the logistic regression in the probability modeling may themselves be of interest. While Kuk and Chen (1992), and Larson and Dinse (1985) have

employed logistic regression mixtures in the context of mixtures of hazard rates we have not seen this method employed in other parametric situations. Furthermore, we believe that neither asymptotic results nor testing for whether a mixture is present has been addressed in the context of logistic mixtures.

1.1 Competing Models

In addition to the standard mixture model with fixed regime probabilities there are two other broad types of time series models we will discuss in this dissertation – mostly as a basis for comparison. Both of these models are similar to the mixture (or switching) models in that they posit the data arises from a combination of data generating processes.

The first type we discuss are threshold models developed by Tong (1983, 1990). There are many generalizations to this simple model, but the basic idea can be given by this two state self exciting threshold autoregressive (SETAR) model:

$$Y_t = \begin{cases} \xi_{01} + \xi_{11} Y_{t-1} + ... + \xi_{p1} Y_{t-p} + \sigma_1 \cdot \epsilon_t \text{ if } Y_{t-d} > \tau \\ \xi_{00} + \xi_{10} Y_{t-1} + ... + \xi_{p0} Y_{t-p} + \sigma_0 \cdot \epsilon_t \text{ if } Y_{t-d} \leq \tau \end{cases}$$

where d (delay parameter), p (lag length), and τ (threshold) are unknown. For fixed τ, d, p the $\xi_{01}, \xi_{11}, ..., \xi_{p1}, \xi_{00}, \xi_{10}, ...,$ and ξ_{p0} are estimated by conditional

least squares, i.e.

$$\widehat{\xi}_{01}, \widehat{\xi}_{11}, ..., \widehat{\xi}_{p1}, \widehat{\xi}_{00}, \widehat{\xi}_{10}, ..., \text{ and } \widehat{\xi}_{p0} =$$

$$\text{Arg} \min_{\xi_{01}, \xi_{11}, ..., \xi_{p1}, \xi_{00}, \xi_{10}, ..., \xi_{p0}} CSS(p, d, \tau, \xi_{01}, \xi_{11}, ..., \xi_{p1}, \xi_{00}, \xi_{10}, ..., \xi_{p0}) \text{ where}$$

$$CSS(p, d, \tau, \xi_{01}, \xi_{11}, ..., \xi_{p1}, \xi_{00}, \xi_{10}, ..., \xi_{p0}) =$$

$$\sum_{t:\, Y_{t-d} \leq \tau} [Y_t - \xi_{00} - \xi_{10} Y_{t-1} - ... - \xi_{p0} Y_{t-p}]^2 +$$

$$\sum_{t:\, Y_{t-d} > \tau} [Y_t - \xi_{01} - \xi_{11} Y_{t-1} - ... - \xi_{p1} Y_{t-p}]^2 .$$

The choices of p, d, and τ may be informed by theory, or by experimenting until one finds $\widehat{p}, \widehat{d}$, and $\widehat{\tau}$ where

$$\widehat{p}, \widehat{d}, \widehat{\tau} = \text{ Arg } \min_{p,d,\tau} CSS(p, d, \tau, \widehat{\xi}_{01}, \widehat{\xi}_{11}, ..., \widehat{\xi}_{p1}, \widehat{\xi}_{00}, \widehat{\xi}_{10}, ..., \widehat{\xi}_{p0})$$

and it is understood the $\widehat{\xi}_{01}, \ldots, \widehat{\xi}_{p0}$ terms in the preceding equation are all dependent upon p, d, and r. In these threshold models the primary feature is that data generation process changes according to whether a variable (in this case Y_{t-d}) exceeds a particular threshold (τ) or not. These models may be generalized to include covariates with coefficients that change depending upon whether or not the threshold is exceeded – Tong calls these TARSO models.

A second class of models we will discuss are hidden Markov model regressions (HMMRs). These were introduced by Goldfeld and Quandt (1973) and Lindgren (1978) and further developed and popularized among econometricians by Hamilton (1990, 1996). These models have been used to describe many economic time series including GNP (Hamilton (1990)), business cycles (Diebold et. al (1994) and Filardo (1994)), stock price volatility (Fridman (1994)), and exchange rate fluctuations (Engel and Hamilton (1990)). These models are similar to the mixture (or switching) models used by economists (Quandt (1958)

and Kiefer (1978)) in that they assume the existence of an unobserved state indicator, I_t. But in the case of hidden Markov models the I_ts are realizations of a Markov chain. We introduce a relatively simple example of such a model involving two states and with normal conditional distributions. As before let

- Y_t be our observed outcome, W_t denote a vector of observable auxiliary information, and I_t be an unobserved indicator of which regime's parameters generate the data.

- Also as before we associate with states 0 and 1 two sets of parameters, α_1 and α_0, such that the conditional distribution of Y_t given $I_t = i$ and $W_t, Y_{t-1}, W_{t-1}, ..., Y_{t-p}, W_{t-p}$ is given by

$$f(y_t \mid W_t, Y_{t-1}, W_{t-1}, ..., Y_{t-p}, W_{t-p}; \alpha_i),$$

  where $I_t \in \{0, 1\}$.

- In this HMMR, I_t is an unobserved stationary 2 state Markov chain with transition matrix P containing elements $[P]_{i,j} = p_{ij} = \mathbb{P}[I_{t+1} = j \mid I_t = i]$. Furthermore, conditional on I_t, I_{t+1} is independent of all the Y_s for $s \in \{-p, -p+1, \ldots, T\}$.

In most econometric applications the conditional densities are taken to be normally distributed:

$$f(y_t \mid W_t, Y_{t-1}, W_{t-1}, ..., Y_{t-p}, W_{t-p}; \alpha_i) = \frac{1}{\sqrt{2\pi\,\phi_i}} \exp\left(\frac{-(y_t - X_{ti}'\beta_i)^2}{2\phi_i}\right)$$

where X_{ti} is a vector composed of functions of $W_t, Y_{t-1}, W_{t-1}, ..., Y_{t-p}, W_{t-p}$ and $\alpha_i = (\beta_i, \phi_i)$. These models generalize the basic mixture model by allowing the state indicator probabilities to correspond to a Markov transition matrix. Maximum likelihood estimation of the parameters (α_1, α_0, P) is usually performed

via the EM (Expectation-Maximization) algorithm. The addition of the Markov structure does complicate estimation as the additional step of a 'backward-forward' algorithm is necessary to employ the EM approach for HMMR models (see Hamilton (1994) or Fridman (1994)).

1.2 Partial Likelihood

Throughout this dissertation we use the terms likelihood, partial likelihood, and conditional likelihood somewhat loosely and at times interchangeably. The difference in these terms largely depends upon how we think about our auxiliary information, W_t. Before going further we define more precisely what we mean by these terms. The following description is drawn from Fokianos (1996) and is based in turn upon Cox (1975), Wong (1986), and Slud and Kedem (1994). Suppose $(Y_t, W_t), t \in \{-p, -p+1, .., -1, 0, 1, ..., T\}$ is a stochastic process and given some initial information set, $\{y_0, w_0, y_{-1}, w_{-1}, ..., y_{-p}, w_{-p}\}$ the joint distribution of our sample is written as

$$g(y_t, w_t, y_{t-1}, w_{t-1}, ..., y_1, w_1 \mid y_0, w_0, y_{-1}, w_{-1}, ..., y_{-p}, w_{-p}).$$

Then we may factor this is into the equivalent products

$$\prod_{t=1}^{T} g(y_t, w_t \mid y_{t-1} w_{t-1}, ..., y_{-p}, w_{-p}) = \prod_{t=1}^{T} g(y_t \mid d_t) \prod_{t=1}^{T} g(w_t \mid c_t) \tag{1.2}$$

where $d_t = (w_t, y_{t-1}, w_{t-1}, ..., y_{-p}, w_{-p})$ and $c_t = (y_{t-1}, w_{t-1}, ..., y_{-p}, w_{-p})$. (In the expressions above we have abused our notation by using $g(\cdot)$ as representing joint as well as conditional probabilities or densities but the meaning should be clear.) For Cox (1975) and Wong (1986) the $\prod g(y_t \mid d_t)$ term in (1.2) is the partial likelihood. Slud and Kedem (1994) provides a more formal definition which includes parameters for a conditional density. We will adopt their usage.

Definition 1.1. *Let $\mathcal{G}_t, t \in \{0, 1, ...\}$ be an increasing sequence of sigma fields, $\mathcal{G}_0 \subseteq \mathcal{G}_1 \subseteq \mathcal{G}_2...$, and let Z_t be a sequence of random variables on some common probability space such that Z_t is $\mathcal{G}_t$-measurable. Denote the density of Z_t given $\mathcal{G}_{t-1}$ by $f(z_t \mid \mathcal{G}_{t-1}; \alpha)$ where $\alpha \in \mathbb{R}^q$ represents a vector of parameters. The partial likelihood function relative to $\alpha, \mathcal{G}_{t-1}$, and the data $z_1, z_2, ..., z_T$ is given by*

$$\prod_{t=1}^{T} g(z_t \mid \mathcal{G}_{t-1}; \alpha). \tag{1.3}$$

Partial likelihood is somewhat different then the general definitions of full and conditional likelihood. Unlike full likelihood, partial likelihood does not require complete knowledge of the joint distribution of the covariates – i.e. we do not concern ourselves with the $\prod g(w_t \mid c_t)$ term in equation (1.2). Unlike conditional likelihood, complete covariate information need not be known throughout the period of observation (from time $t = 1$ through $t = T$). Partial likelihood considers only what is known to the observer up to the time of observation. Often the terms will be the same – we devote a large part of this work to analyzing a situation when there are no W_t terms, only Y_t's. In this case our notions of full and partial likelihood will coincide. The vector α that maximizes (1.3) for a given set of data is called the maximum partial likelihood estimator (MPLE). In the remainder of this study this is what we have in mind when we refer to maximum likelihood estimates.

1.3 Overview of Presentation

In Chapter 2 we introduce our model and detail an EM algorithm approach to finding maximum likelihood estimates (or more precisely, maximum partial like-

lihood estimates) for the parameters in the two component distributions as well as those parameters in the logistic regression that predict the regime probabilities. The time series dependence is modeled through a homogeneous, continuous-state, discrete-time Markov chain that allows us to express the sample likelihood (conditional on some initial set of observations) as a product of mixture densities.

Chapter 3 addresses the large sample properties of a correctly specified logistic mixture model. As might be expected we are able to demonstrate consistency and asymptotic normality under some very general assumptions. Asymptotic results for time dependent models have been developed by Billingsley (1961), Wong (1986), and Kaufmann (1987). Our approach is most similar to Wong's though we tailor it so we may easily demonstrate the conditions for the asymptotic results are met by a logistic mixture of Gaussian AR(1) processes. The primary difficulty in the chapter is to demonstrate that a process evolving from a logistic mixture is asymptotically stationary and ergodic. From this we can apply ergodic theorems to achieve the desired convergence. In Chapter 4 we then illustrate estimation with some simulations and also show our model may be superior to a standard threshold autoregression (SETAR) or a covariate threshold model (TARSO) in that the logistic mixture may be more robust to noise in the threshold variable. We close the chapter by applying the model to rain rate data.

In Chapter 5 we present the most interesting part of this work. Here we raise what has been a difficult question in analysis of mixtures: how to test whether the data are generated by a single parametric distribution or do they arise from a mixture. The application of chi-squared tests to twice the log-likelihood ratio is a popular though incorrect attempt to answer this question –

tionally (another criticism of Hansen's work). However the test is restricted to mixtures with common variance (in the case of Gaussian component densities) and it is not clear the method can incorporate covariates or be used with logistic probabilities (as in (1.1)).

As an alternative we propose a procedure that, like Gong and Mariano's test, yields an exact asymptotic distribution under the null hypothesis yet seems more broadly applicable (e.g. allowing covariates). One drawback for both our test and that of Gong and Mariano is that the class of alternatives is smaller than what seems natural or ideal. Both tests must exclude mixtures with constant regime probabilities (i.e. $\mathbb{P}[I_t = 1 \,|\,$ the past $] = p$, a constant independent of time) from the set of alternatives. Our test is more demanding computationally which can be a drawback when there are a large number of covariates in the logistic regression formulation. We develop the theory behind our test in Chapter 5. Chapter 6 examines out test's performance in simulations and the rain data example. We conclude this study with some thoughts about how this work may be extended.

Chapter 2

Description and Estimation of the Logistic Mixture Model

As discussed in the introductory chapter the basic logistic mixture model contains the following elements:

- Y_t denotes an observed univariate time series, $t \in \{-p, -p+1, 0, 1..T\}$,

- W_t a vector of exogenous random variables,

- Z_t, X_{t0}, X_{t1} sets of observable covariates composed of known functions of $W_t, Y_{t-1}, W_{t-1}, W_{t-2}, Y_{t-2}, ...$ i.e. each of the three vectors is $\in \mathcal{P}_{t-1} = \sigma(W_t, Y_{t-1}, W_{t-1}...Y_0, W_0, ...Y_{-p+1}, W_{-p+1})$, where $\sigma(\cdot)$ denotes the sigma algebra generated by the arguments. For notational simplicity we will assume each vector is $q \times 1$.

- Assume Y_t can obey two different regimes/models where Y_t is generated by the regime 1 distribution if $I_t = 1$ and Y_t is generated by the regime 0 distribution if $I_t = 0$ where

$$\mathbb{P}\left[I_t = 1 \mid \mathcal{P}_{t-1}; \gamma\right] = \exp(Z_t'\gamma)/(1 + \exp(Z_t'\gamma)) \tag{2.1}$$

and γ represents an unknown vector of regression parameters. In other words, I_t is the dependent variable in a logistic regression model with covariate vector Z_t. It is important to note that the I_t's are <u>not</u> observed. (Throughout this work, vectors such as Z_t will be assumed column vectors with transposes denoted by Z_t'.)

- Assume the density of Y_t given the indicator I_t and past values has a canonical GLM distribution with some regime specific covariates, i.e.

$$f(y_t|I_t = 1, \mathcal{P}_{t-1}; \beta_1, \phi_1) = \exp\left([y_t X_{t1}'\beta_1 - b(X_{t1}'\beta_1)]/\phi_1 + c_1(y_t, \phi_1)\right), \tag{2.2}$$

$$f(y_t|I_t = 0, \mathcal{P}_{t-1}; \beta_0, \phi_0) = \exp\left([y_t X_{t0}'\beta_0 - b(X_{t0}'\beta_0)]/\phi_0 + c_0(y_t, \phi_0)\right),$$

and $X_{t1}, X_{t0} \in \mathcal{P}_{t-1}$.

The functions $f(y_t \mid I_t = i, \mathcal{P}_{t-1}; \beta_i, \phi_i), i \in \{0, 1\}$ are considered densities with respect to some σ-finite measure, ν on the real line. As examples, we can obtain the normal and Poisson distributions with the following substitutions:

$$\text{Normal: } b(X_{ti}'\beta_i) = \frac{(X_{ti}'\beta_i)^2}{2}, \phi_i = \sigma_i^2, \text{ and } c(y_t, \phi_i) = -(1/2)\ln 2\pi\phi_i + \frac{-y_t^2}{2\phi_i}$$

$$\text{Poisson: } b(X_{ti}'\beta_i) = \exp(X_{ti}'\beta_i), \phi_i = 1, \text{ and } c(y_t, \phi_i) = -\ln y_t!.$$

The binomial and gamma distributions are other common families with GLM form when modeled with covariates.

We define $\psi = (\beta_1', \beta_0', \phi_0, \phi_1, \gamma')'$. Our goal is to estimate ψ from our incomplete knowledge of the process (i.e. $\{I_t\}$ is not observed). We use the EM algorithm popularized by Dempster, Laird, and Rubin (1977) – see also Wu (1983) or McLachlan (1997). In the next section we focus upon the role of ψ and consequently will often write $f(y_t \mid I_t = i, \mathcal{P}_{t-1}; \psi)$ instead of $f(y_t \mid I_t = i, \mathcal{P}_{t-1}; \beta_i, \phi_i)$

when we want to emphasize the dependence upon ψ. As β_i and ϕ_i are components of ψ there should be no confusion. Also, we will write

$$f(y_t \mid I_t = 1, \mathcal{P}_{t-1}; \beta_1, \phi_1) \text{ as } f(y_t \mid X_{t1}; \beta_1, \phi_1) \text{ or as } f(y_t \mid X_{t1}; \phi)$$

and similar expressions will be used for $f(y_t \mid I_t = 0, \mathcal{P}_{t-1}; \beta_0, \phi_0)$. However we write this conditional distribution it is to be understood that $f(y_t \mid \cdot)$ denotes one of the component densities of the mixture – which density is designated by a 1 or 0 subscript.

2.1 General EM Algorithm

In this section we briefly outline a general EM (Expectation-Maximization) algorithm for optimizing time series partial likelihoods of the type outlined above. It should be stressed that the EM algorithm is primarily just an optimization method that is particularly well-suited to data with missing (or unobserved) components – other optimizing methods could be used (e.g. Newton-Raphson). However, the EM does have the convenient property that the likelihood increases when evaluated at each iteration's new estimate. By this we mean if ψ^k is our EM estimate of ψ^* (the true parameter) after k iterations of the algorithm, then

$$\prod^{T} g(y_t \mid \mathcal{P}_{t-1}; \psi^{k+1}) \geq \prod^{T} g(y_t \mid \mathcal{P}_{t-1}; \psi^k). \tag{2.3}$$

This is clearly not the case for Newton-Raphson or other types of optimization routines. This monotone property of the EM becomes more important as the number of parameters we estimate increases and other methods have trouble converging or finding maxima of the likelihood.

To show how to implement the EM in our case we begin by considering different likelihood products. $\prod_{t=1}^{T} g(y_t \mid \mathcal{P}_{t-1}; \psi)$ is the partial likelihood expression in the previous chapter. $\prod_{t=1}^{T} g(y_t, i_t \mid \mathcal{P}_{t-1}; \psi)$ denotes what we think of as the joint (partial) likelihood of (Y_t, I_t). This corresponds to how we would think of the likelihood if the I_t's were observed. In our model of a logistic mixture

$$g(y_t, i_t \mid \mathcal{P}_{t-1}; \psi) = g(y_t, I_t = 1 \mid \mathcal{P}_{t-1}; \psi)^{i_t} g(y_t, I_t = 0 \mid \mathcal{P}_{t-1}; \psi)^{1-i_t} \tag{2.4}$$

$$= (f(y_t \mid I_t = 1, \mathcal{P}_{t-1}; \beta_1, \phi_1)\mathbb{P}\left[I_t = 1 \mid \mathcal{P}_{t-1}; \gamma\right])^{i_t} \times \tag{2.5}$$

$$(f(y_t \mid I_t = 0, \mathcal{P}_{t-1}; \beta_0, \phi_0)\mathbb{P}\left[I_t = 0 \mid \mathcal{P}_{t-1}; \gamma\right])^{1-i_t}. \tag{2.6}$$

The last product we consider is $\prod_{t=1}^{T} g(i_t \mid y_t, \mathcal{P}_{t-1}; \psi)$ – this is the product of conditional probabilities of the indicator given contemporaneous values of Y_t, W_t and the past history of Y_s, W_s for $s < t$. We may express this in terms of the other two products as

$$\prod_{t=1}^{T} g(i_t \mid y_t, \mathcal{P}_{t-1}; \psi) = \prod_{t=1}^{T} \frac{g(y_t, i_t \mid \mathcal{P}_{t-1}; \psi)}{g(y_t \mid \mathcal{P}_{t-1}; \psi)}. \tag{2.7}$$

In these likelihood products the parameter vector ψ is not assumed to be the true value, ψ^*.

Given ψ and ψ' (possibly identical) we now define

$$M\left(\psi, \underline{y}, \underline{w}, \psi'\right) = \sum_{t=1}^{T} \int \left(\log g(y_t, i_t \mid \mathcal{P}_{t-1}; \psi)\right) g(i_t \mid y_t, \mathcal{P}_{t-1}; \psi') di_t \tag{2.8}$$

We write this as an integral to emphasize that it is an expectation – the expectation of the EM algorithm. The $\underline{y}$ and $\underline{w}$ vectors correspond to our observed sample data $(y_{-p}, y_{-p+1}, ..., y_T)$ and $(w_{-p}, w_{-p+1}, ..., w_T)$. While the w_t's are not shown explicitly on the right-hand side of our definition in (2.8) they are implicitly included in the $\mathcal{P}_{t-1}$ terms. Now, because I_t takes on only two values we

may rewrite our sum of expectations as

$$M\left(\psi, \underline{y}, \underline{w}, \psi'\right) = \sum_{t=1}^{T} \left(\log g(y_t, I_t = 1 \mid y_t, \mathcal{P}_{t-1}; \psi) \cdot \mathbb{P}\left[I_t = 1 \mid \mathcal{P}_{t-1}; \psi'\right] + \right. \\ \left. \log g(y_t, I_t = 0 \mid y_t, \mathcal{P}_{t-1}; \psi) \cdot \mathbb{P}\left[I_t = 0 \mid \mathcal{P}_{t-1}; \psi'\right]\right). \tag{2.9}$$

We also define

$$\begin{aligned} H\left(\psi, \underline{y}, \underline{w}, \psi'\right) &= \sum_{t=1}^{T} \int \left(\log g(i_t \mid y_t, \mathcal{P}_{t-1}; \psi)\right) g(i_t \mid y_t, \mathcal{P}_{t-1}; \psi') di_t \\ &= \sum_{t=1}^{T} \left(\log g(I_t = 1 \mid y_t, \mathcal{P}_{t-1}; \psi) \cdot \mathbb{P}\left[I_t = 1 \mid y_t, \mathcal{P}_{t-1}; \psi'\right] + \right. \\ &\qquad \left. \log g(I_t = 0 \mid y_t, \mathcal{P}_{t-1}; \psi) \cdot \mathbb{P}\left[I_t = 0 \mid y_t, \mathcal{P}_{t-1}; \psi'\right]\right). \end{aligned}$$

With these definitions and the data, $(\underline{y}, \underline{w})$, the algorithm is characterized by the following two step process:

> The E-Step: Given ψ^k compute $M(\psi, \underline{y}, \underline{w}, \psi^k)$. We view this term as a function of ψ.
>
> The M-step: Define $\psi^{k+1} = \text{ Arg } \max_{\psi} M(\psi, \underline{y}, \underline{w}, \psi^k)$.

We are now in a position to prove our statement that the algorithm produces estimates that increase the likelihood in the sense described by equation (2.3). To do so we will first prove the following:

Lemma 2.1. *For any given ψ and ψ' we have*

$$M\left(\psi, \underline{y}, \underline{w}, \psi'\right) - H\left(\psi, \underline{y}, \underline{w}, \psi'\right) = \sum_{t=1}^{T} \log g\left(y_t \mid \mathcal{P}_{t-1}; \psi\right), \tag{2.10}$$

i.e. the difference equals the sample's partial log likelihood evaluated at parameter value ψ.

Proof.

$$M\left(\psi, \underline{y}, \underline{w}, \psi'\right) - H\left(\psi, \underline{y}, \underline{w}, \psi'\right) = \tag{2.11}$$

$$\sum_{t=1}^{T} \int \left[\log g(y_t, i_t \mid \mathcal{P}_{t-1}; \psi) - \log g(i_t \mid y_t, \mathcal{P}_{t-1}; \psi)\right] g(i_t \mid y_t, \mathcal{P}_{t-1}; \psi') di_t =$$

(2.12)

$$\sum_{t=1}^{T} \int \log g(y_t \mid \mathcal{P}_{t-1}; \psi) g(i_t \mid y_t, \mathcal{P}_{t-1}; \psi') di_t. \tag{2.13}$$

The equality in equations (2.12–2.13) follows from the equivalence of the integrands – see the relation in (2.7). But the function $\log g(y_t \mid \mathcal{P}_{t-1}; \psi)$ is measurable with respect to $\sigma\left(y_t, \mathcal{P}_{t-1}\right)$ and hence we may pass the function through the integral in (2.13) and write

$$\sum_{t=1}^{T} \int \log g(y_t \mid \mathcal{P}_{t-1}; \psi) g(i_t \mid y_t, \mathcal{P}_{t-1}; \psi') di_t = \tag{2.14}$$

$$\sum_{t=1}^{T} \log g(y_t \mid \mathcal{P}_{t-1}; \psi) \int g(i_t \mid y_t, \mathcal{P}_{t-1}; \psi') di_t = \tag{2.15}$$

$$\sum_{t=1}^{T} \log g(y_t \mid \mathcal{P}_{t-1}; \psi) = \text{ our expression for partial log-likelihood.} \tag{2.16}$$

□

With this lemma it is easy to demonstrate the increasing likelihood (or equivalently the increasing log-likelihood) associated with the EM algorithm iterates. Let ψ^1 be any starting point and consider the sequence ψ^k generated by following the two step algorithm. Then by our lemma we have

$$\sum_{t=1}^{T} \log g(y_t \mid \mathcal{P}_{t-1}; \psi^{k+1}) = M\left(\psi^{k+1}, \underline{y}, \underline{w}, \psi^k\right) - H\left(\psi^{k+1}, \underline{y}, \underline{w}, \psi^k\right) \tag{2.17}$$

$$\text{and } \sum_{t=1}^{T} \log g(y_t \mid \mathcal{P}_{t-1}; \psi^k) = M\left(\psi^k, \underline{y}, \underline{w}, \psi^k\right) - H\left(\psi^k, \underline{y}, \underline{w}, \psi^k\right). \tag{2.18}$$

If we subtract the two log-likelihoods we obtain

$$\sum_{t=1}^{T} \log g(y_t \mid \mathcal{P}_{t-1}; \psi^{k+1}) - \sum_{t=1}^{T} \log g(y_t \mid \mathcal{P}_{t-1}; \psi^{k}) = \tag{2.19}$$

$$\left[M\left(\psi^{k+1}, \underline{y}, \underline{w}, \psi^{k}\right) - M\left(\psi^{k}, \underline{y}, \underline{w}, \psi^{k}\right)\right] + \tag{2.20}$$

$$\left[H\left(\psi^{k}, \underline{y}, \underline{w}, \psi^{k}\right) - H\left(\psi^{k+1}, \underline{y}, \underline{w}, \psi^{k}\right)\right]. \tag{2.21}$$

By our definition of ψ^{k+1} in the two step algorithm we see that the difference in (2.20) is greater than or equal to zero. If we expand the terms in (2.21) we get

$$H\left(\psi^{k}, \underline{y}, \underline{w}, \psi^{k}\right) - H\left(\psi^{k+1}, \underline{y}, \underline{w}, \psi^{k}\right) = \tag{2.22}$$

$$\sum_{t=1}^{T} \int \left[\log g(i_t \mid y_t, \mathcal{P}_{t-1}; \psi^{k}) - \log g(i_t \mid y_t, \mathcal{P}_{t-1}; \psi^{k+1})\right] g(i_t \mid y_t, \mathcal{P}_{t-1}; \psi^{k}) di_t. \tag{2.23}$$

By the Kullback-Leibler information inequality we have that this term is also greater than or equal to zero. Combining these relations we obtain the desired result that

$$\sum_{t=1}^{T} \log g(y_t \mid \mathcal{P}_{t-1}; \psi^{k+1}) \geq \sum_{t=1}^{T} \log g(y_t \mid \mathcal{P}_{t-1}; \psi^{k}).$$

In practice, ψ^k usually has a finite limit (as k increases) that corresponds to a local maximum of the likelihood surface. There are unusual circumstances where $\lim_k \psi^k$ may be a saddle point or even a local minimum (see Wu (1986) or McLachlan (1997)). These cases are somewhat pathalogical and it is usually the case that by changing the starting value, the sequence will no longer converge to these odd critical points. In general, one should examine various local MLE's that may be obtained by different starting values.

2.2 EM Evaluation in the Context of Logistic Mixtures

Before we apply the EM algorithm to our model we make an assumptions that our process's dependence upon the past is limited to the most recent p periods, i.e.

$$g(y_t \mid w_t, y_{t-1}, w_{t-1}, ..., y_{-p+1}, w_{-p+1}) = g(y_t \mid w_t, y_{t-1}, w_{t-1}, ..., y_{t-p}, w_{t-p}) \text{ and}$$
$$g(i_t \mid w_t, y_{t-1}, w_{t-1}, ..., y_{-p+1}, w_{-p+1}) = g(i_t \mid w_t, y_{t-1}, w_{t-1}, ..., y_{t-p}, w_{t-p}).$$

If we denote $\sigma(W_t, Y_{t-1}, W_{t-1}, ..., Y_{t-p}, W_{t-p})$ by $\mathcal{G}_{t-1}$ in the same way we wrote $\mathcal{P}_{t-1} = \sigma(W_t, Y_{t-1}, W_{t-1}, ..., Y_{-p+1}, W_{-p+1})$ then we may rewrite our likelihood products in terms of $\mathcal{G}_{t-1}$ instead of $\mathcal{P}_{t-1}$, i.e. $\prod g(y_t \mid \mathcal{G}_{t-1}; \psi)$.

2.2.1 Calculating the E-Step

To implement the algorithm we take ψ^k as given and evaluate $M\left(\psi, \underline{y}, \underline{w}, \psi^k\right)$. Our goal is to rewrite this expression in terms of equations (2.1) and (2.2) – densities and probabilities we know how to evaluate. From (2.9) we have

$$M\left(\psi, \underline{y}, \underline{w}, \psi^k\right) = \sum_{t=1}^{T} \left(\log g(y_t, I_t = 1 \mid \mathcal{G}_{t-1}; \psi) \cdot \mathbb{P}\left[I_t = 1 \mid y_t, \mathcal{G}_{t-1}; \psi^k\right] + \right.$$
$$\left. \log g(y_t, I_t = 0 \mid \mathcal{G}_{t-1}; \psi) \cdot \mathbb{P}\left[I_t = 0 \mid y_t, \mathcal{G}_{t-1}; \psi^k\right]\right).$$

From equations (2.4)–(2.6) we have

$$\log g(y_t, I_t = 1 \mid \mathcal{G}_{t-1}; \psi) = \log\left(f(y_t \mid I_t = 1, \mathcal{G}_{t-1}; \beta_1, \phi_1)\mathbb{P}\left[I_t = 1 \mid \mathcal{G}_{t-1}; \gamma\right]\right)$$
$$= \log f(y_t \mid I_t = 1, \mathcal{G}_{t-1}; \beta_1, \phi_1) + \log \mathbb{P}\left[I_t = 1 \mid \mathcal{G}_{t-1}; \gamma\right].$$

Similarly, for the case $I_t = 0$ we have

$$\log g(y_t, I_t = 0 \mid \mathcal{G}_{t-1}; \psi) = \log f(y_t \mid I_t = 0, \mathcal{G}_{t-1}; \beta_0, \phi_0) + \log \mathbb{P}\left[I_t = 0 \mid \mathcal{G}_{t-1}; \gamma\right].$$

If we define $p_t^k = \mathbb{P}\left[I_t = 1 \mid y_t, \mathcal{G}_{t-1}; \psi^k\right]$ then we may write

$$M\left(\psi, \underline{y}, \underline{w}, \psi^k\right) = \tag{2.24}$$

$$\sum_{t=1}^{T} p_t^k \left(\log f(y_t \mid I_t = 1, \mathcal{G}_{t-1}; \beta_1, \phi_1) + \log \mathbb{P}\left[I_t = 1 \mid \mathcal{G}_{t-1}; \gamma\right]\right) + \tag{2.25}$$

$$\sum_{t=1}^{T} \left(1 - p_t^k\right) \cdot \left(\log f(y_t \mid I_t = 0, \mathcal{G}_{t-1}; \beta_0, \phi_0) + \log \mathbb{P}\left[I_t = 0 \mid \mathcal{G}_{t-1}; \gamma\right]\right). \tag{2.26}$$

Equations (2.1) and (2.2) may be used to evaluate the all the terms except for p_t^k. To determine p_t^k we see

$$p_t^k = \mathbb{P}\left[I_t = 1 \mid y_t, \mathcal{G}_{t-1}; \psi^k\right] = \frac{g(y_t, I_t = 1 \mid \mathcal{G}_{t-1}; \psi^k)}{g(y_t \mid \mathcal{G}_{t-1}; \psi^k)} \tag{2.27}$$

$$= \frac{f(y_t \mid I_t = 1, \mathcal{G}_{t-1}; \beta_1^k, \phi_1^k) \cdot \mathbb{P}\left[I_t = 1 \mid \mathcal{G}_{t-1}; \gamma^k\right]}{\sum_{i=0,1} f(y_t \mid I_t = i, \mathcal{G}_{t-1}; \beta_i^k, \phi_i^k) \cdot \mathbb{P}\left[I_t = i \mid \mathcal{G}_{t-1}; \gamma^k\right]}. \tag{2.28}$$

where $f(y_t \mid I_t = i, \mathcal{G}_{t-1}; \beta_i^k, \phi_i^k)$ and $\mathbb{P}\left[I_t = 1 \mid \mathcal{G}_{t-1}; \gamma^k\right]$ may be evaluated using equations (2.1) and (2.2) and $\psi^k = (\beta_1^k, \beta_0^k, \phi_1^k, \phi_0^k, \gamma^k)$. Thus it is easy to express $M\left(\psi, \underline{y}, \underline{w}, \psi^k\right)$ in terms of known functions and parameters.

Before moving forward it is worthwhile to consider the relationship between

$$p_t^k = \mathbb{P}\left[I_t = 1 \mid y_t, \mathcal{G}_{t-1}; \psi^k\right] \text{ and}$$

$$\mathbb{P}\left[I_t = 1 \mid \mathcal{G}_{t-1}; \psi^k\right] = \exp(Z_t'\gamma^k)/(1 + \exp(Z_t'\gamma^k)).$$

p_t^k represents the expectation of I_t conditional on $y_t, w_t, y_{t-1}, w_{t-1}, ..., y_{t-p}, w_{t-p}$ while the second probability gives the expectation conditional on only w_t, y_{t-1}, $w_{t-1}, ..., y_{t-p}$, w_{t-p}. In other words p_t^k updates this second probability by taking into account the contemporaneous value of Y_t. While we would not use the p_t^k for predicting the conditional mean of Y_t (it would be cheating as p_t^k already incorporates knowledge of y_t) it is fine to use it for fitting and estimation purposes.

2.2.2 Calculating the M Step

Recall that given ψ^k and $M\left(\psi, \underline{y}, \underline{w}, \psi^k\right)$ we define

$$\psi^{k+1} = \text{Arg} \max_{\psi} M\left(\psi, \underline{y}, \underline{w}, \psi^k\right)$$

From equations (2.24) – (2.26) we may write $M\left(\psi, \underline{y}, \underline{w}, \psi^k\right) = H_1(\beta_1, \phi_1, p^k) + H_0(\beta_0, \phi_0, p^k) + H_2(\gamma, p^k)$ where

$$H_1(\beta_1, \phi_1, p^k) = \sum_{t=1}^{T} p_t^k \cdot \left[\log f(y_t \mid I_t = 1, \mathcal{G}_{t-1}; \beta_1, \phi_1)\right], \tag{2.29}$$

$$H_0(\beta_0, \phi_0, p^k) = \sum_{t=1}^{T} (1 - p_t^k) \cdot \left[\log f(y_t \mid I_t = 0, \mathcal{G}_{t-1}; \beta_0, \phi_0)\right], \text{ and} \tag{2.30}$$

$$H_2(\gamma, p^k) = \sum_{t=1}^{T} p_t^k \cdot \left[\log \mathbb{P}\left[I_t = 1 \mid \mathcal{G}_{t-1}; \psi\right]\right] \tag{2.31}$$

$$+ (1 - p_t^k) \cdot \left[\log P\left[I_t = 0 \mid \mathcal{G}_{t-1}; \psi\right]\right] \tag{2.32}$$

In the definitions of H_1, H_0, and H_2 above, p^k is shorthand for the vector $(p_1^k, \ldots, p_t^k, \ldots, p_T^k)$. Because the different elements of $\psi = (\beta_1', \beta_0', \phi_1, \phi_0, \gamma')'$ are so neatly separated into distinct terms the notation $\psi^{k+1} = \arg \max_{\psi} M(\psi \mid \psi^k)$ means given p^k choose

(β_1, ϕ_1) to maximize $H_1(\beta_1, \phi_1, p^k)$,

(β_0, ϕ_0) to maximize $H_0(\beta_0, \phi_0, p^k)$, and

γ to maximize $H_2(\gamma, p^k)$.

To maximize the above, note that for equations (2.29) and (2.30) optimization in these case correspond to finding MLEs of standard GLM models with prior weights p_t^k (see p. 29 McCullough and Nelder (1989)).

2.2.3 Maximizing $H_1(\beta_1, \phi_1, p^k)$

Many statistical software packages will perform estimation of weighted GLM distributions. This is one simple option for maximizing $H_1(\beta_1, \phi_1, p^k) = \sum_{t=1}^{T} p_t^k \cdot [\log f(y_t \mid I_t = 1, \mathcal{G}_{t-1}; \beta_1, \phi_1)]$. A second approach is to use a Newton-Raphson method. From elementary differentiation we know $\log f(y_t \mid \mathcal{G}_{t-1}; \beta_1, \phi_1)$, is concave w.r.t. β_1 for GLM distributions. This concavity implies concavity of $H_1(\beta_1, \phi_1, p^k)$ as $H_1(\beta_1, \phi_1, p^k)$ is a weighted sum of concave functions. Consequently, this implies a Newton-Raphson or Fisher scoring method should work well for finding MLE's for β_1^{k+1}.

Once β_1^{k+1} has been obtained, the scale parameter, ϕ_1^{k+1}, can be estimated by using differentiation to minimize

$$\sum_{1}^{T} p_t^k \cdot \left(\frac{\left[y_t X_{t1}' \beta_1^{k+1} - b(X_{t1}' \beta_1^{k+1}) \right]}{\phi_1} + c_1(y_t, \phi_1) \right).$$

In the case of normally distributed component densities we obtain

$$\phi_1^{k+1} = \frac{1}{\sum p_t^k} \sum_{t=1}^{T} (y_t - X_{t1}' \beta_1^{k+1})^2 \cdot p_t^k$$

The estimation of β_0^{k+1} and ϕ_0^{k+1} from $H_0(\beta_0, p^k)$ is completely analogous, except that $1 - p_t^k$ are used as weights, instead of p_t^k.

2.2.4 Maximizing $H_2(\gamma, p^k)$

Recall that $\gamma^{k+1} = \arg \max_{\gamma} H_2(\gamma, p_k) =$

$$\arg \max_{\gamma} \sum_{t=1}^{T} p_t^k \cdot \log P\left[I_t = 1; \mathcal{G}_{t-1}; \gamma\right] + (1 - p_t^k) \cdot \log P\left[I_t = 0; \mathcal{G}_{t-1}; \gamma\right] \quad (2.33)$$

By differentiating twice we can see for logistic regression $\log P\left[I_t = 1 \mid \mathcal{G}_{t-1}; \gamma\right]$ and $\log P\left[I_t = 0 \mid \mathcal{G}_{t-1}; \gamma\right]$ are concave in γ, hence, so is $H_2(\gamma, p_k)$ as it is a sum

of weighted sums of concave functions. Therefore, again a Fisher Scoring type algorithm should quickly find γ_{k+1}. Using (2.1) we may rewrite (2.33) as

$$\gamma_{k+1} = \arg \max_{\gamma} \sum_{t=1}^{T} p_t^k \log \left[\frac{\exp(Z_t'\gamma)}{1+\exp(Z_t'\gamma)} \right] + (1-p_t^k) \log \left[\frac{1}{1+\exp(Z_t'\gamma)} \right]$$

This expression resembles the log likelihood of logistic regression except I_t and $1-I_t$ are replaced by p_t^k and $1-p_t^k$. Consequently, estimates of γ^{k+1} could be obtained through logistic regression with the p_t^ks as the dependent variables with a computer package that allows such substitution. If this is not available then the concavity of $H_2(\gamma, p^k)$ implies Newton-Raphson or Fisher scoring algorithms should work to find γ^{k+1}.

2.3 Estimation procedure

To fit all the parameters in a particular model, the above procedures are combined in the following way:

1. Initialize the model by giving starting values for $\psi = (\beta_0', \beta_1', \phi_0', \phi_1', \gamma')', \psi^1$, and a tolerance level, ϵ.

2. Given ψ^k, compute $\{p_t^k\}$ using (2.28), (2.1), and (2.2). Next, obtain β_0^{k+1}, β_1^{k+1}, ϕ_0^{k+1},and ϕ_1^{k+1} with one of the algorithms described in Section 2.2.3. γ^{k+1} is obtained by following one of the optimizing procedures in Section 2.2.4. Repeat this step until $\left|\left|\psi^{k+1} - \psi^k\right|\right| < \epsilon$.

Chapter 3

Consistency and Asymptotic Normality

In this chapter we discuss large sample properties of the maximum likelihood estimators of a correctly specified logistic mixture model. We will initially assume the model satisfies various conditions that allow us to prove a particular result. Then we choose a more specific model and show how these conditions may be validated for this particular model choice.

3.1 Consistency for General Logistic Mixtures

First we recall the general model:

$$
\begin{aligned}
f(y_t \mid X_t; \beta, \phi) =& \exp\left(\frac{y_t X_t'\beta - b\left(X_t'\beta\right)}{\phi} + c(y_t, \phi)\right), \\
\mathbb{P}\left[I_t = 1 \mid Z_t; \gamma\right] =& \frac{\exp\left(Z_t'\gamma\right)}{1 + \exp\left(Z_t'\gamma\right)}, \\
g(y_t \mid \mathcal{G}_{t-1}; \psi) =& \mathbb{P}\left[I_t = 1 \mid Z_t; \gamma\right] \cdot f(y_t \mid X_{t1}; \beta_1, \phi_1) + \\
& \left(1 - \mathbb{P}\left[I_t = 1 \mid Z_t; \gamma\right]\right) \cdot f(y_t \mid X_{t0}; \beta_0, \phi_0), \\
\text{where } \psi =& (\beta_1', \beta_0', \gamma', \phi_1, \phi_0)' \text{ and } \mathcal{G}_{t-1} = \sigma(X_{t1}, X_{t0}, Z_t).
\end{aligned}
$$

We assume the true conditional distribution of $Y_t \mid \mathcal{G}_{t-1}$ is given by $g(y_t \mid \mathcal{G}_{t-1}; \psi^*)$ for some $\psi^* \in \Psi$, a subset of $\mathbb{R}^{3q+2}$, where q is the common dimension of β_1, β_0, and γ and the last two dimensions are for ϕ_1 and ϕ_0. The conditions we will use for demonstrating consistency are:

3.1.A *$\{Y_t, X_{t1}, X_{t0}, Z_t\}$ is asymptotically stationary with W denoting a random vector in R^{3q+1} that has the joint stationary distribution. Furthermore, $Y_t, X_{t1}, X_{t0},$ and Z_t obey a strong law of large numbers in the sense that if $h(\cdot)$ is a measurable and integrable function of W then*

$$\frac{1}{T}\sum h(Y_t, X_{t1}, X_{t0}, Z_t) \xrightarrow{a.s.} \mathbb{E}\left[h(W)\right].$$

We denote the conformably partitioned components of W as W_Y, W_{X1}, W_{X0} and W_Z. We will be more specific below.

This condition essentially presumes the existence of a stationary distribution that describes the long term behaviour of our process. Such a distribution might arise if we are able to view $\{Y_t, X_{t1}, X_{t0}, Z_t\}$ as a Markov process with W having the invariant distribution. We will discuss this condition at length in the next section, and prove that it holds in a logistic mixture of Gaussian AR(1) processes.

3.1.B *$\mathbb{E}[\log g(W_y \mid W_{X1}, W_{X0}, W_Z; \psi)] < \infty$ for all $\psi \in \Psi$ and is continuous in ψ. By $g(W_y \mid W_{X1}, W_{X0}, W_Z; \psi)$ we mean $g(Y_t \mid \mathcal{G}_{t-1}; \psi)$ with $Y_t, X_{t1}, X_{t0},$ and Z_t replaced by $W_Y, W_{X1}, W_{X0},$ and W_Z. To save space we will write $g(W_y \mid W_{X1}, W_{X0}, W_Z; \psi)$ more concisely as $g(W; \psi)$.*

It is important to note that the expectation above is unconditional – not conditional, i.e. we do not mean

$$\mathbb{E}[\log g(W_y \mid W_{X1}, W_{X0}, W_Z; \psi) \mid W_{X1}, W_{X0}, W_Z].$$

The unconditional expectation is computed with respect to the measure associated with the distribution of W, i.e.

$$\mathbb{E}\left[\log g(W_y \mid W_{X1}, W_{X0}, W_Z; \psi)\right] = \int \log g(w_y \mid w_{X1}, w_{X0}, w_Z; \psi)\mathbb{P}(dw)$$
$$\text{where } \mathbb{P}\left[W \in A\right] = \int_A \mathbb{P}(dw).$$

It is this unconditional expectation that is relevant for the strong law results and the majority of expectations in this chapter are computed this way. Conditional expectations will be denoted in the usual manner – with the vertical bar: $\mathbb{E}\left[\cdot \mid \cdot\right]$.

3.1.C *There exists a compact set $K \subseteq \Psi$, with $\psi^* \in K$ such that if $\psi \in K$ and $\psi \neq \psi^*$ then $\mathbb{E}\left[\log g(W;\psi)\right] < \mathbb{E}\left[\log g(W;\psi^*)\right]$.*

This is an identifiability condition with K chosen to restrict the parameter space. As discussed later (in Section 3.2.3) the likelihood for mixture models is symmetric with respect to some axis in the parameter space. By examining only those parameter values in K we may say that ψ^* is the unique parameter that maximizes $\mathbb{E}\left[\log g(W;\psi)\right]$. This will be addressed at length below.

3.1.D *Given $B_\rho(\psi) = \{\psi' \in K : \ ||\psi' - \psi|| < \rho\}$ we define*

$$g^*(W;\psi,\rho) = \sup_{\psi' \in B_\rho(\psi)} g(W;\psi').$$

This condition is that $\mathbb{E}\left[\log g^*(W;\psi,\rho)\right]$ exists and
$\lim_{\rho\to 0}\mathbb{E}\left[\log g^*(W;\psi,\rho)\right] = \mathbb{E}\left[\log g(W;\psi)\right]$ for all $\psi \in K$.

With the exception of Condition 3.1.A, these conditions are easily met by a broad set of models. Most of the conditions are consequences of $\log g(W;\psi)$ being sufficiently smooth with respect to ψ, and with derivatives that may be

bounded by integrable functions (allowing applications of the dominated convergence theorem).

With this set of conditions we may prove the following theorem:

Theorem 3.2. *Under Conditions* 3.1.A–3.1.D *we can show there exists a sequence of local maximum (partial) likelihood estimates* $\{\hat{\psi}_T\} \in K$ *such that* $\hat{\psi}_T \xrightarrow{a.s.} \psi^*$.

By local maximum likelihood estimates we mean the $\{\hat{\psi}_T\}$ maximize the likelihood only over some fixed neighborhood of ψ^* (K in this case), not necessarily the entire parameter space. The proof of this theorem adapts Wald's approach (1949) to time dependent data.

Proof. Let $\delta > 0$ be given. We want to show

$$\mathbb{P}\left[\lim \left|\left|\hat{\psi}_T - \psi^*\right|\right| > \delta\right] = 0 \text{ where} \tag{3.1}$$
$$\hat{\psi}_T = \text{Arg} \max_{\psi \in K} \sum_{t=1}^{T} \log g(y_t \mid \mathcal{G}_{t-1}; \psi).$$

Here and elsewhere that we discuss the maximum likelihood estimate, $\hat{\psi}_T$, we assume there is some lexicographical rule that allows us to break ties in the event there are two or more elemens of K that maximize the likelihood. For example, we might define $\hat{\psi}_T$ to be that element with the smallest value of $\hat{\beta}_1$ (or $\hat{\phi}_1$ if there is more than one minimizing value of the likelihood with the same smallest value of $\hat{\beta}_1$).

Without loss of generality we may assume δ is sufficiently small so that by Conditions 3.1.B and 3.1.C we can find a $B_\delta(\psi^*)$ such that

$$\sup_{\psi \in K \setminus B_\delta(\psi^*)} \mathbb{E}\left[\log g(W; \psi)\right] < \mathbb{E}\left[\log g(W; \psi^*)\right].$$

Next we choose ϵ such that

$$\mathbb{E}[\log g(W;\psi^*)] - \sup_{\psi \in K \backslash B_\delta(\psi^*)} \mathbb{E}[\log g(W;\psi)] > \epsilon > 0.$$

Now, for each $\psi \in K$ we use Condition 3.1.D to find ρ_ψ such that

$$0 < \mathbb{E}[\log g^*(W;\psi,\rho_\psi)] - \mathbb{E}[\log g(W;\psi)] < \epsilon/2 \text{ which implies}$$

$$\mathbb{E}[\log g^*(W;\psi,\rho_\psi)] - \mathbb{E}[\log g(W;\psi^*)] < -\epsilon/2 \text{ for all } \psi \in K \backslash B_\delta(\psi^*). \quad (3.2)$$

By construction, $\{B_{\rho_\psi}(\psi) \colon \psi \in K \backslash B_\delta(\psi^*)\}$ form an open covering of $K \backslash B_\delta(\psi^*)$ and thus admit a finite subcover we denote as $\{B_{\rho_1}(\psi_1), ..., B_{\rho_L}(\psi_L)\}$. As a bridge to proving Theorem 3.2 we first show

$$\mathbb{P}\left[\lim_T \sup_{\psi \in K \backslash B_\delta(\psi^*)} \sum_{t=1}^T \log g(y_t \mid \mathcal{G}_{t-1};\psi) - \sum_{t=1}^T \log g(y_t \mid \mathcal{G}_{t-1};\psi^*) = -\infty\right] = 1. \quad (3.3)$$

To show this we note that because $\{B_{\rho_1}(\psi_1), ..., B_{\rho_L}(\psi_L)\}$ covers $K \backslash B_\delta(\psi^*)$

$$\sup_{\psi \in K \backslash B_\delta(\psi^*)} \sum_{t=1}^T \log g(y_t \mid \mathcal{G}_{t-1};\psi) \leq$$

$$\max\left\{\sum_{t=1}^T \log g^*(y_t \mid \mathcal{G}_{t-1};\psi_1,\rho_1), ..., \sum_{t=1}^T \log g^*(y_t \mid \mathcal{G}_{t-1};\psi_L,\rho_L)\right\}.$$

So equation (3.3) will follow if we show

$$\mathbb{P}\left[\lim_T \sum_{t=1}^T \log g^*(y_t \mid \mathcal{G}_{t-1};\psi_l,\rho_l) - \sum_{t=1}^T \log g(y_t \mid \mathcal{G}_{t-1};\psi^*) = -\infty\right] = 1 \quad (3.4)$$

for $l = 1...L$. To prove (3.4) we see that because of Condition 3.1.A and equation (3.2) we can conclude

$$\frac{1}{T}\sum_{t=1}^T \log g^*(y_t \mid \mathcal{G}_{t-1};\psi_l,\rho_l) - \frac{1}{T}\sum_{t=1}^T \log g(y_t \mid \mathcal{G}_{t-1};\psi^*) \quad (3.5)$$

$$\xrightarrow{a.s} \mathbb{E}[\log g^*(W;\psi_l,\rho_l)] - \mathbb{E}[\log g(W;\psi^*)] < -\epsilon/2 < 0 \quad (3.6)$$

$$\text{so } \sum_{t=1}^T \log g^*(y_t \mid \mathcal{G}_{t-1};\psi_l,\rho_l) - \sum_{t=1}^T \log g(y_t \mid \mathcal{G}_{t-1};\psi^*) \xrightarrow{a.s} -\infty \quad (3.7)$$

and the demonstration of (3.4) is complete. To show that this proves Theorem 3.2 we note that our definition of $\hat{\psi}_T$ implies

$$\sum_{t=1}^{T} \log g(y_t \mid \mathcal{G}_{t-1}; \hat{\psi}_T) \geq \sum_{t=1}^{T} \log g(y_t \mid \mathcal{G}_{t-1}; \psi^*) \text{ for all } T.$$

Now suppose that $\hat{\psi}_T \overset{a.s.}{\not\to} \psi^*$. This would mean there exists a δ_0 such that

$$\mathbb{P}\left[\left|\left|\hat{\psi}_T - \psi^*\right|\right| > \delta_0 \text{ infinitely often}\right] > 0.$$

Now $\left|\left|\hat{\psi}_T - \psi^*\right|\right| > \delta_0$ implies $\hat{\psi}_T \in K \backslash B_{\delta_0}(\psi^*)$ further implying

$$\begin{aligned}\sup_{\psi \in K \backslash B_{\delta_0}(\psi^*)} \sum_{t=1}^{T} \log g(y_t \mid \mathcal{G}_{t-1}; \psi) &\geq \sum_{t=1}^{T} \log g(y_t \mid \mathcal{G}_{t-1}; \hat{\psi}_T) \\ &\geq \sum_{t=1}^{T} \log g(y_t \mid \mathcal{G}_{t-1}; \psi^*).\end{aligned}$$

So if

$$\mathbb{P}\left[\overline{\lim} \left|\left|\hat{\psi}_T - \psi^*\right|\right| > \delta_0\right] > 0 \text{ this means}$$

$$\mathbb{P}\left[\overline{\lim} \sup_{\psi \in K \backslash B_{\delta_0}(\psi^*)} \sum_{t=1}^{T} \log g(y_t \mid \mathcal{G}_{t-1}; \psi) - \sum_{t=1}^{T} \log g(y_t \mid \mathcal{G}_{t-1}; \psi^*) \geq 0\right] > 0. \tag{3.8}$$

But this contradicts our result in (3.3) with $\delta = \delta_0$. Therefore $\hat{\psi}_T \xrightarrow{a.s.} \psi^*$. □

As we have proven the result under the stated conditions we now consider whether these conditions are valid for logistic mixtures of AR(1) processes.

3.2 Consistency of a Logistic Mixture of Gaussian AR(1) Processes

Here we examine the conditions above in the case of a specific type of logistic mixture model. The mixture of AR(1) normal processes is relatively simple, yet

complex enough to be used in applications as normal AR(p) models are used in threshold and hidden Markov model regressions (see Tong (1983,1990) for threshold models and Hamilton (1990) and Chapter 22 of Hamilton (1994) for hidden Markov models). Although we present results for AR(1) processes we anticipate little difficulty in extending the results to AR(p) mixtures, $p \geq 1$. We place the AR(1) mixture model in our previous notation as follows: let

$$f(y_t \mid Y_{t-1}; \beta, \phi) = \exp\left(-\frac{(y_t - Y_{t-1}\beta)^2}{2\phi} - \frac{1}{2}\ln 2\pi\phi\right) \tag{3.9}$$

$$\mathbb{P}\left[I_t = 1 \mid Y_{t-1}; \gamma\right] = \frac{\exp(\gamma_0 + Y_{t-1}\gamma_1)}{1 + \exp(\gamma_0 + Y_{t-1}\gamma_1)}. \tag{3.10}$$

Our general model of a logistic mixture of this type is

$$g(y_t \mid \mathcal{G}_{t-1}; \psi) = g(y_t \mid Y_{t-1}; \psi) = f(y_t \mid Y_{t-1}; \beta_1, \phi_1)\mathbb{P}\left[I_t = 1 \mid Y_{t-1}; \gamma\right] + \tag{3.11}$$

$$f(y_t \mid Y_{t-1}; \beta_0, \phi_0)\mathbb{P}\left[I_t = 0 \mid Y_{t-1}; \gamma\right], \tag{3.12}$$

$$\text{where } \psi = (\beta_0, \beta_1, \phi_0, \phi_1, \gamma_0, \gamma_1)'. \tag{3.13}$$

We now set about proving the conditions in Section 3.1 are valid in this model. We first define our parameter space, Ψ. We assume the true parameter $\psi^* = (\beta_1^{*\prime}, \phi_1^*, \beta_0^{*\prime}, \phi_0^*, \gamma^{*\prime})'$ lies in the compact space Ψ given by

$$\beta_1, \beta_0 \in [-1 + \epsilon_1, 1 - \epsilon_2]$$

$$\gamma_0 \in [M_1, M_2], \gamma_1 \in [M_3, M_4] \text{ and}$$

$$\phi_0, \phi_1 \in [\phi_{min}, \phi_{max}], 0 < \phi_{min} < \phi_{max} < \infty$$

where ϵ_i denotes an arbitrarily small positive constant, and M_j an arbitrary constant. In practice this seems reasonable as the applications considered should allow the investigator to place bounds on the parameters. The next condition is more difficult to verify.

3.2.1 Condition 3.1.A

Proving the stationarity and ergodicity of $\{Y_t\}$ described in Condition 3.1.A is a lengthy and detailed process. We utilize results of continuous state Markov chains, following the approach discussed in works by Chan (1994), Nummelin (1984), and Tweedie (1975). We first introduce some notation.

Let $\{Y_t\}$ denote a sequence of random variables on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that

$$P(x, A) \stackrel{\triangle}{=} \mathbb{P}[Y_t \in A \mid Y_{t-1} = x] \text{ is independent of } t,$$

$$\mathbb{P}[Y_t \in A \mid Y_{t-1}, Y_{t-2}, ..., Y_0] = \mathbb{P}[Y_t \in A \mid Y_{t-1}]$$

$$P(x, \cdot) \text{ is a probability measure on } (\Omega, \mathcal{F}) \text{ for all } x \in \mathbb{R}$$

$$P(\cdot, A) \text{ is a } \mathcal{F}-\text{measurable function for all A} \in \mathcal{F}.$$

Then we will refer to $\{Y_t\}$ as a homogeneous Markov chain with transition kernel $P(x, A)$. If we set

$$P(x, A) = \int_A g(y_t \mid Y_{t-1} = x; \psi^*)\, dy_t \tag{3.14}$$

where $g(\cdot)$ is defined in (3.11) then we can see that $\{Y_t\}$, our logistic mixture of Gaussian AR(1) processes, is a homogeneous Markov chain on $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P})$. Here $\mathcal{B}(\mathbb{R})$ denotes the Borel measurable sets derived from open intervals of the real line. The first question we address is that of aperiodicity and irreducibility of $\{Y_t\}$. If necessary the reader may consult Nummelin or Chan for these definitions. To prove irreducibility and aperiodicity it is sufficient to prove $P(x, A) > 0$ for all $x \in \mathbb{R}$ and all Borel measurable sets A with positive Lebesgue measure (see Chan (1994) for why this is sufficient). We can easily see this condition is met for our mixture model:

Lemma 3.3. *Let $\{Y_t\}$ have the transition kernel indicated in (3.14) and A be any set in $\mathcal{B}(\mathbb{R})$ with positive Lebesgue measure. Then $P(x, A) > 0$ for all $x \in \mathbb{R}$.*

Proof.

$$P(x, A) = \int_A g(y_t \mid Y_{t-1} = x; \psi^*)\, dy_t \tag{3.15}$$

$$\geq \int_A f(y_t \mid Y_{t-1} = x; \beta_1^*, \phi_1^*)\mathbb{P}[I_t = 1 \mid Y_{t-1} = x; \gamma^*]\, dy_t > 0. \tag{3.16}$$

From its definition in (3.10), $\mathbb{P}[I_t = 1 \mid Y_{t-1} = x; \gamma] > 0$ for any x, and the component density corresponding to regime 1, $f(y_t \mid Y_{t-1} = x; \beta_1, \phi_1)$, is a strictly positive function on A. Therefore the last inequality in (3.16) is strict because the integrand is strictly positive and the set A has positive measure. □

To obtain useful results regarding ergodicity we need to introduce the concept of a small set. In continuous state chains the small sets are analogous to the states in a countable or finite state chain. Their definition is available in the aforementioned works – here we give a sufficient condition for a set to be small.

Theorem 3.4. (Theorem 3.2 Chan (1994)) *Let $\{Y_t\}$ be a Markov chain on $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P})$ such that $P(x, A)$ is continuous in x for every fixed $A \in \mathcal{B}(\mathbb{R})$ and $P(x, A) > 0$ whenever $\mu(A) > 0$ where μ denotes Lebesgue measure. If $\mu(A) > 0$ then A is a small set.*

This theorem does not characterize all small sets for our Markov chain but it is sufficient for our purposes. In our definition of $P(\cdot, A)$ in (3.14) it is clear our process satisfies the continuity condition in Theorem 3.4 so we may view Borel measurable sets with positive measure as small. We also need the concept of positive Harris recurrence to obtain laws of large numbers. A precise definition

is available in Nummelin and Chan though roughly stated positive Harris recurrence means if $\mu(A) > 0$ then $\mathbb{P}[Y_t \in A \text{ infinitely often } | Y_0 = x] = 1$ for all x in $\mathbb{R}$. Positive Harris recurrence is important because such aperiodic Markov chains have a unique invariant distribution such that Y_t is strongly stationary if Y_0 is distributed according to the invariant measure. Also, it will be the case that under very general conditions, the distribution of a positive Harris recurrent Markov chain, Y_t, will converge to the invariant distribution regardless of the distribution of Y_0. Henceforth we will denote this invariant probability measure as π_Y. Nummelin provides a criterion for checking whether an irreducible Markov chain is positive Harris recurrent:

Theorem 3.5. (Proposition 5.10 in Nummelin (1984)) *An irreducible Markov chain $\{Y_t\}$ is positive Harris recurrent if there exist a non-negative measurable function l, a small set C and a constant $\tau > 0$ such that*

$$\mathbb{E}[l(Y_t) \mid Y_{t-1} = x] \leq l(x) - \tau, \text{ for all } x \notin C, \text{ and} \tag{3.17}$$

$$\sup_{x \in C} \int_{C'} l(y)\mathbb{P}(x, dy) < \infty. \tag{3.18}$$

Here the C' denotes the complement of C. We are now in a position to provide theorems that will help us determine when a sum converges to an expected value. This first theorem gives a test for integrability with respect to π_Y:

Theorem 3.6. (Proposition 5.3 in Chan (1994)) *Let $\{Y_t\}$ by aperiodic and positive Harris recurrent and h be a non-negative measurable function. In order that $\int h(y)\pi_Y(dy) < \infty$ it is sufficient that for some small set C with $\mu(C) > 0$ and $\int_C h(x)\pi_Y(x) < \infty$, and some measurable function l with $l(x) \geq h(x), x \in$*

C', the following hold

$$\int_{C'} l(y)P(x,dy) \leq l(x) - h(x), x \in C' \text{ and} \tag{3.19}$$

$$\sup_{x \in C} \int_{C'} l(y)P(x,dy) < \infty. \tag{3.20}$$

These theorems are primarily useful in allowing us to use this final theorem:

Theorem 3.7. (See Proposition 5.6 in Chan (1994)) *Suppose $\{Y_t\}$ is aperiodic and positive Harris recurrent with invariant measure π_Y. The for any π_Y-integrable function f, and any initial distribution λ of Y_0 we have*

$$\frac{1}{T}\sum_{t=1}^{T} f(Y_t) \xrightarrow{a.s.} \int f(y)\pi_Y(dy).$$

In order to apply this useful theorem we must first show our logistic mixture model is positive Harris recurrent.

Theorem 3.8. *Let Y_t have the transition kernel given by* (3.14) *and define $l(x) = |x|$. Then there exists a small set $C = [-c, c]$ and a constant τ such that the conditions of Theorem 3.5 are met.*

Proof. First we find C and τ so (3.17) holds. Examining the left-hand side of this equation we obtain

$$\begin{aligned}\mathbb{E}[l(Y_t) \mid Y_{t-1} = x] =& p_1(x) \int |y_t| \, f(y_t \mid Y_{t-1} = x; \beta_1, \phi_1) dy_t \\ &+ (1 - p_1(x)) \int |y_t| \, f(y_t \mid Y_{t-1} = x; \beta_0, \phi_0) dy_t \end{aligned} \tag{3.21}$$

where f and $p_1(x)$ are shorthand for the terms respectively defined in (3.9) and (3.10). Since conditional on $Y_{t-1} = x$, Y_t has a $N(x\beta, \phi)$ distribution, one can show

$$\int |y_t| \, f(y_t \mid Y_{t-1} = x; \beta, \phi) dy_t = \frac{2\phi^{1/2}}{\sqrt{2\pi}} \exp\left(\frac{-x^2\beta^2}{2\phi}\right) + x\beta\left[1 - 2\Phi\left(\frac{-x\beta}{\sqrt{\phi}}\right)\right], \tag{3.22}$$

where Φ is the cdf for a standard normal random variable. From here it is easy to see that

$$\lim_{|x|\to\infty}\left(|x|\,|\beta_0| - \int |y_t|\, f(y_t \mid Y_{t-1}=x;\beta_0,\phi_0)dy_t\right) = 0 \text{ and}$$
$$\lim_{|x|\to\infty}\left(|x|\,|\beta_1| - \int |y_t|\, f(y_t \mid Y_{t-1}=x;\beta_1,\phi_1)dy_t\right) = 0.$$

From this limiting relationship we see that for ϵ there exists an associated c_ϵ such that

$$\begin{aligned}&\left|\,|x|\,|\beta_0| - \int |y_t|\, f(y_t \mid Y_{t-1}=x;\beta_0,\phi_0)dy_t\right| < \epsilon \text{ and}\\ &\left|\,|x|\,|\beta_1| - \int |y_t|\, f(y_t \mid Y_{t-1}=x;\beta_1,\phi_1)dy_t\right| < \epsilon \text{ for } |x| > c_\epsilon.\end{aligned} \tag{3.23}$$

Define $\beta_\tau = \max\{|\beta_1|, |\beta_0|\}$, and $\tau = \frac{1-\beta_\tau}{2}$. Because $|\beta_1|, |\beta_0| < 1$ we have $\tau > 0$. Choose $c > c_{\tau/2}$ where $c_{\tau/2}$ is defined to satisfy the relations in 3.23 with $\epsilon = \tau/2$. So for $|x| > c > 1$ we see

$$\begin{aligned}\mathbb{E}\left[l(Y_t) \mid Y_{t-1}=x\right] =& p_1(x)\int |y_t|\, f(y_t \mid Y_{t-1}=x;\beta_1,\phi_1)dy_t\\ &+ (1-p_1(x))\int |y_t|\, f(y_t \mid Y_{t-1}=x;\beta_0,\phi_0)dy_t\\ \leq& p_1(x)\,|x|\,|\beta_1| + (1-p_1(x))\,|x|\,|\beta_0|\\ \leq& |x|\,\beta_\tau + \frac{\tau}{2} = |x|\,(1-2\tau) + \frac{\tau}{2}\\ \leq& |x| - 2\tau + \frac{\tau}{2} < |x| - \tau.\end{aligned}$$

This concludes the proof for the first part of Theorem 3.5. The last part involves showing

$$\sup_{x\in C}\int_{C'} l(y)P(x,dy) < \infty.$$

But this is straightforward since

$$\int_{C'} l(y)\mathbb{P}(x,dy) < \int l(y)\mathbb{P}(x,dy) <$$
$$\int_{\mathbb{R}} |y_t|\, f(y_t \mid Y_{t-1} = x; \beta_1, \phi_1)dy_t + \int_{\mathbb{R}} |y_t|\, f(y_t \mid Y_{t-1} = x; \beta_0, \phi_0)dy_t$$

From equation (3.22) we see these integrals are continuous functions of x and hence bounded if x is required to lie in C. This proves the second condition is met and thus the proof is complete. □

Here we have demonstrated that if $\{Y_t\}$ is a logistic mixture of normal AR(1) distributions then $\{Y_t\}$ is ergodic in the sense that $\frac{1}{T}\sum h(Y_t) \xrightarrow{a.s.} \mathbb{E}[h(Y)]$ where Y as the invariant distribution and $\mathbb{E}[h(Y)] < \infty$. While this is useful it is more important to consider $\underline{W}_t \triangleq (Y_t, Y_{t-1})$ and determine if the ergodic property holds. Though we omit the demonstration (see Chan (1994)) one can show that the aperiodicity, irreducibility, and positive Harris recurrence of Y_t imply the same for $\underline{W}_t$ in our logistic mixture with the modified transition kernel:

$$P_2(x,w) \triangleq \begin{cases} 0 & \text{if } w_2 \neq x_1, \\ P(w_1, x_1) \text{ as in (3.14)} & \text{otherwise} \end{cases} \tag{3.24}$$

$$\text{where } x = (x_1, x_2) \text{ and } w = (w_1, w_2). \tag{3.25}$$

This is important because the aperiodicity, irreducibility, and positive Harris recurrence of (Y_t, Y_{t-1}) implies (see the remarks preceding Theorem 3.5) that there exists a stationary distribution, denoted π_W, such that if $W = (W_1, W_0)$ has distribution π_W then

$$\frac{1}{T}\sum_{t=1}^{T} h(Y_t, Y_{t-1}) \xrightarrow{a.s.} \mathbb{E}[h(W)] \equiv \mathbb{E}[h(W_1, W_0)]$$

whenever $\mathbb{E}[h(W)] < \infty$. This is exactly what we want to demonstrate in condition 3.1.A. Thus we consider this condition validated for our case of logistic mixtures of AR(1) processes.

3.2.2 Condition 3.1.B

For the remainder of this chapter $W = (W_1, W_0)$ will denote a random vector with distribution π_W (the stationary distribution associated with (Y_t, Y_{t-1})) and Y will denote a random variable distributed according to π_Y (the stationary distribution for Y_t). It is clear that both W_1 and W_0 have marginal distributions of π_Y.

In demonstrating $\mathbb{E}[\log g(W); \psi] < \infty$ we will need to show $\mathbb{E}[h(Y); \psi]$ exists for various functions $h(\cdot)$. Rather than tackling each instance separately it is more efficient to show Y has a moment generating function – thus ensuring that all the expectations we will need below do exist.

Theorem 3.9. *Let Y_t have the transition kernel given by* (3.14) *and $s \in \mathbb{R}$. Then there exist an $\xi > 0$ and $C = [-c, c]$ such that the conditions of Theorem 3.6 hold with*

$$h(x) = \exp(sx) \text{ and } l(x) = \exp\left(\frac{\xi x^2}{2\phi}\right)$$

where $\phi = \min\{\phi_0, \phi_1\}$.

The proof of this theorem is very similar to that of Theorem 3.8 except it contains more algebra and calculus. Consequently we omit it. Having established the existence of the moment generating function for Y_t we return to proving Condition 3.1.B.

Our notation here is

$$g(W;\psi) = p\left(W_0;\gamma\right) f\left(W_1 \mid W_0;\beta_1,\phi_1\right) + \left(1-p\left(W_0;\gamma\right)\right) f\left(W_1 \mid W_0;\beta_0,\phi_0\right)$$

$$\text{where } p\left(W_0;\gamma\right) = \frac{\exp\left(\gamma_0+W_0\gamma_1\right)}{1+\exp\left(\gamma_0+W_0\gamma_1\right)} \text{ and }, \tag{3.26}$$

$$f(W_1 \mid W_0;\beta,\phi) = \exp\left(-\frac{(W_1-W_0\beta)^2}{2\phi} - \frac{1}{2}\log 2\pi\phi\right). \tag{3.27}$$

Now because $\phi_1, \phi_0 \geq \phi_{min} > 0$ this implies that for all ψ there exists an M such that $\log g(W;\psi) < M$ – i.e. the log density is bounded above by M. Consequently we only need examine the event $-\infty < \log g < -M$ in order to determine integrability of $\log g$. Since

$$\log g = \log(pf_1 + (1-p)f_0) > \log pf_1 \text{ we see}$$

$$|\log g| < M + |\log pf_1|$$

and reduce the problem to examining the integrability of $\log pf_1 = \log f_1 + \log p$. From our expression for f_1 in 3.27 and recalling that $|\beta_1|, |\beta_0| < 1$ we see that

$$|\log f_1| \leq \frac{\left(|W_1|+|W_0|\right)^2}{2\phi_{min}} + \frac{1}{2}\log 2\pi\phi_{max} \text{ and} \tag{3.28}$$

$$\mathbb{E}\left[\log f_1\right] < \frac{4\mathbb{E}{W_1}^2}{2\phi_{min}} + \frac{1}{2}\log 2\pi\phi_{max}. \tag{3.29}$$

Similarly,

$$|\log p| = |\log(\exp(\gamma_0+W_0\gamma_1)) - \log(1+\exp(\gamma_0+\gamma_1 W_0))|$$

$$< 2\left|\gamma_0+\gamma_1 W_0\right| + \log 2$$

$$\text{so } |\log p| < 2\max\{\left|\gamma_0^{min}\right|, |\gamma_0^{max}|\} + \log 2 + 2\max\{\left|\gamma_1^{min}\right|, |\gamma_1^{max}|\}\left|W_0\right| \text{ and} \tag{3.30}$$

$$\mathbb{E}\left[\log p\right] = 2\max\{\left|\gamma_0^{min}\right|, |\gamma_0^{max}|\} + \log 2 + 2\max\{\left|\gamma_1^{min}\right|, |\gamma_1^{max}|\}\mathbb{E}\left|W_0\right| < \infty. \tag{3.31}$$

So we have shown that $\mathbb{E}[\log g(W;\psi)] < \infty$, and that the integrand is uniformly bounded by the integrable function $M+$ the right-hand side of (3.28) + the right-hand side of (3.30). Therefore we may apply the dominated convergence theorem and conclude that $\mathbb{E}[\log g(W;\psi)]$ is a continuous function in ψ – thus proving Condition 3.1.B holds in this case.

3.2.3 Condition 3.1.C

Recall from its definition that $g(W;\psi) \stackrel{\triangle}{=} g(W_1 \mid W_0;\psi)$ where $W = (W_1, W_0)$ has distribution π_W. In this section we will view $g(w_1 \mid W_0 = w_0;\psi)$ as the conditional density function of W_1 given $W_0 = w_0$. Then we see $g(w_1 \mid W_0 = w_0;\psi)$ is the density of a mixture of two normal distributions: one density is $N(w_0\beta_1, \phi_1)$ and the other is $N(w_0\beta_0, \phi_0)$. The associated mixture probabilities

$$\frac{\exp(\gamma_0 + w_0\gamma_1)}{1+\exp(\gamma_0 + w_0\gamma_1)} \text{ and } \frac{1}{1+\exp(\gamma_0 + w_0\gamma_1)} \text{ are in the open interval } (0,1).$$

Let $\psi^* = (\beta_1^*, \beta_0^*, \phi_1^*, \phi_0^*, \gamma_0^*, \gamma_1^*)'$ denote the true parameters with either $\beta_1^* \neq \beta_0^*$ or $\phi_1^* \neq \phi_0^*$. This restriction is necessary to ensure the mixture is not degnenerate, i.e. a single normal distribution. Without such a restriction the logistic mixture is not correctly specified as its γ parameter would be unidentified. The question of degenerate mixtures is addressed in Chapter 5.

Let $\psi^1 = (\beta_1^1, \beta_0^1, \phi_1^1, \phi_0^1, \gamma_0^1, \gamma_1^1)$ with $\psi^* \neq \psi^1$. Teicher (1963) showed that mixtures of normal distributions are identifiable in the sense that if there exists

some w_0 such that for all w_1 we have

$$g\left(w_1 \mid W_0 = w_0; \psi^*\right) = g\left(w_1 \mid W_0 = w_0; \psi^1\right) \text{ then either} \tag{3.32}$$

$$w_0\beta_1^* = w_0\beta_1^1, w_0\beta_0^* = w_0\beta_0^1, \phi_1^* = \phi_1^1, \phi_0^* = \phi_0^1, \text{ and } \gamma_0^* + w_0\gamma_1^* = \gamma_0^1 + w_0\gamma_1^1 \tag{3.33}$$

or

$$w_0\beta_1^* = w_0\beta_0^1, w_0\beta_0^* = w_0\beta_1^1, \phi_1^* = \phi_0^1, \phi_0^* = \phi_1^1, \text{ and } \gamma_0^* + w_0\gamma_1^* = -\gamma_0^1 - w_0\gamma_1^1. \tag{3.34}$$

The relations in (3.32) would hold if $\psi^1 = \psi^*$. The situation in (3.34) corresponds to the idea that if $g(x; p, \alpha_1, \alpha_0)$ is some generic mixture of parametric densities e.g. $g(x; p, \alpha_1, \alpha_0) = p * f(x; \alpha_1) + (1-p) * f(x; \alpha_0)$ then one can 'switch the labels' or permute the component densities and obtain $g(x; p, \alpha_1, \alpha_0) \stackrel{x}{=} g(x; (1-p), \alpha_0, \alpha_1)$. The relations in (3.34) would hold under such label-switching (here multiplying $\gamma_0^* + w_0\gamma_1^*$ by -1 is analogous to substituting $1-p$ for p). To rule out such label-switching let K be a bounded, closed ball in $\mathbb{R}^6$ such that $\psi^* \in K$ but $\psi^{**} \stackrel{\triangle}{=} (\beta_0^*, \beta_1^*, \phi_0^*, \phi_1^*, -\gamma_0^*, -\gamma_1^*) \notin K$. This set K is the one mentioned in our statement of Condition 3.1.C.

Lemma 3.10. *Let $\psi \in K$ with $\psi \neq \psi^*$. Then there exist a w_0 and a w_1 such that $g\left(w_1 \mid W_0 = w_0; \psi\right) \neq g\left(w_1 \mid W_0 = w_0; \psi^*\right)$.*

Proof. Suppose the result is not true and there exists some ψ with components $(\beta_1, \beta_0, \phi_1, \phi_0, \gamma_0, \gamma_1)$ for which $g\left(w_1 \mid W_0 = w_0; \psi\right) = g\left(w_1 \mid W_0 = w_0; \psi^*\right)$ for all w_1 and w_0. Then for each w_0 one of the two relations above (equation (3.32) or (3.34)) must hold. Let us first examine what happens if we assume (3.32) is applicable. Then for $w_0 = 1$ we obtain

$$\beta_1^* = \beta_1, \beta_0^* = \beta_0, \phi_1^* = \phi_1, \phi_0^* = \phi_0, \gamma_0^* + \gamma_1^* = \gamma_0 + \gamma_1.$$

But the relation in (3.32) must also hold for $w_0 = 2$ which implies $\gamma_0^* + 2\gamma_1^* = \gamma_0 + 2\gamma_1$. The only way all these equalities can hold is if $\psi = \psi^*$. This contradicts our assumption that $\psi \neq \psi^*$. Now we consider what happens if (3.34) is applicable. Then for $w_0 = 1$ we obtain

$$\beta_1^* = \beta_0, \beta_0^* = \beta_1, \phi_1^* = \phi_0, \phi_0^* = \phi_1, \gamma_0^* + \gamma_1^* = -\gamma_0 - \gamma_1.$$

If we set $w_0 = 2$ then we obtain $\gamma_0^* + 2\gamma_1^* = -\gamma_0 - 2\gamma_1$. These equalities can only be met if $\psi = \psi^{**}$. But we defined K so that $\psi^{**} \notin K$. Thus we have shown there is no $\psi \in K$, $\psi \neq \psi^*$ such that $g(w_1 \mid W_0 = w_0; \psi) = g(w_1 \mid W_0 = w_0; \psi^*)$ for all w_1 and w_0. □

The next step in proving Condition 3.1.C is to apply the Kullback-Leibler information inequality to the conditional densities. Define

$$\begin{aligned} h(w_0; \psi) &\stackrel{\triangle}{=} \mathbb{E}[\log g(w_1 \mid W_0 = w_0; \psi) \mid W_0 = w_0] \\ &= \int_{\mathbb{R}} \log g(w_1 \mid W_0 = w_0; \psi)\, g(w_1 \mid W_0 = w_0; \psi^*)\, dw_1 \end{aligned}$$

This equality follows because $g(w_1 \mid W_0 = w_0; \psi^*)$ is the true conditional density of W_1 given $W_0 = w_0$. From this representation it is clear that by the Kullback-Leibler information inequality $h(w_0; \psi) \leq h(w_0; \psi^*)$ for all w_0. Now we want to show this inequality is strict for some w_0.

Lemma 3.11. *Let $\psi \in K$, $\psi \neq \psi^*$. Then there exists $w_0 \in \mathbb{R}$ such that $h(w_0; \psi) < h(w_0; \psi^*)$.*

Proof. Choose $\psi \in K$. By Lemma 3.10 there exists a w_0 and w_1 such that $g(w_1 \mid W_0 = w_0; \psi) \neq g(w_1 \mid W_0 = w_0; \psi^*)$. By inspection we see that $g(w_1 \mid W_0 = w_0; \psi)$ and $g(w_1 \mid W_0 = w_0; \psi^*)$ are continuous in w_1. This implies

that for this particular w_0 there exists a compact Borel-measurable set A with positive Lebesgue measure where $g(w_1 \mid W_0 = w_0; \psi) \neq g(w_1 \mid W_0 = w_0; \psi^*)$ for all $w_1 \in A$. Consequently the conditional distributions parameterized by ψ and ψ^* are different and an application of the Kullback-Leibler inequality yields the strict inequality. □

We conclude our proof by showing Lemma 3.11 is sufficient to show for all $\psi \in K, \psi \neq \psi^*$

$$\begin{aligned}\mathbb{E}\left[\log g(W_1 \mid W_0; \psi)\right] <& \mathbb{E}\left[\log g(W_1 \mid W_0; \psi^*)\right], \text{ or equivalently,} \\ \mathbb{E}\left[\log g(W; \psi)\right] <& \mathbb{E}\left[\log g(W; \psi^*)\right].\end{aligned}$$

Let $\psi \in K, \psi \neq \psi^*$. Then through the use of iterated expectations we have

$$\mathbb{E}\left[\log g(W_1 \mid W_0; \psi) - \log g(W_1 \mid W_0; \psi^*)\right] = \mathbb{E}\left[h(W_0; \psi) - h(W_0; \psi^*)\right]. \quad (3.35)$$

Lemma 3.11 shows us that for ψ there exists a w_0 such that $h(w_0; \psi) \neq h(w_0; \psi^*)$. We can find an integrable dominating function and apply the dominated convergence theorem to show that $h(w_0; \psi)$ and $h(w_0; \psi^*)$ are continuous in w_0 and hence there exists a compact Borel set A with positive measure such that $h(w_0; \psi) \neq h(w_0; \psi^*)$ for all $w_0 \in A$. The Kullback-Leibler inequality implies that $h(w_0; \psi) \leq h(w_0; \psi^*)$ for all $w_0 \in \mathbb{R}$. The existence of the set A means we can strengthen this statement to

$$\begin{aligned}\mathbb{E}\left[g(W; \psi)\right] - \mathbb{E}\left[g(W; \psi^*)\right] =& \mathbb{E}\left[h(W_0; \psi) - h(W_0; \psi^*)\right] \\ \leq& \int_A \left[h(w_0; \psi) - h(w_0; \psi^*)\right] d\mathbb{P}(w_0) < 0\end{aligned}$$

In the inequality above we have used the result that $h(w_0; \psi) \leq h(w_0; \psi^*)$ for $w_0 \notin A$ and thus our condition is proven.

3.2.4 Condition 3.1.D

This condition is most easily verified using derivatives of $\log g(W;\psi)$ and the mean value theorem. Let ρ be positive and ψ' be an arbitrary element of K. Consider $\psi \in B_\rho(\psi')$. Then by the mean value theorem

$$|\log g(W;\psi)| \leq |\log g(W;\psi')| + \left\| \left. \frac{\partial \log g(W;\psi)}{\partial \psi'} \right|_{\tilde{\psi}} \right\| \cdot \|\psi - \psi'\| \tag{3.36}$$

where $\tilde{\psi}$ lies on a chord between ψ and ψ' and $\|\cdot\|$ denotes the Euclidean vector norm in $\mathbb{R}^6$. Now suppose that for any ρ and ψ' there exists a $\mathbb{R}^6$ valued function $D(W;\rho,\psi') = (D(W;\rho,\psi')_1, \ldots, D(W;\rho,\psi')_6)'$ such that for all $\psi \in B_\rho(\psi')$

$$D(W;\rho,\psi')_i > \left| \left(\left. \frac{\partial \log g(W;\psi)}{\partial \psi'} \right|_{\tilde{\psi}} \right)_i \right| \text{ and}$$

$$\mathbb{E}[D(W;\rho,\psi')_i] < \infty \text{ for } i \in \{1,\ldots,6\}$$

Then we may deduce from (3.36) that

$$\mathbb{E}[\sup_{\psi \in B_\rho(\psi')} |\log g(W;\psi)|] \leq \mathbb{E}[|\log g(W;\psi')|] + \|\mathbb{E}[D(W;\rho,\psi')]\| \cdot \rho < \infty$$

for all $\psi \in B_\rho(\psi')$. To find such a function, $D(\cdot)$, we examine one of the derivatives – the others follow the same pattern. Consider

$$\log g(W;\psi) = p(W_0;\gamma)\, f(W_1 \mid W_0, \beta_1, \phi_1) + (1 - p(W_0;\gamma))\, f(W_1 \mid W_0, \beta_0, \phi_0)$$

where f is as in (3.27). Then (3.37)

$$\frac{\partial \log g(W;\psi)}{\partial \beta_1} = \frac{f(W_1 \mid W_0, \beta_1, \phi_1) p(W_0;\gamma)}{g(W;\psi)} \frac{(W_1 - W_0\beta_1)W_0}{\phi_1} \tag{3.38}$$

$$< \frac{|W_1 W_0| + |W_0^2|}{\phi_{min}} \tag{3.39}$$

because $\frac{f(W_1|W_0,\beta_1,\phi_1)p(W_0;\gamma)}{g(W;\psi)} < 1$ and $|\beta_1| < 1$. Setting $D(W;\rho,\psi')$ equal to the right-hand side of (3.39) satisfies our requirements for $D(\cdot)$ (at least for the

derivative with respect to β_1 – the other derivatives are similarly bound). Thus $\mathbb{E}[\log g^*(W;\psi',\rho)] < \infty$ for all $\psi' \in K$ and ρ small enough such that $B_\rho(\psi') \subset K$. Also, it should be clear that continuity of $\mathbb{E}[\log g(W;\psi)]$ and the bounding of the derivatives as above is sufficient to show

$$\lim_{\rho \to 0} \mathbb{E}[\log g^*(W;\psi',\rho)] = \mathbb{E}[\log g(W;\psi')]$$

for any $\psi' \in K$ (the dominated convergence theorem is used).

With this we have validated Condition 3.1.D for AR(1) logistic mixtures. Consequently, we have demonstrated all our consistency conditions are met in the case of logistic mixtures of AR(1) models and may conclude there exists a sequence of local maximum (partial) likelihood estimators $\hat{\psi}_T$ such that $\hat{\psi}_T \xrightarrow{a.s.} \psi^*$.

3.3 Asymptotic Normality for General Logistic Mixture

Some of the convenient properties of GLM models include the log concavity of the likelihood with respect to parameter space, and when a canonical link is used, an equivalence between second derivatives of the log likelihood and -1 times the Fisher information when the model's regressors are non-stochastic – see Appendix A.1 in Fahrmeir and Tutz (1994). Unfortunately, adding the complexity of switching destroys these ideals and complicates questions about large sample properties (see Fahrmeir and Kaufmann (1985) for a good treatment of asymptotic theory for standard models). What follows is similar to methods used by Wong (1986) and Cramér (1946). Our general model of a logistic mixture

uses the same notation as described in Section 3.1:

$$
\begin{aligned}
f(y_t \mid X_t; \beta, \phi) =& \exp\left(\frac{y_t X_t'\beta - b\left(X_t'\beta\right)}{\phi} + c(y_t, \phi)\right), \\
\mathbb{P}\left[I_t = 1 \mid Z_t; \gamma\right] =& \frac{\exp\left(Z_t'\gamma\right)}{1 + \exp\left(Z_t'\gamma\right)}, \\
g(y_t \mid \mathcal{G}_{t-1}; \psi) =& \mathbb{P}\left[I_t = 1 \mid Z_t; \gamma\right] \cdot f(y_t \mid X_{t1}; \beta_1, \phi_1) + \\
& \left(1 - \mathbb{P}\left[I_t = 1 \mid Z_t; \gamma\right]\right) \cdot f(y_t \mid X_{t0}; \beta_0, \phi_0), \\
\text{where } \mathcal{G}_{t-1} =& \sigma(X_{t1}, X_{t0}, Z_t).
\end{aligned}
$$

We assume $g(y_t \mid \mathcal{G}_{t-1}; \psi)$ is three times continuously differentiable with respect to ψ and that the conditional density of Y_t given X_{t1}, X_{t0}, and Z_t is $g(y_t \mid \mathcal{G}_{t-1}; \psi^*)$ for some ψ^* in the interior of Ψ, a subset of $\mathbb{R}^v$. (Here $v = 3q + 2$ where q is the common dimension of $X_{t1}, X_{t0},$ and Z_t (the other two dimensions are for ϕ_1 and ϕ_0 if necessary.)

Other notation we use in this section includes

$$
\begin{aligned}
\zeta_t(\psi) =& \frac{\partial \log g(y_t \mid \mathcal{G}_{t-1}; \psi)}{\partial \psi} \text{ a } v \times 1 \text{ vector with elements } \frac{\partial \log g(y_t \mid \mathcal{G}_{t-1}; \psi)}{\partial \psi_r} \\
& \text{for } r \in 1, ..., v, \\
h_t(\psi) =& \text{ a } v \times v \text{ matrix with the (r,s) element given by } \frac{\partial^2 \log g(y_t \mid \mathcal{G}_{t-1}; \psi)}{\partial \psi_r \partial \psi_s}. \\
S_T(\psi) =& \sum_{t=1}^{T} \zeta_t(\psi), \text{ and } H_T(\psi) = \sum_{t=1}^{T} h_t(\psi).
\end{aligned}
$$

Condition 3.12.

3.12.A *Y_t, X_{t1}, X_{t0}, Z_t are strictly stationary and ergodic with the random vector W having the common joint distribution. As in the previous chapter, we mean if $h(\cdot)$ is an integrable function of W then $\frac{1}{T}\sum h(Y_t, X_{t1}, X_{t0}, Z_t) \xrightarrow{a.s.} \mathbb{E}[h(W)]$. We will sometimes write $W_t = (Y_t, X_{t1}, X_{t0}, Z_t)$ and partition*

W as $W = (W_Y, W_{X1}, W_{X0}, W_Z)$ when we need to refer to its components. As in Section 3.1 we define $g(W;\psi) = g(W_Y \mid W_{X1}, W_{X0}, W_Z; \psi)$.

3.12.B *There exist integrable functions, $F_1(W)$ and $F_2(W)$ such that for all $r, s \in \{1, 2, ..., v\}$,*

$$F_1(W) > \left|\frac{\partial g(W;\psi)}{\partial \psi_r}\right|, F_2(W) > \left|\frac{\partial^2 g(W;\psi)}{\partial \psi_r \partial \psi_s}\right|, \text{ and}$$

$$\mathbb{E}[F_1(W) \mid W_{X1}, W_{X0}, W_Z] < \infty, \mathbb{E}[F_2(W) \mid W_{X1}, W_{X0}, W_Z] < \infty. \text{ Also,}$$

$$\mathbb{E}\left[\left(\frac{\partial \log g(W;\psi)}{\partial \psi_r}\right)^2\right] < \infty \text{ and } \mathbb{E}\left[\frac{\partial^2 \log g(W;\psi)}{\partial \psi_r \partial \psi_s}\right] < \infty.$$

All derivatives above are understood to be evaluated at ψ^*.

3.12.C $\frac{1}{T}H_T(\psi) \xrightarrow{a.s.} \mathbb{E}\left[\frac{\partial^2 \log g(W;\psi)}{\partial \psi \partial \psi'}\right]$ *uniformly for all ψ in some closed ball $\bar{\mathcal{O}}$ containing ψ^*. Furthermore, $\mathbb{E}\left[\frac{\partial^2 \log g(W;\psi)}{\partial \psi \partial \psi'}\right]$ is a continuous function of ψ on $\bar{\mathcal{O}}$.*

Again, as in the previous chapter, these conditions are relatively mild (with the exception of Condition 3.12.A) and are consequences of $\log g(W;\psi)$ and $g(W;\psi)$ having enough derivatives that can be uniformly bounded so the dominated convergence theorem may be applied. This will be demonstrated for a logistic mixture of Gaussian AR(1) processes in the next section.

Remark: It is worthwhile to mention a distinction between Conditions 3.1.A and 3.12.A. While both posit the existence of a stationary distribution, Condition 3.12.A assumes $\{Y_t, X_{t1}, X_{t0}, Z_t\}$ have this marginal distribution while Condition 3.1.A only assumes convergence to the stationary distribution. It is not strictly necessary to make this more stringent assumption (one could substitute uniform integrability conditions) but it does make one of the proofs in this chapter a little more straightforward.

Let us define the matrices

$$Q = -\mathbb{E}\left[\left.\frac{\partial^2 \log g(W;\psi)}{\partial\psi\partial\psi'}\right|_{\psi^*}\right] \text{ and } Q_1 = \mathbb{E}\left[\left.\frac{\partial \log g(W;\psi)}{\partial\psi'}\frac{\partial \log g(W;\psi)}{\partial\psi}\right|_{\psi^*}\right]. \tag{3.40}$$

Remark: We assume that ψ^* is such that Q is invertible. As will be discussed at length in Chapter 5, this excludes the case that ψ^* correspond to a degenerate mixture, i.e. $\beta_1^* = \beta_0^*, X_{t1} = X_{t0}$, and $\phi_1^* = \phi_0^*$. For the remainder of this chapter we assume that ψ^* does not correspond to a degenerate mixture and Q is invertible.

Theorem 3.13. *Let $\hat{\psi}_T$ be a consistent sequence of local maximum (partial) likelihood estimates of ψ^* as described in* 3.1. *Then under Conditions* 3.12.A–3.12.C *we have that*

$$\sqrt{T}\left(\hat{\psi}_T - \psi^*\right) \xrightarrow{\mathcal{D}} N\left(\underline{0}, Q^{-1}\right). \tag{3.41}$$

Furthermore, $Q = Q_1$.

Proof. We begin by expanding a Taylor series about ψ^* and obtain:

$$S_T(\hat{\psi}_T) = 0 = S_T(\psi^*) + H_T(\bar{\psi}_T)(\hat{\psi}_T - \psi^*) \tag{3.42}$$

where $\bar{\psi}_T$ lies on the chord between $\hat{\psi}_T$ and ψ^*. Dividing by T we see

$$0 = \frac{S_T(\psi^*)}{T} + \left[\frac{H_T(\psi^*)}{T} + \frac{\left(H_T(\bar{\psi}_T) - H_T(\psi^*)\right)}{T}\right](\hat{\psi}_T - \psi^*). \tag{3.43}$$

At this point we want to show the expression in brackets $\xrightarrow{P} -Q$. From Condition 3.12.C we need only demonstrate $\frac{1}{T}\left(H_T(\bar{\psi}_T) - H_T(\psi^*)\right) \xrightarrow{P} \underline{0}$ (a matrix of zeros). To prove this let $\epsilon, \delta > 0$ be given. We want to find a T_0 such that

$$\mathbb{P}\left[\frac{1}{T}\left\|H_T(\bar{\psi}_T) - H_T(\psi^*)\right\| > \delta\right] < \epsilon$$

for all $T > T_0$ where $\|\cdot\|$ is a Euclidean matrix norm. Now,

$$\mathbb{P}\left[\left\|\frac{H_T(\bar{\psi}_T) - H_T(\psi^*)}{T}\right\| > \delta\right] \tag{3.44}$$

$$\leq \mathbb{P}\left[\left\|\frac{H_T(\bar{\psi}_T)}{T} - \mathbb{E}\left[h\left(W;\psi\right)\right]\big|_{\bar{\psi}_T}\right\| > \delta/3\right] \tag{3.45}$$

$$+\mathbb{P}\left[\left\|\mathbb{E}\left[h\left(W;\psi\right)\right]\big|_{\bar{\psi}_T} - \mathbb{E}\left[h\left(W;\psi\right)\right]\big|_{\psi^*}\right\| > \delta/3\right] \tag{3.46}$$

$$+\mathbb{P}\left[\left\|\mathbb{E}\left[h\left(W;\psi\right)\right]\big|_{\psi^*} - \frac{H_T(\psi^*)}{T}\right\| > \delta/3\right]. \tag{3.47}$$

Now, without loss of generality we may take $\bar{\mathcal{O}}$ small enough such that

$$\sup_{\psi',\psi''\in\bar{\mathcal{O}}}\left\|\mathbb{E}\left[h\left(W;\psi\right)\right]\big|_{\psi'} - \mathbb{E}\left[h\left(W;\psi\right)\right]\big|_{\psi''}\right\| < \delta/3. \tag{3.48}$$

Also, because $\hat{\psi}_T \xrightarrow{P} \psi^*$ there exists a T_1 such that

$$\mathbb{P}\left[\bar{\psi}_T \notin \bar{\mathcal{O}}\right] < \epsilon/3 \text{ for } T > T_1.$$

Next, the uniform convergence condition (3.12.C) allows us to say there exists a T_2 such that

$$\mathbb{P}\left[\sup_{\psi\in\bar{\mathcal{O}}}\left\|\frac{H_T(\psi)}{T} - \mathbb{E}\left[h\left(W;\psi\right)\right]\big|_{\psi}\right\| > \delta/3\right] < \epsilon/3 \text{ for } T > T_2.$$

Now, if we choose $T_0 > \max\{T_1, T_2\}$ and combine the results in equations (3.45) – (3.48) we see T_0 satisfies the desired property and we conclude the bracketed term in (3.43) $\xrightarrow{P} -Q$. After inverting this term and multiplying by $\sqrt{T}$ we find

$$\sqrt{T}\left(\hat{\psi}_T - \psi^*\right) = -\left[\frac{H_T(\psi^*)}{T} + \frac{\left(H_T(\bar{\psi}_T) - H_T(\psi^*)\right)}{T}\right]^{-1}\frac{S_T(\psi^*)}{\sqrt{T}}, \tag{3.49}$$

and so we need only show

$$\frac{S_T(\psi^*)}{\sqrt{T}} \xrightarrow{\mathcal{D}} N\left(\underline{0}, Q\right) \tag{3.50}$$

to prove the theorem. By construction (and Condition 3.12.B) $\{\zeta_t\}$ is a martingale difference sequence adapted to $\{\mathcal{G}_t\}$. To prove this, the critical point is to establish

$$\mathbb{E}\left[\left.\left.\frac{\partial \log g(y_t \mid \mathcal{G}_{t-1};\psi)}{\partial \psi'}\right|_{\psi^*}\right| \mathcal{G}_{t-1}\right] = \underline{0}.$$

$$\begin{aligned}
\text{But the left-hand side above } &= \int \left.\frac{\partial \log g(y_t \mid \mathcal{G}_{t-1};\psi)}{\partial \psi'}\right|_{\psi^*} g(y_t \mid \mathcal{G}_{t-1};\psi^*)dy_t \\
&= \int \left.\frac{\partial g(y_t \mid \mathcal{G}_{t-1};\gamma,\psi)}{\partial \psi'}\right|_{\psi^*} dy_t \\
&= \left.\frac{\partial}{\partial \psi'}\left(\int g(y_t \mid \mathcal{G}_{t-1};\gamma,\psi)\right)\right|_{\psi^*} = \underline{0}.
\end{aligned}$$

The interchange of differentiation and integration is justified by the properties of the $F_1(W)$ function of Condition 3.12.B. Since we have shown $\{\zeta_t\}$ is a martingale difference sequence then the same is true for $\{c'\zeta_t\}$ where c is an arbitrary non-random element in $\mathbb{R}^v$. At this point we introduce a martingale central limit theorem we will need here and elsewhere. The theorem is drawn from McLeish (1974).

Theorem 3.14. *Let $D_{T,t}$ be a real-valued triangular martingale difference array on $(\Omega, \mathcal{F}, \mathbb{P}), t = 1...T, T \in \mathbb{N}$ such that*

$$\text{for all } \epsilon > 0 \lim_{T\to\infty} \sum_{t=1}^{T} \int_{|D_{T,t}|>\epsilon} D_{T,t}^2 d\mathbb{P} = 0, \text{ and} \tag{3.51}$$

$$\sum_{t=1}^{T} D_{T,t}^2 \xrightarrow{P} 1. \tag{3.52}$$

Then $\sum_{t=1}^{T} D_{T,t} \xrightarrow{\mathcal{D}} N(0,1)$.

The proof comes from McLeish's second theorem and the discussion following it.

To apply this theorem in our case we let ϵ be given and define

$$m_t = m(Y_t, X_{t1}, X_{t0}, Z_t) \stackrel{\triangle}{=} c'\zeta_t(\psi^*) \text{ and} \tag{3.53}$$

$$D_{T,t} \stackrel{\triangle}{=} \frac{m(Y_t, X_{t1}, X_{t0}, Z_t)}{\sqrt{Tv_c}} \text{ where } v_c = c'Qc. \tag{3.54}$$

Then

$$\lim_{T\to\infty} \sum_{t=1}^{T} \int_{|D_{T,t}|>\epsilon} D_{T,t}^2 d\mathbb{P} = \lim_{T\to\infty} \frac{1}{Tv_c} \sum_{t=1}^{T} \int_{m_t^2 > \epsilon^2 T v_c} m_t^2 \ d\mathbb{P}.$$

To show this limit is zero we first prove for all τ we have

$$\lim_{T\to\infty} \frac{1}{Tv_c} \sum_{t=1}^{T} \int_{m_t^2>\epsilon^2 T v_c} m_t^2 d\mathbb{P} \leq \lim_{T\to\infty} \frac{1}{Tv_c} \sum_{t=1}^{T} \int_{m_t^2>\tau} m_t^2 d\mathbb{P}. \tag{3.55}$$

To see this let $T_\tau = \sup\{T : \epsilon^2 T v_c \leq \tau\}$. Then

$$\lim_{T\to\infty} \frac{1}{Tv_c} \sum_{t=1}^{T} \int_{m_t^2>\epsilon^2 T v_c} m_t^2 d\mathbb{P} \tag{3.56}$$

$$= \lim_{T\to\infty} \frac{1}{Tv_c} \left(\sum_{t=1}^{T_\tau} \int_{m_t^2>\epsilon^2 T v_c} m_t^2 d\mathbb{P} + \sum_{t=T_\tau+1}^{T} \int_{m_t^2>\epsilon^2 T v_c} m_t^2 d\mathbb{P} \right)$$

$$= \lim_{T\to\infty} \frac{1}{Tv_c} \sum_{t=T_\tau+1}^{T} \int_{m_t^2>\epsilon^2 T v_c} m_t^2 d\mathbb{P} \text{ (since the } \lim_{T\to\infty} \frac{1}{Tv_c} \sum_{t=1}^{T_\tau} \text{ term } = 0)$$

$$\leq \lim_{T\to\infty} \frac{1}{Tv_c} \sum_{t=T_\tau+1}^{T} \int_{m_t^2>\tau} m_t^2 d\mathbb{P} \tag{3.57}$$

From our Condition 3.12.A we know the m_t variables are identically distributed with the variable $m(W)$ having the common distribution. So

$$\lim_{T\to\infty} \frac{1}{Tv_c} \sum_{t=T_\tau+1}^{T} \int_{m_t^2>\tau} m_t^2 d\mathbb{P} = \frac{1}{v_c} \mathbb{E}\left[m(W)^2 I_{[m(W)^2>\tau]} \right] \tag{3.58}$$

if the expectation on the right-hand side exists. But we know this expectation does exist because from Conditions 3.12.A and 3.12.B

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} m_t^2 \xrightarrow{P} c'Q_1c$$

and therefore

$$\mathbb{E}\left[m(W)^2\right] = c'Q_1c \text{ and } \mathbb{E}\left[m(W)^2 I_{[m(W)^2>\tau]}\right] \text{ exists.}$$

Putting together our results from equations (3.55) and (3.58) we obtain

$$\lim_{T\to\infty} \frac{1}{Tv_c} \sum_{t=1}^{T} \int_{m_t^2>\epsilon^2 T v_c} m_t^2 d\mathbb{P} \leq \mathbb{E}\left[m(W)^2 I_{[m(W)^2>\tau]}\right]$$

where the right-hand side can be made as small as desired by choosing τ large enough. Thus

$$\lim_{T\to\infty} \sum_{t=1}^{T} \int_{|D_{T,t}|>\epsilon} D_{T,t}^2 d\mathbb{P} = \lim_{T\to\infty} \frac{1}{Tv_c} \sum_{t=1}^{T} \int_{m_t^2>\epsilon^2 T v_c} m_t^2 \ d\mathbb{P} = 0,$$

and we see the Lindeberg condition (3.51) holds. To check the second condition of the theorem (3.52) we note that

$$\sum_{t=1}^{T} D_{T,t}^2 = \frac{1}{v_c}\frac{1}{T}\sum_{t=1}^{T} m_t^2 \xrightarrow{a.s.} \frac{1}{v_c} c'Q_1c = \frac{c'Q_1c}{c'Qc}.$$

To show this is 1 we examine the $(r,s)th$ elements of Q_1 and Q (we want to show $Q_1 = Q$). We now demonstrate

$$\mathbb{E}\left[\frac{\partial \log g(W_Y \mid W_{X1}, W_{X0}, W_Z; \psi)}{\partial \psi_r} \frac{\partial \log g(W_Y \mid W_{X1}, W_{X0}, W_Z; \psi)}{\partial \psi_s}\right] = -\mathbb{E}\left[\frac{\partial^2 \log g(W_Y \mid W_{X1}, W_{X0}, W_Z; \psi)}{\partial \psi_r \partial \psi_s}\right]. \tag{3.59}$$

where the derivatives are evaluated at ψ^*. This is straightforward because from Condition 3.12.B we know the result holds for conditional expectation, i.e.

$$\mathbb{E}\left[\frac{\partial \log g(W_Y|W_{X1},W_{X0},W_Z;\psi)}{\partial \psi_r} \frac{\partial \log g(W_Y|W_{X1},W_{X0},W_Z;\psi)}{\partial \psi_s} \middle| W_{X1}, W_{X0}, W_Z\right] = -\mathbb{E}\left[\frac{\partial^2 \log g(W_Y|W_{X1},W_{X0},W_Z;\psi)}{\partial \psi_r \partial \psi_s} \middle| W_{X1}, W_{X0}, W_Z\right].$$

The two sides of the equation are functions of $W_{X1}, W_{X0},$ and W_Z that are equal almost everywhere, and thus the expectations of the functions are equal. So by

taking expectations of these conditional expectations we get the result in (3.59). Thus we see $Q_1 = Q$ and consequently

$$\sum_{t=1}^{T} D_{T,t}^2 = 1.$$

Now the two conditions of McLeish's theorem hold so we have shown

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} c'\zeta_t \xrightarrow{\mathcal{D}} N(0, c'Qc)$$

for arbitrary $c \in \mathbb{R}^q$. By the Cramér-Wold Theorem we may conclude

$$\frac{S_T(\psi^*)}{\sqrt{T}} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \zeta_t \xrightarrow{\mathcal{D}} N(\underline{0}, Q)$$

and our demonstration of asymptotic normality is complete. □

3.4 Asymptotic Normality for Logistic Mixtures of Gaussian AR(1) Processes

As in Section 3.2 we examine the conditions just presented and try to verify them for our logistic mixture of AR(1) models. The model specification is the same as that in Section 3.2. In this section we will freely use our results from Section 3.2 which demonstrated the following:

- Let $\underline{W}_t = (Y_t, Y_{t-1})$. Then there exists a random vector W with distribution π_W such that if $h(\cdot)$ is an integrable function of W then

  $$\frac{1}{T} \sum h(W_t) \xrightarrow{a.s.} \mathbb{E}\left[h(W)\right].$$

  The components of W we will denote by W_1 and W_0. Both W_1 and W_0 have a common marginal distribution π_Y.

- For Y a random variable with distribution π_Y we have $\mathbb{E}[\exp(sY)] < \infty$ for all $s \in \mathbb{R}$.

We now set about examining the three components in Condition 3.12.

3.4.1 Condition 3.12.A

The existence of an appropriate stationary distribution was established in Section 3.2. If we additionally assume (Y_0, Y_{-1}) have π_W as an initial distribution then all the (Y_t, Y_{t-1}) will be identically distributed as π_W. Alternatively, one could imagine our time series has been running long enough so that we may consider our sample to be approximately identically distributed according to the stationary distribution.

3.4.2 Condition 3.12.B

There are two types of expectations we want to check. The first type concerns showing the conditional expectation of derivatives of $g(\cdot)$ is finite. Though there are several derivatives to check we present analysis for only one; the others follow a similar pattern. As our example we consider

$$\begin{aligned}\left.\frac{\partial^2 g(Y_t \mid Y_{t-1};\psi)}{\partial\gamma_1\partial\phi_0}\right|_{\psi^*} &= -\frac{\exp(\gamma_0^* + Y_{t-1}\gamma_1^*)Y_{t-1}}{(1+\exp(\gamma_0^* + Y_{t-1}\gamma_1^*))^2}\frac{1}{2\phi_0^*}\left(\frac{(Y_t - Y_{t-1}\beta_0^*)^2}{\phi_0^*} - 1\right)\times \\ &\qquad \exp\left(-\frac{(Y_t - Y_{t-1}\beta_0^*)^2}{2\phi_0^*} - \frac{1}{2}\ln 2\pi\phi_0^*\right) \\ &< Y_{t-1}\frac{1}{2\phi_{min}}\left(\frac{Y_t^2 + Y_{t-1}^2 + 2\left|Y_tY_{t-1}\right|}{\phi_{min}} + 1\right)\frac{1}{\sqrt{2\pi\,\phi_{min}}}\end{aligned} \tag{3.60}$$

and wish to show the $\mathbb{E}[RHS \text{ of } (3.60) \mid Y_{t-1}] < \infty$. For k a non-negative integer we have

$$\mathbb{E}\left[|Y_t|^k \mid Y_{t-1}\right] < \int_{\mathbb{R}} |Y_t|^k \exp\left(-\frac{(Y_t - Y_{t-1}\beta_0^*)^2}{2\phi_0^*} - \frac{1}{2}\ln 2\pi\phi_0^*\right) dY_t + \tag{3.61}$$

$$\int_{\mathbb{R}} |Y_t|^k \exp\left(-\frac{(Y_t - Y_{t-1}\beta_1^*)^2}{2\phi_1^*} - \frac{1}{2}\ln 2\pi\phi_1^*\right) dY_t \tag{3.62}$$

which is clearly finite since a normally distributed r.v. has all finite moments. This fact that $\mathbb{E}\left[|Y_t|^k \mid Y_{t-1}\right] < \infty$ for all k is sufficient to show the right-hand side of (3.60) is conditionally integrable.

We also wish to show that unconditional expectation of derivatives of $\log g(\cdot)$ are integrable. As an example we consider

$$\left.\frac{\partial^2 \log g(Y_t \mid Y_{t-1};\psi)}{\partial\phi_1\partial\beta_0}\right|_{\psi^*} = -\frac{(1-p_t(\gamma^*))f(Y_t \mid Y_{t-1};\beta_0^*,\phi_0^*)}{g(Y_t \mid Y_{t-1};\psi^*)}\frac{(Y_t - Y_{t-1}\beta_0^*)}{\phi_0^*}Y_{t-1} \times$$
$$\frac{p_t(\gamma^*)f(Y_t \mid Y_{t-1};\beta_1^*,\phi_1^*)}{g(Y_t \mid Y_{t-1};\psi^*)}\frac{1}{2\phi_1^*} \times$$
$$\left(\frac{(Y_t - Y_{t-1}\beta_1^*)^2}{\phi_1^*} - 1\right) \text{ where}$$
$$p_t(\gamma^*) = \frac{\exp(\gamma_0^* + Y_{t-1}\gamma_1^*)}{1+\exp(\gamma_0^* + Y_{t-1}\gamma_1^*)} \text{ and}$$
$$f(Y_t \mid Y_{t-1};\beta,\phi) = \exp\left(-\frac{(Y_t - Y_{t-1}\beta)^2}{2\phi} - \frac{1}{2}\ln 2\pi\phi\right).$$

Because $\frac{p_t(\gamma^*)f(Y_t|Y_{t-1};\beta_1^*,\phi_1^*)}{g(Y_t|Y_{t-1};\psi^*)} < 1$ and $\frac{(1-p_t(\gamma^*))f(Y_t|Y_{t-1};\beta_0^*,\phi_0^*)}{g(Y_t|Y_{t-1};\psi^*)} < 1$ we have

$$\left|\left.\frac{\partial^2 \log g(Y_t \mid Y_{t-1};\psi)}{\partial\phi_1\partial\beta_0}\right|_{\psi^*}\right| < \left|\frac{(Y_t - Y_{t-1}\beta_0^*)}{\phi_0^*}\frac{Y_{t-1}}{2\phi_1^*}\left(\frac{(Y_t - Y_{t-1}\beta_1^*)^2}{\phi_1^*} - 1\right)\right| \tag{3.63}$$

The function on the right-hand side of (3.63) is clearly integrable if Y_t and Y_{t-1} have enough finite moments. This is certainly the case as they are identically distributed with a common distribution that has a finite moment generating function over the real line. As the other derivatives are handled the same way we see that the mild moment conditions of 3.12.B are met.

3.4.3 Condition 3.12.C

The proof of this condition requires a theorem for uniform convergence. We will digress from this discussion to examine such a theorem.

A Uniform Convergence Theorem

Here we present the primary theorem we will use throughout to demonstrate a uniform law of large numbers over a compact parameter set. The theorem is an adaptation of one presented in Andrews (1987).

To introduce notation let $\{W_t\}$ be a sequence of $\mathbb{R}^p$-valued random variables on a probability space $(\Omega, \mathcal{F}, P)$. We assume that the W_t's are asymptotically stationary and ergodic in the sense described in Condition 3.1.A (the W_t's could also be considered identically distributed with the stationary distribution as in Condition 3.12.A). Here we let W denote the $\mathbb{R}^p$-valued vector with the stationary distribution, $\psi \in \Psi$ (a subset of $\mathbb{R}^v$) and $B(\psi, \rho) = \{\psi' \in \Psi : \ ||\psi' - \psi|| < \rho\}$. Further, suppose $q(W_t; \psi) \colon \mathbb{R}^p \times \Psi \to \mathbb{R}$ and $\mathbb{E}q(W; \psi) < \infty$ for all ψ . Define,

$$q^*(W_t; \psi, \rho) = \sup_{\psi'} \{q(W_t; \psi) \colon \psi' \in B(\psi, \rho)\} \tag{3.64}$$

$$q_*(W_t; \psi, \rho) = \inf_{\psi'} \{q(W_t; \psi) \colon \psi' \in B(\psi, \rho)\}. \tag{3.65}$$

The conditions for our theorem are:

3.15.A *The parameter space Ψ is compact.*

3.15.B *The rv's $q(W_t; \psi), q^*(W_t; \psi, \rho)$ and $q_*(W_t; \psi, \rho)$ satisfy pointwise strong laws of large numbers for sufficiently small ρ that may depend upon ψ.*

3.15.C *For all $\psi \in \Psi$,*

$$\lim_{\rho \to 0} \mathbb{E} \sup_{\psi' \in B(\psi, \rho)} |q(W; \psi') - q(W; \psi)| = 0.$$

Theorem 3.16. *Under Conditions* 3.15.A – 3.15.C *we have*

$$\sup_{\psi\in\Psi}\left|\frac{1}{T}\sum_{t=1}^{T}q(W_t;\psi)-\mathbb{E}q(W;\psi)\right|\to 0 \tag{3.66}$$

almost surely. In other words $\frac{1}{T}\sum_{t=1}^{T}q(W_t;\psi)$ *converges uniformly to* $\mathbb{E}q(W;\psi)$*, almost surely.*

Remark: If the pointwise convergence in Condition 3.15.B is weak instead of strong than the conclusions should be changed from 'almost surely' to 'in probability'.

Proof. Condition 3.15.C shows that given ϵ and $\psi\in\Psi$ there exists a $\rho(\psi)$ such that

$$\mathbb{E}q(W;\psi)-\epsilon\le\mathbb{E}q_*(W;\psi,\rho(\psi))\le\mathbb{E}q^*(W;\psi,\rho(\psi))\le\mathbb{E}q(W;\psi)+\epsilon. \tag{3.67}$$

For the fixed ϵ the collection $\{B(\psi,\rho(\psi))\}$ form an open cover of Ψ. Under Condition 3.15.A we obtain a finite subcover, $\{B(\psi_i,\rho(\psi_i)):i=1...L\}$. Consider a fixed i and the associated $B(\psi_i,\rho(\psi_i))$. For ψ in $B(\psi_i,\rho(\psi_i))$ we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}(q(W_t;\psi)-\mathbb{E}q(W;\psi))\le&\frac{1}{T}\sum_{t=1}^{T}(q^*(W_t;\psi_i,\rho(\psi_i))-\mathbb{E}q_*(W;\psi_i\,\rho(\psi_i)))\\ \le&\frac{1}{T}\sum_{t=1}^{T}(q^*(W_t;\psi_i,\rho(\psi_i))-\mathbb{E}q^*(W;\psi_i\,\rho(\psi_i)))+2\epsilon,\end{aligned}$$

and similarly,

$$\frac{1}{T}\sum_{t=1}^{T}(q(W_t;\psi)-\mathbb{E}q(W;\psi))\ge\frac{1}{T}\sum_{t=1}^{T}(q_*(W_t;\psi_i,\rho(\psi_i))-\mathbb{E}q_*(W;\psi_i,\rho(\psi_i)))-2\epsilon.$$

Now we consider arbitrary ψ in Ψ which must be in one of the $B(\psi_i,\rho(\psi_i))$ balls.

Then

$$\min_{1\le i\le L}\frac{1}{T}\sum_{t=1}^{T}(q_*(W_t;\psi_i,\rho(\psi_i))-\mathbb{E}q_*(W;\psi_i,\rho(\psi_i)))-2\epsilon \tag{3.68}$$

$$\le\frac{1}{T}\sum_{t=1}^{T}(q(W_t;\psi)-\mathbb{E}q(W;\psi)) \tag{3.69}$$

$$\le\max_{1\le i\le L}\frac{1}{T}\sum_{t=1}^{T}(q^*(W_t;\psi_i,\rho(\psi_i))-\mathbb{E}q^*(W;\psi_i\,\rho(\psi_i)))+2\epsilon. \tag{3.70}$$

With an application of Condition 3.15.B it is easy to see we may find a value T_0 such that for all $T>T_0$ the expressions in (3.68) and (3.70) are arbitrarily close to -2ϵ and 2ϵ, with arbitrarily high probability. This implies that the right side of (3.69) is arbitrarily close to $0\pm 2\epsilon$ with high probability for all ψ and $T>T_0$. Because ϵ and ψ are arbitrary the theorem is proven. □

We may supply relatively mild conditions to replace Conditions 3.15.B and 3.15.C.

Condition 3.17. *If $q(W;\psi)$ has continuous derivatives with respect to ψ and for ρ small enough there exists a function, $D(W;\psi,\rho)$ such that if $\left\|\tilde{\psi}-\psi\right\|<\rho$ then*

$$\left\|\left.\frac{\partial q(W;\psi)}{\partial\psi}\right|_{\tilde{\psi}}\right\|<D(W;\psi,\rho)\ \textit{and}\ \mathbb{E}D(W;\psi,\rho)<\infty.$$

To show Condition 3.17 implies 3.15.B we first show that $\mathbb{E}q^*(W;\psi,\rho)$ is finite. Let ϵ and ψ be given. Then from Condition 3.17 we note there exists a ρ such that $q^*(W;\psi,\rho)\le q(W;\psi')+\epsilon$ where, $\|\psi'-\psi\|<\rho$. Now

$$q(W;\psi')=q(W;\psi)+\left.\frac{\partial q(W;\psi)}{\partial\psi'}\right|_{\tilde{\psi}}(\psi'-\psi)\text{ where, }\tilde{\psi}\in B(\psi,\rho).$$

Taking expectations we obtain,

$$\begin{aligned}\mathbb{E}\,\|q^*(W;\psi,\rho)\| \leq & \mathbb{E}\,\|q(W;\psi')\| + \mathbb{E}\left\| \left.\frac{\partial q(W;\psi)}{\partial \psi'}\right|_{\tilde{\psi}} (\psi'-\psi)\right\| + \epsilon \\ \leq & \mathbb{E}\,\|q(W;\psi')\| + \mathbb{E}D(W;\psi,\rho)\cdot\rho + \epsilon < \infty \end{aligned} \tag{3.71}$$

So we have shown $\mathbb{E}q^*(W;\psi,\rho) < \infty$. From our condition of almost sure convergence we obtain

$$\frac{1}{T}\sum_{t=1}^{T} q^*(W_t;\psi,\rho) \xrightarrow{a.s.} \mathbb{E}q^*(W;\psi,\rho), \frac{1}{T}\sum_{t=1}^{T} q_*(W_t;\psi,\rho) \xrightarrow{a.s.} \mathbb{E}q_*(W;\psi,\rho), \text{ and}$$
$$\frac{1}{T}\sum_{t=1}^{T} q(W_t;\psi) \xrightarrow{a.s.} \mathbb{E}q(W;\psi).$$

To show Condition 3.17 implies 3.15.C we see that

$$\begin{aligned}\sup_{\psi'\in B(\psi,\rho)} \|q(W;\psi') - q(W;\psi)\| \leq & \sup_{\tilde{\psi}\in B(\psi,\rho)} \left\| \left.\frac{\partial q(W;\psi)}{\partial \psi'}\right|_{\tilde{\psi}} \right\| \cdot \rho \\ \leq & D(W;\psi,\rho)\cdot\rho.\end{aligned}$$

Taking expectations and limits as $\rho \to 0$ shows that Condition 3.15.C follows. Thus we have proven the following:

Theorem 3.18. *Under Conditions* 3.15.A *and* 3.17 *we have*

$$\sup_{\psi\in\Psi} \left\| \frac{1}{T}\sum_{t=1}^{T} q(W_t;\psi) - \mathbb{E}q(W;\psi)\right\| \xrightarrow{a.s.} 0.$$

Now we return to our example of logistic mixtures. To show uniform convergence of $\frac{1}{T}\sum h_t(\psi)$ we successively apply this theorem to the real valued elements of $h_t(\psi^*)$, i.e. we set $q(W_t;\psi) = \frac{\partial^2 \log g(Y_t\,|Y_{t-1};\psi)}{\partial\psi_r\partial\psi_s}$ for $r,s \in \{1,...,v\}$. To apply our result we verify the two conditions in Theorem 3.18. As we previously explained in discussing Condition 2.3.B, it is clear $\mathbb{E}\,[q(W;\psi)] < \infty$ for each (r,s) pair. It is left to verify Condition 3.17, that is show that all third derivatives may be

bound by integrable functions. Rather than demonstrate the technique for each we will use one of the derivatives as an example of how to find the bounding function $D(W_t;\psi,\rho)$ in Condition 3.17. In this case it will turn out that the bounding function is independent of ψ (though it does depend on boundaries of Ψ.) All the other derivatives may be bounded in the same manner. To reduce notational clutter we will use the following abbreviations:

$$f_0 = f(Y_t \mid Y_{t-1}; \beta_0^*, \phi_0^*) \tag{3.72}$$

$$f_1 = f(Y_t \mid Y_{t-1}; \beta_1^*, \phi_1^*) \tag{3.73}$$

$$p = \frac{\exp(\gamma_0^* + Y_{t-1}\gamma_1^*)}{1+\exp(\gamma_0^* + Y_{t-1}\gamma_1^*)} \tag{3.74}$$

$$g = pf_1 + (1-p)f_0. \tag{3.75}$$

The example we work with is

$$\left.\frac{\partial^3 \log g(Y_t \mid Y_{t-1};\psi^*}{\partial\gamma_1\partial\phi_1\partial\beta_0}\right|_{\psi^*} = -\frac{Y_{t-1}}{\phi_1^*}\left(\frac{Y_t - Y_{t-1}\beta_0^*}{\phi_0^*}\right)\left(\frac{(Y_t - Y_{t-1}\beta_1^*)^2}{\phi_1^*} - 1\right) \times \frac{f1f0}{g^2}p(1-p)\left[1-2p-2\left(\frac{f1-f0}{g}\right)p(1-p)\right]$$

As before, the key to bounding the expression is to note that whenever there is a g term in the denominator, there is an offsetting $(1-p)f_0$, or pf_1 term in the numerator. In the case of $(1-p)f_0$ in the numerator we note that

$$\frac{(1-p)f_0}{g} = \frac{(1-p)f_0}{pf_1 + (1-p)f_0} < 1. \tag{3.76}$$

A similar inequality holds for pf_1/g. Using these facts and noting $0 < p, (1-p) < 1$ we can see

$$\left|\left.\frac{\partial^3 \log g}{\partial\gamma_1\partial\phi_1\partial\beta_0}\right|_{\psi^*}\right| < \frac{5\,|Y_{t-1}|}{\phi_{min}}\left(\frac{|Y_t - Y_{t-1}\beta_0^*|}{\phi_{min}}\right)\left(\frac{(Y_t - Y_{t-1}\beta_1^*)^2}{\phi_{min}} + 1\right).$$

Because $|\beta_1^*|, |\beta_0^*| < 1$ and both Y_t and Y_{t-1} have a finite moment generating function it is clear the right-hand side above is finite, and hence this third derivative is bounded; the other third derivative are bounded similarly. Thus we may find bounding integrable functions that allow us to apply Theorem 3.18 and the first part of Condition 3.12.C has been proven.

The second part of this condition requires us to show $\mathbb{E}[h(W;\psi)]$ is continuous in ψ. We could show this by applying the same techniques we used for the bounding the third derivatives of $\log g(W;\psi)$ to find integrable functions that bound the second derivatives. Once these bounding function are found we may appeal to the dominated convergence theorem and conclude the continuity condition holds. With this we conclude the proof of Condition 3.12.C

Conclusion

In this chapter we showed that the maximum (partial) likelihood estimates of a correctly specified logistic mixture model are both consistent and asymptotically normal if general conditions are met. We then showed these conditions are satisfied for the case of AR(1) mixtures.

Chapter 4

Numerical Results

In this chapter we present simulations and an example of fitting the model to rain rate data. The results from the first set of simulations adhere to our theory regarding consistency and the asymptotic variance structure we developed in the previous chapter. We see the β and ϕ parameters in the component densities are estimated relatively well though there appears to be bias in estimating γ. This bias is attenuated in large sample sizes. A second set of simulations suggests that a logistic mixture model may be a superior model in circumstances in which one might use a two regime threshold model. These simulations indicate the logistic mixtures yield robust estimates when the threshold variable in a threshold model is not directly observed but instead only a noisy approximation is available. The results from the threshold model are biased and not robust in the presence of noise. We close the chapter with an application of our model to rain rate data that suggests a logistic mixture with variable regime probabilities may be superior to a mixture model with constant regime probabilities.

4.1 Simulation I – Consistency and Asymptotic Variance

In our first simulation component densities (as presented in equation (2.2)) correspond to Gaussian AR(2) models with parameters chosen such that each regime is, by itself, a stationary process. The logistic regression model in the form of (2.1) contains only a slope and intercept parameter with the lagged value of the switching process as a covariate:

$$\begin{aligned}\mathbb{P}\left[I_t = 1 \mid \mathcal{G}_{t-1}\right] =& \mathbb{P}\left[I_t = 1 \mid Y_{t-1}\right] \\ &= \exp(-2 + y_{t-1})/(1 + \exp(-2 + y_{t-1})), \qquad (4.1)\end{aligned}$$

$$f(y_t \mid I_t = 1, \mathcal{G}_{t-1}) = N(.5y_{t-1} + .3y_{t-2}, .25), \text{ and} \qquad (4.2)$$

$$f(y_t \mid I_t = 0, \mathcal{G}_{t-1}) = N(-.5y_{t-1} - .15y_{t-2}, 1). \qquad (4.3)$$

A total of 200 simulations were analyzed. For each simulation a time series of 500 observations was generated according to the model described above. For each observation, the probability of drawing from the first regime was computed using (4.1) and past values of Y_t. A Bernoulli random variable with this mean was then generated. When that variable was 1 the observation for Y_t was drawn using the distribution given by (4.2); otherwise Y_t was obtained from the distribution in (4.3). Initial values, y_0 and y_{-1} were set equal to 0.

Each simulation produced not only an estimate of

$$\beta_1 = \begin{pmatrix} .5 \\ .3 \end{pmatrix}, \beta_0 = \begin{pmatrix} -.5 \\ -.15 \end{pmatrix}, \begin{pmatrix} \phi_1 \\ \phi_0 \end{pmatrix} = \begin{pmatrix} .25 \\ 1 \end{pmatrix}, \text{ and } \gamma = \begin{pmatrix} -2 \\ 1 \end{pmatrix},$$

but also generated estimates of the standard errors of $\hat{\beta}_0, \hat{\beta}_1$, and $\hat{\gamma}$. These estimated standard errors were obtained using the large sample results from the

Parameter	True Value	Average	$\bar{\hat{\sigma}}$(asymp.)	$\bar{\hat{\sigma}}$(simul.)
β_{11}	.5	.490	.0613	.0797
β_{12}	.3	.292	.0756	.0898
β_{01}	-.5	-.509	.0659	.0684
β_{02}	-.15	-.154	.0575	.0568
γ_0	-2	-2.14	.646	.835
γ_1	1	1.11	.370	.443
$\phi_1 = \sigma_1^2$	.25	.241	.0753	.088
$\phi_0 = \sigma_0^2$	1	.987	.0841	.105

Table 4.1: 200 Simulations of 500 Observations

previous section (see (3.40) with Q_1 estimated by the data using sample means in place of expectations and estimated parameters in place of their true values). These standard errors for each simulation were averaged, and the resulting means are included in the table below under the heading '$\bar{\hat{\sigma}}$(asymp.)'

A second method for estimating these standard errors is also presented. The 200 values of a given parameter, say $\{\beta_{11}^{(i)}\}, i = 1 \ldots 200$ form an i.i.d. sequence of random variables with a common standard error that may be estimated by

$$\frac{1}{199} \sum_{i=1}^{200} (\beta_{11}^{(i)} - \bar{\beta}_{11})^2. \tag{4.4}$$

Here $\beta_{11}^{(i)}$ is the estimate of β_{11} from the i^{th} simulation and $\bar{\beta}_{11}$ is the mean value for the 200 simulations. The standard errors obtained in this manner are included in the column headed '$\bar{\hat{\sigma}}$(simul.)'.

The model appears to have performed relatively well in generating point estimates of the regime specific parameters, β and ϕ, but performs worse when

estimating γ. To some extent the difficulty with estimating γ is to be expected given the small sample bias of logistic regression estimates in even optimal circumstances (see Chapter 7 of McCullagh (1987), and Chapter 15 of McCullagh and Nelder (1989)). These problems are probably exacerbated in the present context, when, instead of estimating γ from an observed sequence of ones and zeros (as is the case for standard logistic regression), we use probabilities associated with an unobserved process as a basis for estimating γ (i.e. the probabilities p_t^k in (2.33)).

In addition to the problems in the point estimates of γ, it is worth noting that most of the standard errors derived from asymptotic results of the preceding section underestimate the 'true' standard error (as derived from the empirical sample of 200 simulations). In each case corresponding to $\beta_{01}, \beta_{11}, \beta_{12}, \gamma_1,$ and γ_2, the asymptotic standard error is somewhat less than that derived from (4.4) (the standard errors for β_{02} are an exception). We are not sure why this discrepancy arises but it may occur because we lack enough observations to appropriately apply the asymptotic normality results. In order to check this hypothesis we performed a second set of 200 simulations with 2500, instead of 500 observations. Table 4.2 gives results from this second set of simulations that are consistent with our hypothesis that the percent differences should shrink (at least with respect to the β and ϕ terms). As expected, these estimates are better, but the asymptotic standard errors are still generally smaller than those derived from the sample. These simulations were run primarily to verify our programming and asymptotic results were accurate. They also indicate that the estimates of γ suffer bias and have relatively large standard errors. This suggests caution in placing much emphasis upon point estimates of γ.

Parameter	True Value	Average	$\bar{\hat{\sigma}}$(asymp.)	$\bar{\hat{\sigma}}$(simul.)
β_{11}	.5	.497	.0276	.0287
β_{12}	.3	.299	.0328	.0326
β_{01}	-.5	-.503	.0287	.0272
β_{02}	-.15	-.153	.0251	.0264
γ_0	-2	-2.03	.268	.308
γ_1	1	1.02	.151	.168
$\phi_1 = \sigma_1^2$	.25	.254	.0385	.0420
$\phi_0 = \sigma_0^2$	1	.995	.0330	.0346

Table 4.2: 200 Simulations of 2500 Observations

4.2 Simulation II - Comparison to Threshold Method

As a second example we consider simulations drawn from a threshold autoregressive model of the type developed by Tong (1983,1990). As before, we assume our observed process, Y_t, comes from a mixture of AR(2) processes, but in this case the choice of regime is determined by whether the lagged value of a second variable is above or below a threshold value: regime 1 is applicable if $X_{t-2} < .1$, regime 0 is relevant otherwise. As constructed here X_{t-2} is correlated with Y_{t-2}. In this simulation X_{t-2} is not observed directly – only its noisy proxy, Y_{t-2}, is available to the observer. The relationship between Y_{t-2} and X_{t-2} is

$$X_t = Y_t + .8 \cdot \eta_t$$

where η_t is i.i.d. $N(0,1)$ noise term that is independent of Y_s for $s \leq t$ and X_s for $s < t$. The regime indicator I_t is determined by

$$I_t = \begin{cases} 1 \text{ if } X_{t-2} < .1 \\ 0 \text{ otherwise,} \end{cases}$$

With this structure we now define how the Y_t process evolves:

$$Y_t = \begin{cases} .5Y_{t-1} + .2Y_{t-2} + .5 * \epsilon_t \text{ if } I_t = 1 \\ .1Y_{t-1} - .1Y_{t-2} + .7 * \epsilon_t \text{ if } I_t = 0, \end{cases}$$

where ϵ_t is i.i.d. $N(0,1)$, independent of η_s for all s.

Two methods were used to estimate

$$\beta_0 = \begin{pmatrix} .1 \\ -.1 \end{pmatrix}, \beta_1 = \begin{pmatrix} .5 \\ .2 \end{pmatrix}, \text{ and } \begin{pmatrix} \phi_1 \\ \phi_0 \end{pmatrix} = \begin{pmatrix} .25 \\ .49 \end{pmatrix}.$$

In applying the logistic mixture model we use 1 and Y_{t-2} as covariates in the Z_t vector of (2.1). It should be noted that the logistic mixture model is incorrectly specified. The true model uses a threshold to select the relevant regime, while this model posits the selection is made by the outcome of a Bernoulli random variable with mean $\exp(Z_t'\gamma)/(1 + \exp(Z_t'\gamma))$. In this sense the model is misspecified and the γ coefficients associated with the logistic regression do not have corresponding 'true' values. Nevertheless, as will be seen below, this model performs very well in estimating β_0, β_1, ϕ_0, and ϕ_1.

As an alternative to our model we present a threshold autoregressive (TAR) model that estimates $\beta_0, \beta_1, \phi_0, \phi_1$, and the unknown threshold, denoted by τ, in the following manner (the true value of τ is .1):

- For a fixed τ let $C_1^\tau = \{t : Y_{t-2} < \tau\}$ and $C_0^\tau = \{t : Y_{t-2} \geq \tau\}$. These sets partition the data into two groups. The data in C_1^τ has observations in

which Y_{t-2} is less than the hypothesized threshold, τ. The Y_{t-2} values of those observations in C_0^τ exceed τ.

- Using the C_0^τ set, estimate β_0 and ϕ_0 by conditional least squares using those values of Y_t, Y_{t-1}, and Y_{t-2} corresponding to observations in C_0^τ. Similarly, estimate β_1 and ϕ_1 from the observations corresponding to C_1^τ. Ordinary least squares is typically the estimation method.

- Add the residual sum of squares from the two regressions to obtain an overall sum of squares, $CSS(\tau)$. We follow this procedure for several choices of τ and choose $\hat{\tau}$ as the threshold that minimizes $CSS(\tau)$ The $\hat{\beta}_1, \hat{\beta}_0, \hat{\phi}_1$, and $\hat{\phi}_0$ estimates that are associated with $\hat{\tau}$ are the final estimates under the TAR model.

In this instance we produced 200 simulations where each simulation contained a time series of 500 observations. Each simulation was generated by the parameters indicated above. In each simulation $\hat{\tau}$ was chosen from the set $\{0, .025, .050, .075, .100, .125, .150, .175, .200\}$. In Table 4.3, estimates produced by the TAR method are included in the column headed 'TAR - Y_{t-2}'. The column headed 'LM - Y_{t-2}' displays estimates derived from our logistic mixture model. As before, the columns headed '$\bar{\hat{\sigma}}$(simul.)' indicate estimates of the parameters' standard errors derived from the empirical sample. For the LM estimates, there is a an additional column, '$\bar{\hat{\sigma}}$(asymp.)' that gives the average standard error derived from the approximation to the Fisher information matrix – see (3.40). The results show that the LM model has performed significantly better than the TAR model in determining the regime specific parameters. We reiterate that both models use Y_{t-2} instead of X_{t-2}; the 'TAR - Y_{t-2}' model

uses Y_{t-2} as the threshold variable and the 'LM - Y_{t-2}' model uses Y_{t-2} as a covariate in the logistic regression. It is surprising that although the LM model is misspecified, the standard error estimates for the β and ϕ terms derived from the asymptotic normality approximation ($\bar{\hat{\sigma}}$(asymp.)) agree closely with those derived from the empirical sample ($\bar{\hat{\sigma}}$(simul.)). We do not claim this agreement will hold in general, but it is nonetheless encouraging.

When we examine the results corresponding to the TAR model we suspect that because the threshold variable, X_{t-2}, is not directly observed, its imperfect proxy, Y_{t-2}, occasionally misclassifies observations into C_1 and C_0. Thus if we denote by f_1 the conditional density associated with $X_{t-2} < .1$ and f_0 the density associated with $X_{t-2} \geq .1$ the class C_1^T incorrectly contains some observations that were generated by f_0. Consequently, the estimates of β_1 and ϕ_1 will be

		LM - Y_{t-2}			TAR - Y_{t-2}	
Parameter	True	Average	$\bar{\hat{\sigma}}$ (asymp.)	$\bar{\hat{\sigma}}$(simul.)	Average	$\bar{\hat{\sigma}}$(simul.)
β_{11}	.5	.500	.0680	.0796	.415	.0523
β_{12}	.2	.197	.0580	.0590	.193	.0538
β_{01}	.1	.0740	.139	.169	.197	.0868
β_{02}	-.1	-.115	.106	.119	-.0652	.0812
ϕ_1	.25	.249	.0366	.042	.330	.0284
ϕ_0	.49	.483	.0743	.074	.446	.0501
γ_0	NA	.197	1.56	2.30	NA	NA
γ_1	NA	-3.70	3.89	7.11	NA	NA

Table 4.3: 200 Simulations of 500 Observations Using Y_{t-2} as Threshold Variable

biased because they are not based on the appropriate set of observations. A similar consequence holds for estimates of β_0 and ϕ_0 based on the class C_0^τ. The misclassifications act to pull the estimated parameter groups, (β_1, ϕ_1) and (β_0, ϕ_0), closer together. This bias would be more pronounced if the two sets of parameters were further apart. Were the parameters closer together the bias would be attenuated though still present. Also, the more noise in the threshold variable (and hence the more likely that misclassification occurs) the greater the bias in the estimates. Though not presented here we have performed additional simulations that exhibit this behavior. Finally, we have presented the model as a fixed threshold (.1 in this case) with an imperfectly observed threshold variable. The same results would be had if we considered the threshold variable perfectly observed but the threshold level varying randomly about some mean in an unobservable manner. If the statistician models such a process with a fixed, constant threshold the same types of misclassification and consequent bias would arise.

We produce one more table based on this set of simulations. Using the same data realizations as above, we use X_{t-2} instead of Y_{t-2} in computing the TAR coefficients. This corresponds to the unlikely occurrence that the analyst observers the threshold variable, X_{t-2}, without error. We reproduce the LM estimates from above as a basis for comparison, reiterating that *these LM estimates were produced using the proxy* Y_{t-2} *instead of* X_{t-2} in the logistic regression. The most noteworthy result in Table 4.4 is that the LM estimates for the regime-specific parameters (the β and ϕ terms) are not much worse than those derived from the correctly specified model with perfectly observed data. The estimates corresponding to the regime for $X_{t-2} < .1$ are quite good while those for the $X_{t-2} \geq .1$

		LM - Y_{t-2}			TAR - X_{t-2}	
Parameter	True	Average	$\bar{\hat{\sigma}}$ (asymp.)	$\bar{\hat{\sigma}}$(simul.)	Average	$\bar{\hat{\sigma}}$(simul.)
β_{11}	.5	.500	.0680	.0796	.501	.0503
β_{12}	.2	.197	.0580	.0590	.200	.0494
β_{01}	.1	.0740	.139	.169	.091	.0774
β_{02}	-.1	-.115	.106	.119	-.104	.0792
ϕ_1	.25	.249	.0366	.042	.253	.0217
ϕ_0	.49	.483	.0743	.074	.496	.0501
γ_0	NA	.197	1.56	2.30	NA	NA
γ_1	NA	-3.70	3.89	7.11	NA	NA

Table 4.4: 200 Simulations of 500 Observations Using X_{t-2} as Threshold Variable

regime are not as good.

This particular set of simulations indicates that in some instances, the LM estimation procedure may produce superior regime specific estimates in a two regime threshold model. This is particularly true if either the threshold variable is subject to noise and thus imperfectly measured or if the threshold value is variable though modeled as constant.

4.3 Application to Rain Rates

In this section we fit a logistic mixture to rain rate data. The Global Atmospheric Research Program's Atlantic Tropical Experiment (GATE) data provides ship-based radar measurements of rainfall collected from the South Atlantic during the summer of 1974. Details regarding the data collection are available from

Hudlow and Patterson (1974).

The selection of data used in this section is drawn from the GATE Phase I dataset. Every 15 minutes a radar snapshot of rain was obtained. This grid was divided into pixels of size 4 km by 4 km and an average rain rate was computed for each pixel. Our dataset contains only positive rain rates and the hour in which the observation was made. Pixels in our dataset were selected so that a minimum of 32 km separated them from any other pixel in the dataset. Our data consists of 860 such observations.

For modeling purposes we treat the data as independently distributed. While this assumption may not be entirely justified it is common with GATE data – see Kedem, Pfeiffer, and Short (1997) and Bell and Suhasini (1994). This independence is in contrast to our general model formulation in which Y_t may depend upon $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}$ as well as some exogenous covariates. However, our general model does accommodate independent and/or identically distributed data as a special case.

Some analysts have suggested that a log-normal or gamma distribution may provide a reasonable parametric model for rain rate, given that it is raining (e.g. Kedem, Pfeiffer, Short (1997) and Kedem, Chiu, and North (1990)). Others have suggested there are different types of rain (Houze (1981)) which indicates that a mixture density may be appropriate (Sansom and Thomson (1992) and Bell and Suhasini (1994)). The GATE dataset is particularly well suited for analysis via mixtures as the tropical weather patterns exhibit both longer periods of moderate rainfall (termed stratiform rain) as well as shorted periods of more intense rain (convective rain).

We decided to model the data as a two regime mixture of log-normal distri-

bution with the component distributions corresponding to stratiform and convective rain patterns. A logistic mixture model allows us to parameterize the regime probabilities using a logistic regression model. Bell and Suhasini (1994) have suggested the regime probabilities should follow a daily, or diurnal, cycle. Our model is well suited for this purpose: if $h_t \in \{1, 2, \ldots, 24\}$ denotes the hour the t^{th} observation is made we model

$$p_t^s(h_t; a, b, d) = \mathbb{P}\left[t^{th} \text{ observation is stratiform } \mid h_t; a, b, d\right] = \frac{\exp\left(a \sin\left(w h_t + b\right) + d\right)}{1 + \exp\left(a \sin\left(w h_t + b\right) + d\right)} \text{ and}$$

$$p_t^c(h_t; a, b, d) = 1 - p_t^s(h_t; a, b, d) = \mathbb{P}\left[t^{th} \text{ observation is convective} \mid h_t; a, b, d\right] = \frac{1}{1 + \exp\left(a \sin\left(w h_t + b\right) + d\right)} \text{ where } w = \frac{2\pi}{24}.$$

Remark: As an aside we note that the parameterization of the probabilities as written above is not of the $Z_t'\gamma$ form of our general model. This $Z_t'\gamma$ form is necessary to implement the EM algorithm, as we developed it, in Chapter 2. We may reparameterize the probabilities in the $Z_t'\gamma$ form in the following way:

$$a \sin(w h_t + b) + d = a \cos(b) \sin(w h_t) + a \sin(b) \cos(w h_t) + d \text{ and setting}$$

$$Z_t = (1, \sin(w h_t), \cos(w h_t))' \text{ and } \gamma = (\gamma_0, \gamma_1, \gamma_2)' = (d, a\cos(b), a\sin(b))'.$$

Then point estimates of a and b may be obtained by considering $\hat{b} = \tan^{-1}(\hat{\gamma}_2/\hat{\gamma}_1)$ and $\hat{a} = \hat{\gamma}_1/\cos(\hat{b})$. Standard errors can be derived using a multivariate delta method approach (see page 402 in Billingsley (1986)). Alternatively, one could estimate the model using a general non-linear minimization package and estimate the Fisher information matrix in an obvious way.

The way we model p_t^s and p_t^c imposes a diurnal cycle as long as $a \neq 0$. The a parameter controls the variability, or amplitude in the cycle. The b gives

freedom to the phase shift of the cycle and the d term allows these probabilities to fluctuate about some average value different from .5. If we denote the log of our observed rain rate by Y_t then our logistic mixture model is given by

$$g\left(y_t \mid h_t; \psi = (\mu_s, \phi_s, \mu_c, \phi_c, a, b, d)\right) = p_t^s(h_t; a, b, d) f(y_t; \mu_s, \phi_s) + \tag{4.5}$$

$$p_t^c(h_t; a, b, d) f(y_t; \mu_c, \phi_c) \tag{4.6}$$

where $f(\cdot; \mu, \phi)$ denotes the density of a $N(\mu, \phi)$ random variable and the s and c subscripts are labels designating stratiform and convective. As a basis of comparison we include results for two nested models:

$$g(y_t; \mu_s, \phi_s, \mu_c, \phi_c, p) = p f(y_t; \mu_s, \phi_s) + (1-p) f(y_t; \mu_c, \phi_c) \text{ and} \tag{4.7}$$

$$g(y_t; \mu, \phi) = f(y_t; \mu, \phi). \tag{4.8}$$

The model in (4.7) corresponds to a standard mixture model with fixed regime probabilities and the model in (4.8) is a one regime, or no mixture model. In the table below we present point estimates and standard errors (in parentheses) for the three models. The column headed 'LM' corresponds to results for logistic mixtures, the '2R' heading denotes the mixture with constant regime probabilities in (4.7) and the '1R' indicates results for the one regime model.

The results suggest that in passing to each of the more restrictive nested model the explanatory power is significantly weakened though we discuss no formal test of such hypotheses until the next chapter. What may be most surprising is the apparent power that is gained by parameterizing the regime probabilities according to a daily cycle. The addition of the b and a parameters increases the log likelihood by approximately 36 over what was obtained from the 2R model.

Parameter	LM	2R	1R
μ_s	-.0930	-1.06	.497
	(.0831)	(.0984)	(.0436)
ϕ_s	.907	.175	1.63
	(.104)	(.0827)	(1.81)
μ_c	1.74	.713	NA
	(.160)	(.0944)	NA
ϕ_c	.870	1.45	NA
	(.157)	(.134)	NA
d	1.12	1.974	NA
	(.334)	(.418)	NA
b	.766	NA	NA
	(.116)	NA	NA
a	1.48	NA	NA
	(.234)	NA	NA
Log Likelihood	-1380.97	-1417.05	- 1430.09

Table 4.5: Three Models of Rain Rate

One might want to conclude that under the null hypothesis that the 2R model is correctly specified, then

$$2 * (\text{Log-Likelihood(LM)} - \text{Log-Likelihood(2R)}) \xrightarrow{\mathcal{D}} \chi^2_2.$$

This would be incorrect as the b term is not identified under the null hypothesis that $p^s_t(h_t; a, b, d) = p$, a constant. By this we mean while it is clear that $a = 0$ under the null hypothesis, any value of b would suffice, and hence b is not

identified. A similar though more difficult identification problem arises if we want to compare either the LM or 2R model to the 1R model. These identifiability problems invalidate traditional approaches to hypothesis testing. We will take up this question at length in the next chapter.

For now we will interpret the result of fitting the LM model. From the parameter estimates we can obtain mean rain rates for each of the two densities. The mean for the stratiform regime is $\exp(-.0930 + .5 * .907) = 1.43$ mm/hr. That for the convective regime is 8.82 mm/hr. From these means and the derived hourly regime probabilities we produce estimated hourly rain rates. A plot of the stratiform regime probabilities and the estimated hourly rates is included in Figure 4.1.

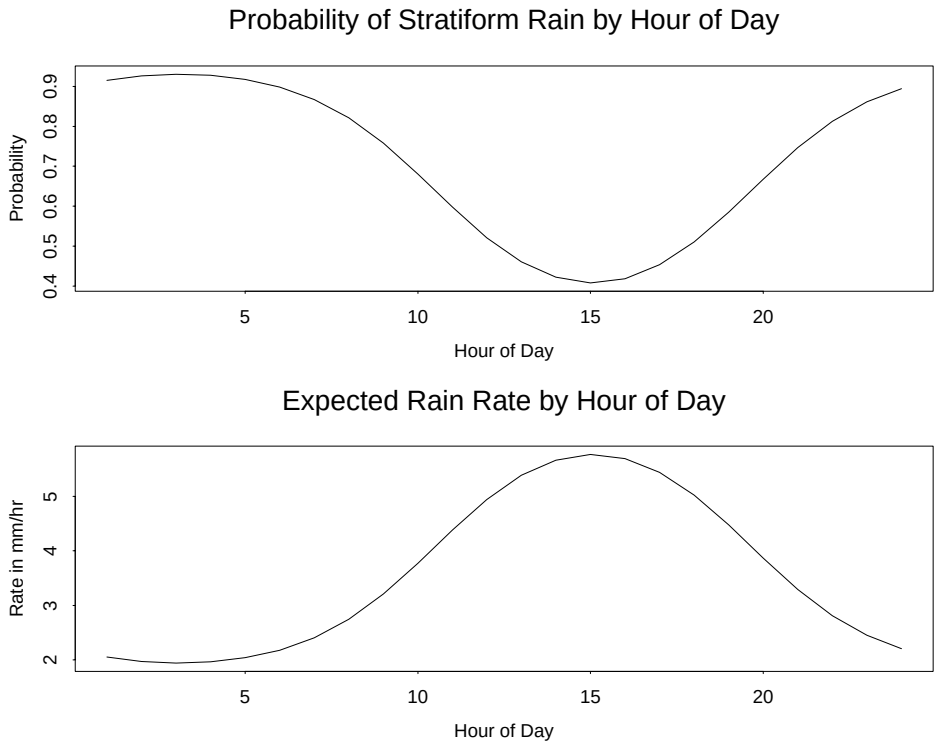

Figure 4.1: Fitted Hourly Estimates

The plots indicate that the more intensive rain rates are associated with the afternoon, when warmer temperatures may make such events more likely.

The results of Table 4.5 include the surprising degree to which our estimates of stratiform and convective parameters differ between the LM and 2R model. This discrepancy may best be explained through examination of a histogram of the log rain rates (see Figure 4.2).

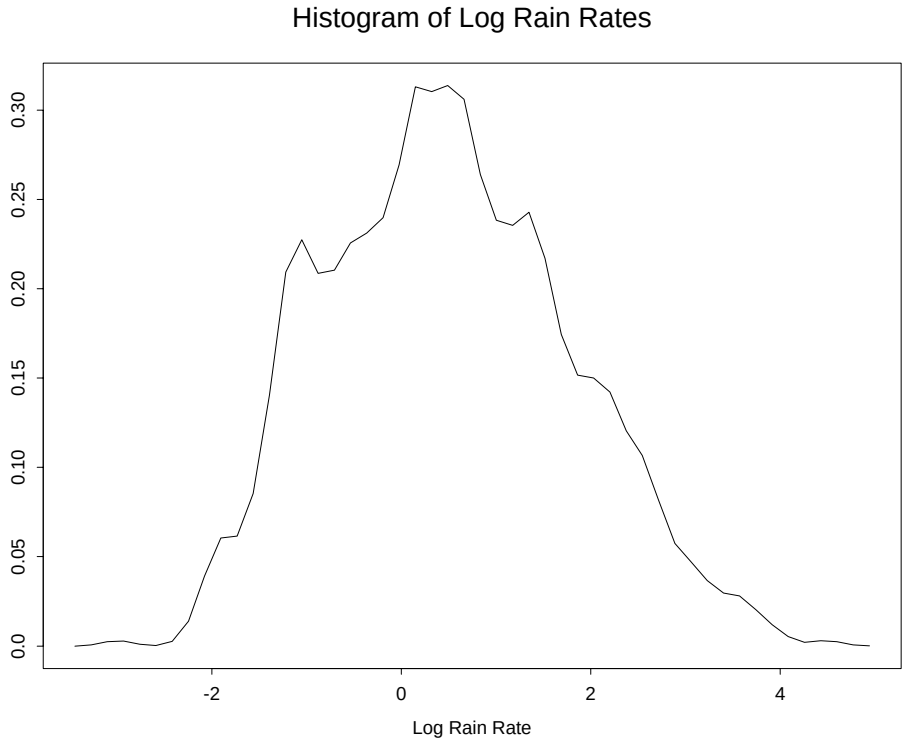

Figure 4.2: Log Rain Rate Distribution

The histogram indicates there may be more than two modes in our mixture distribution and that the 2R log-likelihood is maximized around a different pair of modes than those that maximize the LM log-likelihood. This finding suggests we should be diligent in making sure the log-likelihood is maximized at the reported values. In our investigations the 2R and LM estimates were both robust

to different starting values as well as different optimization methods (both the EM algorithm approach outlined in Chapter 2 as well as generic minimization routines were used). We are not sure how to explain why the two models pick out such different component densities other than to surmise that the parameterization of the diurnal cycle captures effects that are obscured when constant regime probabilities are imposed.

These results also suggest that perhaps investigation of a three regime model is warranted. Bell and Suhasini used a non-parametric principal components approach to estimating a mixture of densities and found support for 2 distributions (with mean rain rates of 2.6 and 8.8 mm/hr – very similar to our results) but that the data was not better fit by allowing for 3 distributions. We performed some estimation of 3 regime models with fixed probabilities but the estimates were not robust (i.e. different starting points led to different estimates) and the greatest log likelihood we were able to achieve was -1410.5 – still inferior to that obtained under the 2 regime LM model (-1381). The development of logistic mixture models with 3 or more regimes should be part of future work.

In this chapter we have made no attempt to formally test which of the three models (1R, 2R, or LM) may be better. In the next chapter we address some aspects of how to make such comparisons.

Chapter 5

A Likelihood Ratio Test of One vs. Two Regimes

In this chapter we develop a likelihood ratio test for determining if the data come from a logistic mixture of distinct distributions or arise from a single distribution (i.e. no mixture). It is well known that this test is non-standard in the sense that some parameters are not identifiable under the null hypothesis of no mixture (Ghosh and Sen (1985) and McLachlan and Basford (1988)). Among econometricians there has been recent work on likelihood ratio tests with non-identifiable parameters (Andrews (1993), Andrews and Ploberger (1994), and Hansen (1996a)) though, as they point out, their techniques are not directly applicable to mixture data. These tests have been applied to threshold and change point models but the unobserved (or latent) variable structure of mixture models make these techniques inappropriate as singular Fisher information matrices are encountered. This singularity is the basis for the difficulties with mixtures.

There have been a number of recent approaches to constructing a likelihood ratio test for the number of regimes in a mixtures with constant transition probabilities and i.i.d. random variables (Garel (1996), Lemdani and Pons (1995), and

Dacunha-Castelle and Gassiat (1997)). These papers are generally extensions of the ideas in Ghosh and Sen though the Dacunha-Castelle and Gassiat paper introduces new techniques and a broader level of generality. The logistic mixture models we consider have two primary differences from standard mixtures: 1) the transition probabilities vary, and 2) we introduce correlation through time. The variable transition probabilities in the logistic mixture do not pose any new significant problems – indeed the variation in these probabilities will prove to be of great use to us. However, the time series aspect invalidates the moment conditions necessary for the application of weak and uniform convergence theorems cited in the papers using i.i.d. data. At least this is true for our test case of a logistic mixture of Gaussian AR(1) processes. This problem will be spelled out below. As a consequence we develop a new test that has the flavor of the approaches used for i.i.d. data with constant regime probabilities but can accommodate the test case of mixing AR(1) processes. Unfortunately this extension comes at a cost. The mathematical feature of the model that allows us to implement our test is the variation (non-constancy) in the regime probabilities. This means we are only able to test the hypothesis of no mixture against the alternatives of mixtures with varying regime probabilities – we are unable to include mixtures with fixed regime probabilities as part of the set of alternatives. We will be more clear below.

In Section 5.1 we will first discuss the problems with testing for mixtures in the simpler case of i.i.d. data with constant regime probabilities. Section 5.2 introduces our approach to the problem using the variable regime probabilities – first in the context of i.i.d. data and then in our time dependent situation. Here we derive the asymptotic distribution for the likelihood ratio statistic of

models that obey a series of conditions. In the subsequent section we show that our logistic mixture of normally distributed AR(1) processes satisfies these conditions. In the following chapter we address the question of implementing the test in practice, check its performance via simulations, and apply the test to the rain data introduced in Chapter 4.

5.1 Problems Associated with Tests for Mixtures

The set of problems one encounters when testing for mixtures are extensive and generally invalidate conventional approaches. To introduce the problems we begin by discussing a simple mixture model with i.i.d. data and constant regime probabilities.

Suppose $\{x_t\}, t = 1...T$ are i.i.d. and $f(x;\alpha)$ is a parametric family of densities with $\alpha \in A$, a parameter space. We want to test

H_0: $X \sim f(x;\alpha^*)$ for some unknown $\alpha^* \in A$ versus

H_1: $X \sim g(x;\alpha_1^*, \alpha_0^*, p^*)$ where $g(x;\alpha_1,\alpha_0,p) = pf(x;\alpha_1) + (1-p)f(x;\alpha_0)$, $\alpha_1^* \neq \alpha_0^*$, α_1^*, α_0^* are unknown elements in A, and p^* an unknown point in $(0,1)$.

Let $\psi = (\alpha_1, \alpha_0, p)$. A naive likelihood ratio approach might be to examine

$$\Lambda_T \stackrel{\triangle}{=} 2\left(\sup_{\alpha_1,\alpha_0,p}\sum_{t=1}^{T}\log g(x_t;\alpha_1,\alpha_0,p) - \sup_{\alpha}\sum_{t=1}^{T}\log f(x_t;\alpha)\right)$$

and suppose that $\Lambda_T \xrightarrow{\mathcal{D}} \chi_2^2$ or χ_1^2 distribution. An examination of the proof of such results (e.g. Theorem 5.6.3 in Sen and Singer (1993) or Wilks (1937)) shows

that to apply this result there must exist 'true' values $\psi^* = (\alpha_1^*, \alpha_0^*, p^*)$ such that for $\widehat{\psi} \stackrel{\Delta}{=} \text{Arg max}_\psi \sum \log g(x_t; \alpha_1, \alpha_0, p)$ we have $\widehat{\psi} \xrightarrow{P} \psi^*$. (Throughout this section it will be understood that $\widehat{\psi}$ and its components depend upon T though we suppress this notation.) Under the null hypothesis of no mixture such true values do not exist. To see why this is true suppose that we further restrict our parameter space under the alternative to satisfy $p \in [\epsilon, 1-\epsilon]$ for some small, positive ϵ. Then if we estimate the mixture model when the data are generated under the null hypothesis with $\alpha = \alpha^*$ we would expect for large T to have $\widehat{\alpha}_1 \approx \alpha^*$ and $\widehat{\alpha}_0 \approx \alpha^*$ (i.e. one can show $\widehat{\alpha}_1 \xrightarrow{P} \alpha^*$ and $\widehat{\alpha}_0 \xrightarrow{P} \alpha^*$ under mild conditions). But while we can estimate $\widehat{p}$ for any fixed T, it will never converge to any fixed value. This is so because $\sum \log g(x_t; \alpha_1 = \alpha^*, \alpha_0 = \alpha^*, p) = \sum \log f(x_t; \alpha^*)$ for any p – in other words p is not identified. As $\widehat{\alpha}_1$ and $\widehat{\alpha}_0$ converge to α^*, $\widehat{p}$ will randomly move in $[\epsilon, 1-\epsilon]$ as the sample size grows and not approach any particular value.

Alternatively, suppose we restrict the parameter space to have $|\alpha_1 - \alpha_0| > \delta, p \geq \epsilon > 0$. Then for large T we would obtain $\widehat{\alpha}_1 \approx \alpha^*$ and $\widehat{p} \approx 1$ but then α_0 will be unidentified because $\sum \log g(x_t; \alpha_1 = \alpha^*, \alpha_0, p = 1) = \sum \log f(x_t; \alpha^*)$ for any α_0. In this case $\widehat{\alpha}_0$ will not converge to any true value. We obtain a similar result if we require $|\alpha_1 - \alpha_0| > \delta, p \leq 1 - \epsilon$ in which case α_1 will be unidentified when $p = 0$. If we do not restrict the parameter space at all than it is not clear that any of the components of $\widehat{\psi}$ will converge. This example illustrates that any approach to the mixture problem that supposes the existence of true parameters under the null hypothesis will fail unless it is altered to take into account the identifiability problem. This same problem is present in conventional Akaike Information Criterion (Akaike (1973)), Lagrange multiplier, Wald (see

Chapter 5 in Sen and Singer), and Generalized Method of Moments (chapter 14 in Hamilton (1994) and Hansen (1982)) test procedures. Mixture models are not unique in this aspect of having parameters that are identifiable only under the alternative hypothesis. Change point and threshold models have the same type of difficulty. Andrews (1993) and Andrews and Ploberger (1994) have developed and summarized an empirical processes approach to these types of problems utilizing likelihood ratio, Wald, and Lagrange multiplier tests. Although their methods do not work for mixture models they are of the same general type.

The motivation for our approach will arise from an appreciation of some of the mathematical difficulties we encounter when we try to apply a conventional approach. To make matters concrete let us retain our example above. Suppose the data X_t are generated by $f(x;\alpha^*)$ for some unknown $\alpha^* \in A$. Without restrictions on our parameter space there are three ways to write $f(x;\alpha^*)$ in terms of our mixture distribution $g(x;\psi)$:

1. $f(x;\alpha^*) = g(x;\alpha_1 = \alpha^*, \alpha_0 \text{ unspecified}, p = 1)$, or

2. $f(x;\alpha^*) = g(x;\alpha_1 \text{ unspecified}, \alpha_0 = \alpha^*, p = 0)$, or

3. $f(x;\alpha^*) = g(x;\alpha_1 = \alpha^*, \alpha_0 = \alpha^*, p \text{ unspecified})$.

By restricting our parameter space we may choose one of these representations of f in terms of the mixture, g. The reason for this is that we may specify how the estimated parameters should behave if the null hypothesis is true. For instance, if we require $p \in [\epsilon, 1-\epsilon]$ then we would obtain $\hat{\alpha}_1, \hat{\alpha}_0 \xrightarrow{P} \alpha^*$. If we were to try to analyze such statistics in a conventional way we would be interested in calculating the Fisher information. If we let $l(x;\psi) = \log g(x;\psi)$ then we would

need to compute

$$\mathbb{E}\left[\left(\frac{\partial l}{\partial p}\right)^2\Bigg|_{H_0}\right] = \mathbb{E}\left[\frac{\left(f(x;\alpha_1)-f(x;\alpha_0)\right)^2}{g(x;\psi)^2}\Bigg|_{H_0}\right]$$
$$= \mathbb{E}\left[\frac{\left(f(x;\alpha^*)-f(x;\alpha^*)\right)^2}{f(x;\alpha^*)^2}\right] = 0$$

because the integrand is identically 0. Consequently,

$$\frac{\partial^2 l}{\partial\psi\partial\psi'}$$

is not invertible for any p and thus the conventional variance-covariance matrix does not exist. This variance matrix is very important in all of the empirical processes approaches and without it we cannot move forward. We encounter this same type of problem if we use either of the other two restrictions on the parameter space (i.e. $p \in [0, 1-\epsilon]$ or $p \in [\epsilon, 1]$ with $|\alpha_1 - \alpha_0| > \delta$).

Redner (1981) and Feng and McCulloch (1996) found that if we choose one of these three restrictions and the null hypothesis is true, then the identified parameters will be consistent. For example if we restrict $p \in [\epsilon, 1-\epsilon]$ then they show $\widehat{\alpha}_1 \xrightarrow{P} \alpha^*$ and $\widehat{\alpha}_0 \xrightarrow{P} \alpha^*$ though $\widehat{p}$ is not identified and will randomly wander as the sample size grows. However, these authors were unable to make conclusions about asymptotic inference that would help in testing whether a mixture is present. In other words, while it is possibly true there exists a Q such that

$$\sqrt{T}\begin{pmatrix}\widehat{\alpha}_1 - \alpha^* \\ \widehat{\alpha}_0 - \alpha^*\end{pmatrix} \xrightarrow{\mathcal{D}} N(\underline{0}, Q)$$

under the null hypothesis and the restriction that $p \in [\epsilon, 1-\epsilon]$, no one has been able to determine Q in a general case. If such a Q was determined then we could use the relation above as a basis for hypothesis testing.

Ghosh and Sen (1985) used empirical process results to approach this problem. They analyzed the case of $f(x;\alpha), \alpha \in \mathbb{R}$. They had in mind $f(x;\alpha)$ corresponding to a normal distribution with unknown mean $\alpha \in \mathbb{R}$ and a known variance of 1. In their approach they chose to restrict $|\alpha_1 - \alpha_0| > \delta, p \in [0, 1-\epsilon]$ and thus the null hypothesis distribution $f(x;\alpha^*)$ has the mixture representation

$$g(x;\alpha_1 = \text{ unspecified}, \alpha_0 = \alpha^*; p = 0) = 0 \cdot f(x;\alpha_1) + 1 \cdot f(x;\alpha_0 = \alpha^*).$$

Their idea was to perform profile (or concentrated) likelihood, holding fixed the non-identified parameter and then treating the resulting quantity as an empirical process in that parameter. In this case α_1 is the non-identified parameter. We sketch their argument as follows: for fixed $\alpha_1 \in A$ let

$$L_T^g(\alpha_1) = \sup_{(\alpha_0,p)\in A(\alpha_1,\delta,\epsilon)} \sum_{t=1}^{T} \log g(x_t;\alpha_1,\alpha_0,p) \text{ and}$$

$$(\widehat{\alpha}_0(\alpha_1), \widehat{p}(\alpha_1)) = \text{Arg} \sup_{(\alpha_0,p)\in A(\alpha_1,\delta,\epsilon)} \sum_{t=1}^{T} \log g(x_t;\alpha_1,\alpha_0,p).$$

where $A(\alpha_1,\delta,\epsilon) = \{(\alpha_0,p) : \alpha_0 \in A, |\alpha_1 - \alpha_0| > \delta, 0 \le p \le 1-\epsilon\}$. Then under some mild conditions they show

$$(\widehat{\alpha}_0(\alpha_1), \widehat{p}(\alpha_1)) \xrightarrow{P} (\alpha^*, 1) \text{ for all } \alpha_1 \in A, \text{ and}$$

$$2\left(L_T^g(\alpha_1) - \sup_{\alpha\in A} \sum_{t=1}^{T} \log f(x_t;\alpha)\right) = (D_T(\alpha_1))^2 \cdot I_{[D_T(\alpha_1)\ge 0]} \qquad (5.1)$$

where $D_T(\alpha_1) \xrightarrow{\mathcal{D}} N(0,1)$ for all α_1 under the null hypothesis. (The indicator $I_{[\cdot]}$ arises from the fact that p=0 lies on the boundary of the parameter space – see Chernoff (1954)). At this point we do not concern ourselves with the particular form of D_T other than its limiting distribution. Then the log-likelihood ratio

statistic

$$2\left(\sup_{\alpha_1\in A,(\alpha_0,p)\in A(\alpha_1,\delta,\epsilon)}\sum\log g(x_t;\alpha_1,\alpha_0,p)-\sup_{\alpha\in A}\sum\log f(x_t;\alpha)\right) \text{ reduces to}$$

$$2\left(\sup_{\alpha_1\in A}L_T^g(\alpha_1)-\sup_{\alpha\in A}\sum\log f(x_t;\alpha)\right)=\sup_{\alpha_1}D_T^2(\alpha_1)\cdot I_{[D_T(\alpha_1)\geq 0]}$$

To find the distribution of this last quantity they show there exists a mean zero Gaussian process, $D(\cdot)$, indexed by α_1 such that 1) for any fixed $\alpha_1\in A, D(\alpha_1)$ has a $N(0,1)$ distribution, and 2) $D_T(\cdot)\xrightarrow{\mathcal{W}}D(\cdot)$ where we interpret $D_T(\cdot)$ to be a stochastic processes on A indexed by α_1 and $\xrightarrow{\mathcal{W}}$ denotes weak convergence (see Pollard (1984) or Billingsley (1968) for extensive discussion of weak convergence). By the continuous mapping theorem we have

$$2\left(\sup_{\alpha_1\in A,(\alpha_0,p)\in A(\alpha_1,\delta,\epsilon)}\sum\log g(x_t;\alpha_1,\alpha_0,p)-\sup_{\alpha\in A}\sum\log f(x_t;\alpha)\right)$$
$$\xrightarrow{\mathcal{D}}\sup_{\alpha_1}\left(D(\alpha_1)\right)^2\cdot I_{[D(\alpha_1)\geq 0]}.$$

Determining the distribution of the right-hand side above is difficult because the supremum's distribution will depend upon the covariance kernel of $D(\alpha_1)$ which is not easily determined in general and often must be estimated through simulation. If one can simulate the $\left(D(\alpha_1)\right)^2\cdot I_{[D(\alpha_1)\geq 0]}$ process then one can use the suprema of the simulations to obtain an empirical distribution and consequently critical points for a test under the null hypothesis. Dacunha-Castelle and Gassiat (1997) have extended these ideas to more general tests of mixtures in the i.i.d. case (for example, testing a mixture with p components versus one of q components).

To this point in our work, the extension of techniques for i.i.d. data to time dependent data has not created much difficulty as we have found central limit theorems and laws of large numbers to use that are valid for dependent, non-

identically distributed data. But here we do run into surprising problems for some time series models. In both the Ghosh and Sen and Dacunha-Castelle and Gassiat papers a basic quantity of analysis is

$$\mathbb{E}\left[\left(\frac{f(x;\alpha_1)-f(x;\alpha^*)}{f(x;\alpha^*)}\right)^k\right]=\mathbb{E}\left[\left(\frac{f(x;\alpha_1)}{f(x;\alpha^*)}-1\right)^k\right] \tag{5.2}$$

for various powers of k where α_1 is an arbitrary element of A and α^* is the true parameter under the null hypothesis. In the work of Ghosh and Sen this term arises from moments of

$$\left.\frac{\partial \log g(x;\alpha_1,\alpha_0,p)}{\partial p}\right|_{\alpha_0=\alpha^*,p=0}$$

and Dacunha-Castelle and Gassiat restrict attention to parameter combinations of α_1 and α^* for which the expectations exist with k at least 2. Both sets of authors assume these expectations exist for all $(\alpha_1,\alpha^*)\in A\times A$ (or except for an arbitrarily small area of $A\times A$ in Ghosh and Sen's case). In the time series case these expectations do not always exist. To see this suppose we revisit our test case of a logistic mixture of AR(1) processes. We will take the mixture probabilities as constant (p) and suppose the component conditional densities have common known variance 1 (the analysis remains the same if we allow the more general circumstances of varying probabilities and different unknown variances). Within this framework we are interested in finding

$$\mathbb{E}\left[\left(\frac{f(Y_t\mid Y_{t-1};\alpha_1)}{f(Y_t\mid Y_{t-1};\alpha^*)}\right)^k\right].$$

Then under the null hypothesis we assume $Y_t=\alpha^*\cdot Y_{t-1}+\epsilon_t$ where ϵ_t is i.i.d. $N(0,1)$. Then assuming the Y_t are identically distributed with the common

stationary distribution $N(0,(1-(\alpha^*)^2)^{-1})$ we see

$$\mathbb{E}\left[\left(\frac{f(Y_t \mid Y_{t-1};\alpha_1)}{f(Y_t \mid Y_{t-1};\alpha^*)}\right)^k\right] = \mathbb{E}\left[\mathbb{E}\left[\left(\frac{f(Y_t \mid Y_{t-1};\alpha_1)}{f(Y_t \mid Y_{t-1};\alpha^*)}\right)^k \middle| Y_{t-1}\right]\right] \tag{5.3}$$

$$=\mathbb{E}\left[\int_{\mathbb{R}} \left[\varphi\left(y_t - Y_{t-1}\alpha_1\right)\varphi\left(y_t - Y_{t-1}\alpha^*\right)^{-1}\right]^k \varphi\left(y_t - Y_{t-1}\alpha^*\right) dy_t\right] \tag{5.4}$$

where $\varphi(x) = (2\pi)^{-1/2}\exp(-x^2/2)$. After some algebra this is simplified to

$$\int_{\mathbb{R}} \sqrt{\frac{1-(\alpha^*)^2}{2\pi}} \exp\left(Y_{t-1}^2((k^2-k)(\alpha_1-\alpha^*)^2 - (1-(\alpha^*)^2))\right) dY_{t-1}$$

which is finite iff $(k^2-k)(\alpha_1-\alpha^*)^2 - (1-(\alpha^*)^2) < 0$. This restriction greatly reduces our initial parameter space of $(\alpha_1,\alpha^*) \in [-1+\epsilon, 1-\epsilon] \times [-1+\epsilon, 1-\epsilon]$. In Figure 5.1 below the area between the curves show the allowable (α_1,α^*) combinations for $k=2$.

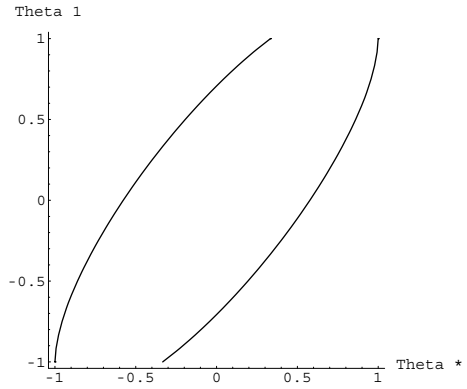

Figure 5.1: Allowable α^* and α_1 Combinations for $k=2$

The graph indicates that for $\alpha^* = .5$ this expectation is finite only for $\alpha_1 \in (0,1)$, approximately. The next figure shows the allowable region if we need a finite expectation for $k=4$ (which may be convenient for using the Cauchy-Schwarz inequality). Here we see for $\alpha^* = .5$ we require $\alpha_1 \in (.3,.7)$, approximately.

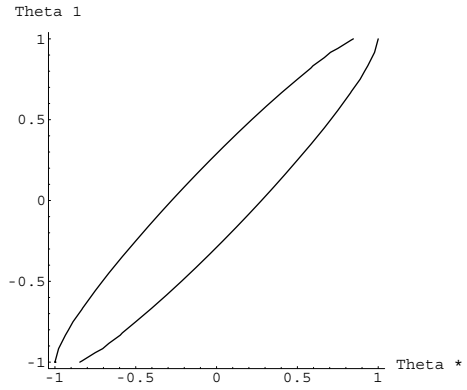

Figure 5.2: Allowable α^* and α_1 Combinations for $k = 4$

For us these parameter boundaries seem too restrictive to be of use in testing the mixture of AR(1) processes – the set of mixtures under the alternative hypothesis seems too small. These figures contrast sharply to the i.i.d. case where for normal component densities these expectations exist for all $(\alpha_1, \alpha^*) \in \mathbb{R} \times \mathbb{R}$, and k as long as the variances ϕ_1 and ϕ^* are the same or 'close' to one another (the more moments that are required, the closer they must be).

As we feel any method we try should apply to our test case of an AR(1) logistic mixture we are forced to develop a different test. In Ghosh and Sen α_1 was initially held constant and profiled maximization was performed with respect to p and α_0. Our approach is analogous to initially holding constant p and maximizing with respect to α_1 and α_0. This gives rise to a different set of derivatives that avoid expectations of the form

$$\mathbb{E}\left[\left(\frac{f(Y_t \mid Y_{t-1}; \alpha_1)}{f(Y_t \mid Y_{t-1}; \alpha^*)}\right)^k\right].$$

While this choice of parameterization leads to some other problems that force us to restrict the parameter space under the alternative hypothesis, we are able

to obtain results for a reasonably large set of alternatives to the null hypothesis.

5.2 The Likelihood Ratio Test

In this section we describe a likelihood ratio test for a restricted set of alternative hypotheses that allow us to obtain an asymptotic test for this smaller set of alternatives. Our restricted set of alternatives consists of those logistic mixtures with non-constant regime probabilities (minus a bit more to maintain a compact parameter space). We show below that it is the constancy of regime probabilities that causes the problems with the information matrix which invalidate an empirical process approach to the problem. With non-constant probabilities the problem is eliminated.

Before describing our test we think it useful to re-examine the simple mixture model in the context of i.i.d. data to indicate the problem and its solution. Here we sketch the overall idea – we defer the proofs until we discuss the problem in the context of time dependent logistic mixtures. We return to the situation we investigated in Section 5.1. Suppose $\{X_t\}, t = 1...T$ are i.i.d. and $f(x;\alpha)$ is a parametric family of densities with $\alpha \in A$, a subset of R^d for some d. We want to test

H_0: $X \sim f(x;\alpha^*)$ for some unknown $\alpha^* \in A \subset \mathbb{R}^d$ versus

H_1: $X \sim g(x;\alpha_1^*, \alpha_0^*, p^*)$ where $g(x;\alpha_1,\alpha_0,p) = pf(x;\alpha_1) + (1-p)f(x;\alpha_0)$, $p^* \neq 0, 1, \alpha_1^* \neq \alpha_0^*$, and α_1^*, α_0^* are unknown elements in A. Furthermore we restrict p^* to be in $[\epsilon, 1-\epsilon]$.

Here we adopt a profile likelihood technique; we first hold p fixed and examine

the likelihood ratio associated with that p. Let us define

$$(\hat{\alpha}_1(p), \hat{\alpha}_0(p)) = \text{Arg} \max_{\alpha_1, \alpha_0} L_T^g(p, \alpha_1, \alpha_0) \text{ where} \tag{5.5}$$

$$L_T^g(p, \alpha_1, \alpha_0) = \sum_{t=1}^{T} \log g(x_t; \alpha_1, \alpha_0, p). \tag{5.6}$$

We assume the null hypothesis holds, i.e. there exists a $\alpha^* \in A$ such that the data's true density is given by $f(x; \alpha^*)$.

We also assume the set of one and two regime mixtures are identified in the sense that if $p \notin \{0, 1\}$ then $\mathbb{E}[\log g(X; \alpha_1, \alpha_0, p) - \log f(X; \alpha^*)] \leq 0$ and equality holds iff $\alpha_1 = \alpha_0 = \alpha^*$. As discussed in Chapter 3 this is the case for many exponential family distributions. This strict inequality is the critical condition in a Wald-like approach to consistency. From this relation (and some other mild conditions) it can be shown that

$$(\hat{\alpha}_1(p), \hat{\alpha}_0(p)) \xrightarrow{P} (\alpha^*, \alpha^*).$$

From the usual likelihood ratio expansion about the values (α^*, α^*) we obtain

$$L_T^g(p, \hat{\theta}) = L_T^g(p, \theta^*) - \tag{5.7}$$

$$\frac{1}{2} \frac{1}{\sqrt{T}} \frac{\partial L_T^g(p, \theta)}{\partial \theta'}\bigg|'_{\theta=\theta^*} \left[\frac{1}{T} \frac{\partial^2 L_T^g(p, \theta)}{\partial \theta \partial \theta'}\bigg|_{\theta=\bar{\theta}} \right]^{-1} \frac{1}{\sqrt{T}} \frac{\partial L_T^g(p, \theta)}{\partial \theta'}\bigg|_{\theta=\theta^*} \tag{5.8}$$

where $\theta = (\alpha_1', \alpha_0')', \hat{\theta} = (\hat{\alpha}_1', \hat{\alpha}_0')', \theta^* = (\alpha^{*\prime}, \alpha^{*\prime})'$, and $\bar{\theta}$ lies on a chord between $\hat{\theta}$ and θ^*. In the expression above we have assumed that

$$\left[\frac{1}{T} \frac{\partial^2 L_T^g(p, \theta)}{\partial \theta \partial \theta'}\bigg|_{\theta=\bar{\theta}} \right]$$

is invertible. Under the usual type of regularity conditions (Cramér (1946),

Chapter 5 in Sen and Singer (1993)) and a weak law of large numbers we have

$$\left[\frac{1}{T}\left.\frac{\partial^2 L_T^g(p,\theta)}{\partial\theta\partial\theta'}\right|_{\theta=\bar{\theta}}\right] \xrightarrow{P} \tag{5.9}$$

$$-\mathbb{E}\left[\begin{array}{cc} \left.\frac{\partial \log g(X,\alpha_1,\alpha_0)}{\partial\alpha_1}\right|^2_{\theta=\theta^*} & \left.\frac{\partial \log g(X,\alpha_1,\alpha_0)}{\partial\alpha_1}\frac{\partial \log g(X,\alpha_1,\alpha_0)}{\partial\alpha_0}\right|_{\theta=\theta^*} \\ \left.\frac{\partial \log g(X,\alpha_1,\alpha_0)}{\partial\alpha_1}\frac{\partial \log g(X,\alpha_1,\alpha_0)}{\partial\alpha_0}\right|_{\theta=\theta^*} & \left.\frac{\partial \log g(X,\alpha_1,\alpha_0)}{\partial\alpha_1}\right|^2_{\theta=\theta^*} \end{array}\right]. \tag{5.10}$$

In the case of our mixture this equals

$$-\mathbb{E}\left[\begin{array}{cc} p^2\left[\left.\frac{\partial \log f(X;\alpha)}{\partial\alpha}\right|_{\alpha=\alpha^*}\right]^2 & p(1-p)\left[\left.\frac{\partial \log f(X;\alpha)}{\partial\alpha}\right|_{\alpha=\alpha^*}\right]^2 \\ p(1-p)\left[\left.\frac{\partial \log f(X;\alpha)}{\partial\alpha}\right|_{\alpha=\alpha^*}\right]^2 & (1-p)^2\left[\left.\frac{\partial \log f(X;\alpha)}{\partial\alpha}\right|_{\alpha=\alpha^*}\right]^2 \end{array}\right]. \tag{5.11}$$

The matrix above (without the negative sign) we will denote as $I(\alpha^*)$. If $I(\alpha^*)$ is invertible then it may be shown that (again under some regularity conditions)

$$2\left(L_T^g(p,\hat{\theta}) - L_T^g(p,\theta^*)\right) = \tag{5.12}$$

$$\frac{1}{\sqrt{T}}\left.\frac{\partial L_T^g(p,\theta)}{\partial\theta'}\right|'_{\theta=\theta^*}[I(\alpha^*)]^{-1}\frac{1}{\sqrt{T}}\left.\frac{\partial L_T^g(p,\theta)}{\partial\theta'}\right|_{\theta=\theta^*} + o_p(1) \xrightarrow{\mathcal{D}} \chi^2_{2d}(p) \tag{5.13}$$

where $o_p(1)$ means uniform convergence to 0 in probability and the chi-square random variable has 2d degrees of freedom because α is assumed to be a d dimensional vector. By uniform convergence we mean

$$\sup_{p\in[\epsilon,1-\epsilon]}\left\{\left[\frac{1}{T}\left.\frac{\partial^2 L_T^g(p,\theta)}{\partial\theta\partial\theta'}\right|_{\theta=\bar{\theta}}\right] + I\left(\alpha^*\right)\right\} \xrightarrow{P} \underline{0}.$$

From here we try try to continue analysis of the problem using the Ghosh and Sen approach outlined in the introductory section of of this chapter. By this we mean we treat $\chi^2_{2d}(p)$ as a stochastic process indexed by p in $[\epsilon, 1-\epsilon]$. The problem is that $I(\alpha^*)$ is not invertible if p is constant. In this case the second d rows are obtained by multiplying the first d rows by $(1-p)/p$. But what happens if p in (5.11) is not constant but instead is random? We claim that in this case $I(\alpha^*)$ is 'usually' invertible. For the moment we will not be concerned with how

our expression for $I(\alpha^*)$ was derived but instead focus upon whether or not it is invertible if p is considered as a random variable.

Let p be a random variable in $(0,1)$, $q = 1-p$, and V be $d \times 1$ a random vector satisfying $\mathbb{E}[VV']$ is positive definite. In this context we consider

$$V = \left. \frac{\partial \log f(X;\alpha)}{\partial \alpha'} \right|_{\alpha=\alpha^*}.$$

Here p and V are defined on a common probability space.

Lemma 5.1. *Let* $\Omega = \{p : \mathbb{E}[VV' \mid p] \text{ is positive definite } \}$. *Assume* $\mathbb{P}[\Omega] > 0$ *and* $Var(p \mid \Omega) > 0$ *(i.e. p is not equal to a constant for* $p \in \Omega$*). Then the* $2d \times 2d$ *matrix* $\mathbb{E}[(pV', qV')'(pV', qV')]$ *is positive definite.*

Remark: We will see (in Section 5.3) that $\mathbb{P}[\Omega] = 1$ for our test case of Gaussian AR(1) mixtures and hence the condition $Var(p \mid \Omega) > 0$ reduces to p not constant. We expect this to be a common finding. This will be discussed more in Section 5.3.

Proof. Define $Q = (pV', qV')'(pV', qV')$, a $2d \times 2d$ symmetric matrix (the expectation of this matrix is $I(\alpha^*)$ when p and V have the interpretation described above). Consider $z \in \mathbb{R}^{2d}$ and partition z as $z = (z_1', z_2')'$ where z_1 and z_2 are elements in R^d. We want to show for $z \neq 0_{2d}, z'\mathbb{E}[Q]\, z > 0$ (here 0_{2d} denotes a $2d$ dimensional vector of zeros). Because of the special form of Q we see that

$$z'\mathbb{E}[Q]\, z = \mathbb{E}[z'Qz] = \mathbb{E}\left[\left((pz_1 + qz_2)' V \right)^2 \right].$$

First we consider z_1, z_2 such that $z_1 \neq cz_2$ for any $c \in \mathbb{R}$ (i.e. z_1 and z_2 are not collinear). Then, regardless of the distribution of p we have $pz_1 + qz_2 \neq 0_d$ (a d

dimensional vector of zeros). Now

$$\begin{aligned}\mathbb{E}\left[\left((pz_1+qz_2)'V\right)^2\right] &= \mathbb{E}\left[\mathbb{E}\left[\left((pz_1+qz_2)'V\right)^2 \mid p\right]\right] \\ &\geq \mathbb{E}\left[\mathbb{E}\left[\left((pz_1+qz_2)'V\right)^2 \mid p\right]\cdot 1_\Omega\right] > 0\end{aligned}$$

because we see from the definition of Ω that the outer expectation is the integral of a positive function over a set with positive measure. So for z such that z_1 and z_2 are not collinear we have $z'\mathbb{E}[Q]z > 0$.

We finish the proof by considering the case when z_1 and z_2 are collinear, i.e. $z_2 = cz_1$ for some $c \in \mathbb{R}$. Without loss of generality assume $z_1 \neq 0_d$. Then we may write $pz_1 + qz_2 = h(p)z_1$ where $h(p) = (p + qc)$. Then we still have

$$z'\mathbb{E}[Q]z = \mathbb{E}[z'Qz] = \mathbb{E}\left[\left((pz_1+qz_2)'V\right)^2\right] = \mathbb{E}\left[\left((h(p)z_1)'V\right)^2\right]$$

Now because $\mathbb{P}[\Omega] > 0$ it must be the case that

$$\mathbb{E}[h(p)z_1'VV'z_1h(p)] \geq \mathbb{E}[\mathbb{E}[h(p)z_1'VV'z_1h(p) \mid p]\cdot 1_\Omega] > 0$$

if $\mathbb{P}[h(p) \neq 0 \mid \Omega] > 0$. At this point we consider the special case of $c = 1$, i.e. $z_1 = z_2$. Then $h(p) = 1$ for all p and it follows that $\mathbb{P}[h(p) \neq 0 \mid \Omega] > 0$. If $c \neq 1$ then the derivative of h with respect to p is $(1 - c)$, a non-zero constant and thus $Var(p \mid \Omega) > 0$ implies $\mathbb{P}[h(p) \neq 0 \mid \Omega] > 0$. Consequently, if z_1 and z_2 are collinear but not both equal to 0_d and if $Var(p \mid \Omega) > 0$, we have $z'\mathbb{E}[Q]z > 0$. When combined with our earlier result for the case of z_1 and z_2 not collinear we obtain the desired result. □

As we mentioned in the remark preceding our proof, it will often be the case that $\mathbb{P}[\Omega] = 1$ and hence the matrix will be invertible if p is not constant with probability 1. While we are hesitant to claim this holds for all logistic

mixtures of GLM time series we will, for the remainder of this chapter, assume the information matrix is invertible as long as p is not constant – thus allowing us to continue in an empirical process approach. We have illustrated our ideas using i.i.d. data because it simplifies the problem's presentation and solution. We now develop our ideas for the general logistic mixture model for time series data.

Adaptation for a Logistic Mixture of Time Series Data

Here we reintroduce our general model for logistic mixtures. Let

$$f(y_t \mid X_t; \beta, \phi) = \exp\left(\frac{y_t X_t'\beta - b\left(X_t'\beta\right)}{\phi} + c(y_t, \phi)\right) \text{ and}$$
$$\mathbb{P}\left[I_t = 1 \mid Z_t; \gamma\right] = \frac{\exp\left(Z_t'\gamma\right)}{1 + \exp\left(Z_t'\gamma\right)}.$$

The logistic mixture densities are given by

$$\begin{aligned} g(y_t \mid \mathcal{G}_{t-1}; \psi) =& g(y_t \mid X_{t1}, X_{t0}, Z_t; \beta_1, \phi_1, \beta_0, \phi_0, \gamma) \\ =& \mathbb{P}\left[I_t = 1 \mid Z_t; \gamma\right] \cdot f(y_t \mid X_{t1}; \beta_1, \phi_1) + \\ & \left(1 - \mathbb{P}\left[I_t = 1 \mid Z_t; \gamma\right]\right) \cdot f(y_t \mid X_{t0}; \beta_0, \phi_0), \\ & \text{where } \mathcal{G}_{t-1} = \sigma(X_{t1}, X_{t0}, Z_t), \text{ and } \psi = (\beta_1', \phi_1, \beta_0', \phi_0, \gamma')'. \end{aligned}$$

Our hypotheses of interest are

H_0: $Y_t \mid \mathcal{G}_{t-1} \sim f(y_t \mid \mathcal{G}_{t-1}; \beta^*, \phi^*)$ for some (β^*, ϕ^*) in the interior of $\mathcal{B} \times [\epsilon_1, M_1]$ (where $\mathcal{B}$ is a compact subset of R^{d-1} and $M_1 > \epsilon_1 > 0$) versus

H_1: $Y_t \mid \mathcal{G}_{t-1} \sim g(y_t \mid \mathcal{G}_{t-1}; \beta_1^*, \phi_1^*, \beta_0^*, \phi_0^* \gamma^*)$

where (β_1^*, ϕ_1^*) and $(\beta_0^*, \phi_0^*) \in \mathcal{B} \times [\epsilon_1, M_1]$ and Z_t and γ^* are such that $Z_t'\gamma^* = \gamma_0^* + Z_{t1}\gamma_1^* + ... + Z_{tr}\gamma_r^*$ is not constant. The $\gamma_0^*, \gamma_1^*, ..., \gamma_r^*$ are required to lie in a compact

set, Γ, in $\mathbb{R}^{r+1}$ that excludes the points given by $\{(\gamma_0, 0, ..., 0) \in \mathbb{R}^{r+1} : \gamma_0 \in \mathbb{R}\}$. The points must be excluded to ensure the regime probabilities are not constant. We let Ψ denote the set of allowable parameter points under the alternative hypothesis:

$$\Psi = \{(\beta_1, \phi_1, \beta_0, \phi_0, \gamma) \in (\mathcal{B} \times [\epsilon_1, M_1]) \times (\mathcal{B} \times [\epsilon_1, M_1]) \times \Gamma\}.$$

Furthermore, we restrict the covariates X_{t1} and X_{t0} to be identical for all t. We denote the common covariates by X_t. This allows us to say that given β and ϕ we have

$$g(y_t \mid \mathcal{G}_{t-1}; \beta_1 = \beta, \phi_1 = \phi, \beta_0 = \beta, \phi_0 = \phi, \gamma) = f(y_t \mid \mathcal{G}_{t-1}; \beta, \phi) \tag{5.14}$$

for all $\gamma \in \Gamma$. In this way all the no mixture models (i.e. one regime models) are nested within the logistic mixtures.

As we want to derive a test statistic's distribution under the null hypothesis of no mixture we assume the data Y_t have conditional distribution $f(y_t \mid \mathcal{G}_{t-1}; \beta^*, \phi^*)$ for some (β^*, ϕ^*) in the interior of $\mathcal{B} \times [\epsilon_1, M_1]$. We define Ψ^*, a subset of Ψ by

$$\Psi^* = \{(\beta^*, \phi^*, \beta^*, \phi^*, \gamma) : \gamma \in \Gamma\},$$

i.e. the collection of parameters values with the β and ϕ terms set at β^* and ϕ^* and the γ terms allowed to vary. In light of equation (5.14) we see that for $\psi \in \Psi^*$ it is the case that $g(y_t \mid \mathcal{G}_{t-1}; \psi^*) \equiv f(y_t \mid X_t; \beta^*, \phi^*)$. Sometimes we will write $f(y_t \mid \mathcal{G}_{t-1}; \beta^*, \phi^*)$ instead of $f(y_t \mid X_t; \beta^*, \phi^*)$.

To obtain asymptotic results we assume the following additional conditions are met. These conditions are analogous to those we used in Chapter 3 to show consistency of the maximum likelihood estimator.

5.2.A $\{Y_t, X_t, Z_t\}$ is stationary and ergodic with W denoting a random vector having the joint stationary distribution. By this we mean $Y_t, X_t,$ and Z_t obey a strong law of large numbers in the sense that if $h(\cdot)$ is a measurable and integrable function of W then $\frac{1}{T}\sum h(Y_t, X_t, Z_t) \xrightarrow{a.s.} \mathbb{E}[h(W)]$. We denote the associated components of W as W_Y, W_X, and W_Z. We will sometimes write $g(W_Y \mid W_X, W_Z; \psi)$ as $g(W; \psi)$. This condition was discussed at length in Chapter 3.

5.2.B $\mathbb{E}[\log g(W; \psi)] < \infty$ for all $\psi \in \Psi$ and is continuous in ψ.

5.2.C For any $\psi \in \Psi \backslash \Psi^*$(elements in Ψ but not in Ψ^*) and $\tilde{\psi} \in \Psi^*$ we assume $\mathbb{E}[\log g(W; \psi)] < \mathbb{E}\left[\log g(W; \tilde{\psi})\right]$. From equation (5.14) it is clear that $\mathbb{E}\left[\log g(W; \tilde{\psi})\right]$ is constant for all $\tilde{\psi} \in \Psi^*$. This constant value is

$$\mathbb{E}[\log f(W_Y \mid W_X; \beta^*, \phi^*)].$$

Here $f(W_Y \mid W_X; \beta^*, \phi^*)$ is $f(y_t \mid \mathcal{G}_{t-1}; \psi)$ with y_t and X_t replaced by W_Y and W_X. This is analogous to the identifiability condition 3.1.C in Chapter 3 but is modified to account for the fact that γ is not identifiable under the null hypothesis.

5.2.D Given $B_\rho(\psi) = \{\psi' \in K : \ ||\psi' - \psi|| < \rho\}$ we define

$$g^*(W; \psi, \rho) = \sup_{\psi' \in B_\rho(\psi)} g(W; \psi').$$

We assume that for any $\psi \in \Psi$ $\mathbb{E}[\log g^*(W; \psi, \rho)]$ exists for ρ sufficiently small and

$\lim_{\rho \to 0} \mathbb{E}[\log g^*(W; \psi, \rho)] = \mathbb{E}[\log g(W; \psi)]$ for all $\psi \in \Psi$.

As in the case of i.i.d. data we seek to use profile likelihood techniques – we first fix $\gamma \in \Gamma$ and compute the log-likelihood ratio statistic at this point. We begin

by setting

$$\theta = (\beta_1', \phi_1, \beta_0', \phi_0)' \tag{5.15}$$

$$\theta^* = (\beta^{*\prime}, \phi^*, \beta^{*\prime}, \phi^*)' \tag{5.16}$$

$$\hat{\theta}(\gamma) = \text{ Arg } \max_{\theta} L_T^g(\gamma, \theta) \text{ where} \tag{5.17}$$

$$L_T^g(\gamma, \theta) = \sum \log g(y_t \mid \mathcal{G}_{t-1}; \gamma, \theta) \tag{5.18}$$

$$\theta(\gamma) = \text{ Arg } \max_{\theta} \mathbb{E}\left[\log g(W_Y \mid W_X, W_Z; \gamma, \theta)\right]. \tag{5.19}$$

In the notation of the previous section $\alpha = (\beta', \phi)'$. With these definitions we may prove the following:

Lemma 5.3. *Under Conditions* 5.2.A – 5.2.D *we have* $\hat{\theta}(\gamma) \xrightarrow{a.s} \theta(\gamma)$. *Furthermore,* $\theta(\gamma) = (\beta^*, \phi^*, \beta^*, \phi^*)$ *for all* $\gamma \in \Gamma$.

Proof. There are two statements to prove. The first assertion is that $\hat{\theta}(\gamma) \xrightarrow{a.s} \theta(\gamma)$. A comparison of the conditions in this section with those used in Section 3.1 to prove Theorem 3.2 show we can apply the results of that theorem to our case and obtain the desired conclusion. All that is necessary is noting that ψ in Section 3.1 corresponds to θ in this section, K corresponds to $(\mathcal{B} \times [\epsilon_1, M_1]) \times (\mathcal{B} \times [\epsilon_1, M_1])$, and ψ^* corresponds to θ^*.

The second statement to prove is $\theta(\gamma) = (\beta^{*\prime}, \phi^*, \beta^{*\prime}, \phi^*)$ for all $\gamma \in \Gamma$. This statement follows from the identifiability condition (5.2.C) and the strong law of large numbers (Condition 5.2.A). It may be formally proved by following the structure of the proof in Section 3.1. □

Now we try to explicitly find $L_T^g(\gamma,\hat{\theta}(\gamma)) - L_T^g(\gamma,\theta^*)$. Using a Taylor series expansion about θ^* we have

$$L_T^g(\gamma,\theta) - L_T^g(\gamma,\theta^*) = \left.\frac{\partial L_T^g(\gamma,\theta)}{\partial\theta'}\right|_{\theta^*}(\theta-\theta^*) + \frac{1}{2}(\theta-\theta^*)'\left.\frac{\partial^2 L_T^g(\gamma,\theta)}{\partial\theta\partial\theta'}\right|_{\bar{\theta}}(\theta-\theta^*) \tag{5.20}$$

with $\bar{\theta}$ lies on the chord between θ^* and θ. If we maximize the right-hand side to find $\hat{\theta}(\gamma)$ then using calculus and assuming the matrix of second partials is invertible we obtain

$$\left(\hat{\theta}(\gamma)-\theta^*\right) = -\left[\left.\frac{\partial^2 L_T^g(\gamma,\theta)}{\partial\theta\partial\theta'}\right|_{\bar{\theta}}\right]^{-1}\left.\frac{\partial L_T^g(\gamma,\theta)}{\partial\theta'}\right|_{\theta^*}. \tag{5.21}$$

Upon substituting this back into (5.20) we see

$$L_T^g(\gamma,\hat{\theta}(\gamma)) - L_T^g(\gamma,\theta^*) =$$
$$-\frac{1}{2}\frac{1}{\sqrt{T}}\left.\frac{\partial L_T^g(\gamma,\theta)}{\partial\theta}\right|_{\theta^*}\left[\frac{1}{T}\left.\frac{\partial^2 L_T^g(\gamma,\theta)}{\partial\theta\partial\theta'}\right|_{\bar{\theta}}\right]^{-1}\frac{1}{\sqrt{T}}\left.\frac{\partial L_T^g(\gamma,\theta)}{\partial\theta'}\right|_{\theta^*} \tag{5.22}$$

Here we add more conditions to our model. Almost all of these conditions can be established if $\log g(W;\psi)$ is sufficiently smooth with respect to ψ and has derivatives that may be bounded by integrable functions.

5.2.E *For all* $\gamma\in\Gamma$

$$\mathbb{E}\left[\left.\frac{\partial\log g(W;\gamma,\theta)}{\partial\theta'}\frac{\partial\log g(W;\gamma,\theta)}{\partial\theta}\right|_{\theta^*}\right] = -\mathbb{E}\left[\left.\frac{\partial^2\log g(W;\gamma,\theta)}{\partial\theta\partial\theta'}\right|_{\theta^*}\right] \textit{ exists.}$$

Also we assume the matrix is a continuous function of γ *and* θ *and invertible for* $\theta=\theta^*$. *Much of this chapter concerns the invertibility of these matrices.*

5.2.F *From condition* 5.2.A *we have that for any* γ *and* θ *such that* $(\theta,\gamma)=\psi$ *is an element of* Ψ

$$\left[\frac{1}{T}\frac{\partial^2 L_T^g(\gamma,\theta)}{\partial\theta\partial\theta'}\right] \xrightarrow{a.s} \mathbb{E}\left[\frac{\partial^2\log g(W;\gamma,\theta)}{\partial\theta\partial\theta'}\right].$$

Furthermore, this convergence is uniform *for* $\psi \in \Psi$*, i.e.*

$$\sup_{\psi \in \Psi} \left\| \frac{1}{T} \frac{\partial^2 L_T^g(\gamma,\theta)}{\partial\theta\partial\theta'} - \mathbb{E}\left[\frac{\partial^2 \log g(W;\gamma,\theta)}{\partial\theta\partial\theta'} \right] \right\| \xrightarrow{a.s.} \underline{0}.$$

5.2.G $\left. \frac{1}{\sqrt{T}} \frac{\partial L_T^g(\gamma,\theta)}{\partial\theta} \right|_{\theta^*}$ *is* uniformly *bounded in probability, i.e.* $\sup_{\gamma\in\Gamma} \left. \frac{1}{\sqrt{T}} \frac{\partial L_T^g(\gamma,\theta)}{\partial\theta} \right|_{\theta^*}$ *is* $O_p(1)$.

Now using these conditions and (5.22) we can derive the following series of equations:

$$2(L_T^g(\gamma,\hat{\theta}(\gamma)) - L_T^g(\gamma,\theta^*)) = \tag{5.23}$$

$$- \frac{1}{\sqrt{T}} \frac{\partial L_T^g(\gamma,\theta)}{\partial\theta}\bigg|_{\theta^*} \left[\frac{1}{T} \frac{\partial^2 L_T^g(\gamma,\theta)}{\partial\theta\partial\theta'}\bigg|_{\bar{\theta}} \right]^{-1} \frac{1}{\sqrt{T}} \frac{\partial L_T^g(\gamma,\theta)}{\partial\theta'}\bigg|_{\theta^*} = \tag{5.24}$$

$$\frac{1}{T} \frac{\partial L_T^g(\gamma,\theta)}{\partial\theta}\bigg|_{\theta^*} \left[\mathbb{E}\, \frac{\partial \log g(W;\gamma,\theta)}{\partial\theta'} \frac{\partial \log g(W;\gamma,\theta)}{\partial\theta}\bigg|_{\theta^*} \right]^{-1} \frac{\partial L_T^g(\gamma,\theta)}{\partial\theta'}\bigg|_{\theta^*} + o_p(1). \tag{5.25}$$

Conditions 5.2.F and 5.2.G allow us to conclude that the $o_p(1)$ term in (5.25) is uniform for $\gamma \in \Gamma$. This will be important for us below. We demonstrate this uniformity as follows:

$$\begin{aligned}
&\sup_\gamma \left\| \tfrac{1}{T} \tfrac{\partial L_T^g(\gamma,\theta)}{\partial\theta}\big|_{\theta^*} \left[\left[\tfrac{1}{T} \tfrac{\partial^2 L_T^g(\gamma,\theta)}{\partial\theta\partial\theta'}\big|_{\bar{\theta}} \right]^{-1} - \left[\mathbb{E}\, \tfrac{\partial \log g(W;\gamma,\theta)}{\partial\theta'} \tfrac{\partial \log g(W;\gamma,\theta)}{\partial\theta}\big|_{\theta^*} \right]^{-1} \right] \tfrac{\partial L_T^g(\gamma,\theta)}{\partial\theta'}\big|_{\theta^*} \right\| \\
&\leq \sup_\gamma \left\| \tfrac{1}{\sqrt{T}} \tfrac{\partial L_T^g(\gamma,\theta)}{\partial\theta}\big|_{\theta^*} \right\|^2 \sup_\gamma \left\| \left[\left[\tfrac{1}{T} \tfrac{\partial^2 L_T^g(\gamma,\theta)}{\partial\theta\partial\theta'}\big|_{\bar{\theta}} \right]^{-1} - \left[\mathbb{E}\, \tfrac{\partial \log g(W;\gamma,\theta)}{\partial\theta'} \tfrac{\partial \log g(W;\gamma,\theta)}{\partial\theta}\big|_{\theta^*} \right]^{-1} \right] \right\| \\
&= O_p(1) \cdot o_p(1) = o_p(1)
\end{aligned}$$

where $\|\cdot\|$ denote either vector or matrix Euclidean norms. Now, keeping in mind our condition that we note that $\{Y_t, X_t, Z_t\}$ have the common stationary distribution given by W we want to verify that

$$\frac{1}{\sqrt{T}} \frac{\partial L_T^g(\gamma,\theta)}{\partial\theta'}\bigg|_{\theta^*} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \frac{\partial \log g(y_t \mid \mathcal{G}_{t-1};\gamma,\theta)}{\partial\theta'}\bigg|_{\theta^*} \tag{5.26}$$

is a martingale. This will be implied by the following additional condition:

5.2.H *$g(W_Y, W_X, W_Z; \psi)$ is three times continuously differentiable with respect to ψ. Furthermore there exist integrable functions $F_1(W)$ and $F_2(W)$ such that for all $r, s \in \{1, 2, ..., q\}$,*

$$F_1(W) > \left| \frac{\partial g(W;\psi)}{\partial \psi_r} \right|, F_2(W) > \left| \frac{\partial^2 g(W;\psi)}{\partial \psi_r \partial \psi_s} \right|, \text{ and}$$

$$\mathbb{E}[F_1(W) \mid W_X, W_Z] < \infty, \mathbb{E}[F_2(W) \mid W_X, W_Z] < \infty.$$

This condition implies the first part of Condition 5.2.E.

To show how this implies $\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left. \frac{\partial \log g(y_t \mid \mathcal{G}_{t-1}; \gamma, \theta)}{\partial \theta} \right|_{\theta^*}$ is a martingale the critical point is to establish

$$\mathbb{E}\left[\left. \frac{\partial \log g(y_t \mid \mathcal{G}_{t-1}; \gamma, \theta)}{\partial \theta'} \right|_{\theta^*} \middle| \mathcal{G}_{t-1} \right] = \underline{0}.$$

$$\begin{aligned} \text{The left-hand side above } &= \int \left. \frac{\partial \log g(y_t \mid \mathcal{G}_{t-1}; \gamma, \theta)}{\partial \theta'} \right|_{\theta^*} g(y_t \mid \mathcal{G}_{t-1}; \gamma, \theta^*) dy_t \\ &= \int \left. \frac{\partial g(y_t \mid \mathcal{G}_{t-1}; \gamma, \theta)}{\partial \theta'} \right|_{\theta^*} dy_t \\ &= \frac{\partial}{\partial \theta'} \left(\int g(y_t \mid \mathcal{G}_{t-1}; \gamma, \theta) \right) \Bigg|_{\theta^*} = \underline{0}. \end{aligned}$$

As the interchange of differentiation and integration is justified by Condition 5.2.H, the martingale property is established. From here we can apply the martingale central limit theorem (Theorem 3.14) to show

$$\begin{aligned} &\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left. \frac{\partial \log g(y_t \mid \mathcal{G}_{t-1}; \gamma, \theta)}{\partial \theta'} \right|_{\theta^*} \xrightarrow{\mathcal{D}} \\ &N\left(\underline{0}, \mathbb{E}\left[\left. \frac{\partial \log g(W; \gamma, \theta)}{\partial \theta'} \frac{\partial \log g(W; \gamma, \theta)}{\partial \theta} \right|_{\theta^*} \right] \right). \end{aligned} \tag{5.27}$$

To demonstrate how to use Theorem 3.14 we first establish the following lemma:

Lemma 5.4. *Let Q_t be a real-valued stationary martingale difference sequence such that $\mathbb{E}[Q_t^2] = v < \infty$ and $\frac{1}{T} \sum Q_t^2 \xrightarrow{a.s.} v$. Then*

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} Q_t \xrightarrow{\mathcal{D}} N(0, v).$$

Proof. The proof is nearly immediate from Theorem 3.14. We define $D_{T,t} = \frac{1}{\sqrt{Tv}}Q_t$. Then the conditions in Theorem 3.14 are satisfied. □

With Lemma 5.4 we can demonstrate the convergence in 5.27 with an application of the Cramér-Wold device. From this result we may reexamine (5.25) and conclude

$$2(L_T^g(\gamma, \hat{\theta}(\gamma)) - L_T^g(\gamma, \theta^*)) \xrightarrow{\mathcal{D}} \chi^2_{2d} \tag{5.28}$$

where d is the dimension of $(\beta', \phi)'$.

At this point we define our log likelihood ratio statistic as

$$LR_T \stackrel{\triangle}{=} \sup_{\psi \in \Psi} L_T^g(\psi) - \sup_{\beta,\phi} L_T^f(\beta, \phi) \text{ where} \tag{5.29}$$
$$L_T^f(\beta, \phi) = \sum_{t=1}^{T} \log f(y_t \mid \mathcal{G}_{t-1}; \beta, \phi).$$

In order to tie this statistic to what we have developed thus far we note for $\left(\hat{\beta}, \hat{\phi}\right) =$ Arg $\max_{\beta,\phi} L_T^f(\beta, \phi)$ we see that

$$2LR_T = 2\left(\sup_{\gamma} L_T^g(\gamma, \hat{\theta}(\gamma)) - L_T^f(\hat{\beta}, \hat{\phi})\right) \tag{5.30}$$
$$= 2\left(\sup_{\gamma} L_T^g(\gamma, \hat{\theta}(\gamma)) - L_T^g(\gamma, \theta^*)\right) - 2\left(L_T^f(\hat{\beta}, \hat{\phi}) - L_T^f(\beta^*, \phi^*)\right). \tag{5.31}$$

The last equality above holds because we showed in Lemma 5.3 that $\theta(\gamma) = \theta^* = (\beta^{*\prime}, \phi^*, \beta^{*\prime}, \phi^*)'$ and so $L_T^g(\gamma, \theta(\gamma)) = L_T^f(\beta^*, \phi^*)$ as implied by (5.14). Using the

same Taylor series techniques we can show that under the null hypothesis

$$2\left(L_T^f(\hat{\beta},\hat{\phi}) - L_T^f(\beta^*,\phi^*)\right) = o_p(1) + \tag{5.32}$$

$$\rho'\left[\mathbb{E}\left[\begin{array}{cccc} \frac{\partial \log f(y_t|\mathcal{G}_{t-1};\beta,\phi)}{\partial \beta'} & \frac{\partial \log f(y_t|\mathcal{G}_{t-1};\beta,\phi)}{\partial \beta} & \frac{\partial \log f(y_t|\mathcal{G}_{t-1};\beta,\phi)}{\partial \beta'} & \frac{\partial \log f(y_t|\mathcal{G}_{t-1};\beta,\phi)}{\partial \phi} \\ \frac{\partial \log f(y_t|\mathcal{G}_{t-1};\beta,\phi)}{\partial \beta} & \frac{\partial \log f(y_t|\mathcal{G}_{t-1};\beta,\phi)}{\partial \phi} & \frac{\partial \log f(y_t|\mathcal{G}_{t-1};\beta,\phi)}{\partial \phi} & \frac{\partial \log f(y_t|\mathcal{G}_{t-1};\beta,\phi)}{\partial \phi} \end{array}\right]\right]^{-1}\rho \tag{5.33}$$

$$\xrightarrow{\mathcal{D}} \chi^2_d, \text{ where } \rho = \left[\begin{array}{c} \frac{1}{\sqrt{T}}\sum \frac{\partial \log f(y_t|\mathcal{G}_{t-1};\beta,\phi)}{\partial \beta'} \\ \frac{1}{\sqrt{T}}\sum \frac{\partial \log f(y_t|\mathcal{G}_{t-1};\beta,\phi)}{\partial \phi} \end{array}\right] \tag{5.34}$$

and all derivatives are evaluated at (β^*,ϕ^*). The chi-square r.v. in this case has d instead of $2d$ degrees of freedom (the matrix in (5.33) is $d \times d$). So while we know

$$2\left(L_T^f(\hat{\beta},\hat{\phi}) - L_T^f(\beta^*,\phi^*)\right) \xrightarrow{D} \chi^2_d \tag{5.35}$$

and

$$2\left(L_T^g(\gamma,\hat{\theta}(\gamma)) - L_T^g(\gamma,\theta^*)\right) \xrightarrow{D} \chi^2_{2d} \tag{5.36}$$

for fixed γ, finding the asymptotic distribution of the right-hand side of (5.31) is considerably more complicated as the chi-square r.v.s in (5.35) and (5.36) are correlated and a supremum is involved. To proceed further we appeal to the

theory of empirical processes. We begin by defining a $3d$ dimensional vector

$$s_t(\gamma) = \begin{bmatrix} \left.\frac{\partial \log g(y_t|\mathcal{G}_{t-1};\gamma,\theta)}{\partial \beta_1'}\right|_{\theta^*} \\ \left.\frac{\partial \log g(y_t|\mathcal{G}_{t-1};\gamma,\theta)}{\partial \phi_1}\right|_{\theta^*} \\ \left.\frac{\partial \log g(y_t|\mathcal{G}_{t-1};\gamma,\theta)}{\partial \beta_0'}\right|_{\theta^*} \\ \left.\frac{\partial \log g(y_t|\mathcal{G}_{t-1};\gamma,\theta)}{\partial \phi_0}\right|_{\theta^*} \\ \left.\frac{\partial \log f(y_t|\mathcal{G}_{t-1};\beta,\phi)}{\partial \beta'}\right|_{\beta^*,\phi^*} \\ \left.\frac{\partial \log f(y_t|\mathcal{G}_{t-1};\beta,\phi)}{\partial \phi}\right|_{\beta^*,\phi^*} \end{bmatrix} = \begin{bmatrix} p_t(\gamma)\left(\frac{y_t X_t - \dot{b}(X_t'\beta^*)}{\phi^*}\right) \\ p_t(\gamma)\left(\frac{(y_t X_t'\beta^* - b(X_t'\beta^*))}{(\phi^*)^2} + \left.\frac{\partial c(y_t,\phi)}{\partial \phi}\right|_{\phi^*}\right) \\ (1-p_t(\gamma))\left(\frac{y_t X_t - \dot{b}(X_t'\beta^*)}{\phi^*}\right) \\ (1-p_t(\gamma))\left(\frac{(y_t X_t'\beta^* - b(X_t'\beta^*))}{(\phi^*)^2} + \left.\frac{\partial c(y_t,\phi)}{\partial \phi}\right|_{\phi^*}\right) \\ \left(\frac{y_t X_t - \dot{b}(X_t'\beta^*)}{\phi^*}\right) \\ \left(\frac{(y_t X_t'\beta^* - b(X_t'\beta^*))}{(\phi^*)^2} + \left.\frac{\partial c(y_t,\phi)}{\partial \phi}\right|_{\phi^*}\right) \end{bmatrix} \tag{5.37}$$

where $\dot{b}(X_t'\beta^*)$ denotes the derivative of $b(X_t'\beta^*)$ with respect to β evaluated at β^*. As was partially demonstrated above, we can show that $s_t(\gamma)$ is a martingale difference sequence (above we showed only that the first $2d$ elements of $s_t(\gamma)$ was a martingale difference). We can extend these ideas to the whole of the $3d$ dimensional vector, $s_t(\gamma)$. In light of Condition 5.2.A we see these martingale differences have a common stationary distribution. From here we apply our usual martingale central limit theorem (combined with the Cramér-Wold device) to obtain

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} s_t(\gamma) \xrightarrow{\mathcal{D}} N\left(\underline{0}, \mathbb{E}[s_t(\gamma)s_t(\gamma)']\right) \tag{5.38}$$

for any $\gamma \in \Gamma$. However, we will need a stronger statement regarding the convergence of finite dimensional distributions. Let $\gamma^1, \gamma^2, \ldots, \gamma^L$ be elements of Γ. We want to show

$$\frac{1}{\sqrt{T}}\begin{pmatrix} \sum s_t(\gamma^1) \\ \vdots \\ \sum s_t(\gamma^L) \end{pmatrix} \xrightarrow{\mathcal{D}} N\left(\underline{0}, \begin{pmatrix} K(\gamma^1,\gamma^1), \ldots, K(\gamma^1,\gamma^L) \\ \vdots \\ K(\gamma^L,\gamma^1), \ldots, K(\gamma^L,\gamma^L) \end{pmatrix}\right) \tag{5.39}$$

where is the $K(\gamma^i, \gamma^j)$ is the $3d \times 3d$ matrix, $\mathbb{E}[s_t(\gamma^i)s_t(\gamma^j)]$. To show this we use the Cramér-Wold device along with Lemma 5.4. Consequently, convergence of finite dimensional distributions to a multivariate normal has been established.

Now suppose we can show the following condition is met (and we will show this later for our special mixture of Gaussian AR(1) processes)

5.2.I $\frac{1}{\sqrt{T}} \sum_{t=1}^{T} s_t(\gamma)$ is a stochastically equicontinuous sequence of functions in $\mathcal{C}^{3d}[\Gamma]$, the space of continuous functions h such that $h: \Gamma \to \mathbb{R}^{3d}$.

By definition (see Andrews (1993)) this would mean that for every $\epsilon > 0$ there exists a δ such that

$$\overline{\lim_{T\to\infty}} \mathbb{P}\left[\sup_{\substack{\gamma^1,\gamma^2\in\Gamma: \\ \|\gamma^1-\gamma^2\|<\delta}} \frac{1}{\sqrt{T}} \left\| \sum^{T} s_t(\gamma^1) - \sum^{T} s_t(\gamma^2) \right\| > \epsilon \right] < \epsilon.$$

With the finite dimensional convergence of (5.39), compactness of Γ, and the stochastic equicontinuity condition above we can conclude via an empirical processes theorem (e.g. Theorem 10.2 in Pollard (1990)) that there exists a unique stochastic process ,$S(\cdot)$, taking values in $\mathcal{C}^{3d}[\Gamma]$ (with probability 1) such that

1). For fixed γ, $S(\gamma) \sim N(\underline{0}, \mathbb{E}[s_t(\gamma)s_t(\gamma)'])$, and

2). If l is a continuous functional such that $l: \mathcal{C}^{3d}[\Gamma] \to \mathbb{R}$ then

$$l\left(\left\{\frac{1}{\sqrt{T}}\sum_{t}^{T} s_t(\gamma) : \gamma \in \Gamma\right\}\right) \xrightarrow{D} l(\{S(\gamma) : \gamma \in \Gamma\}). \tag{5.40}$$

With this result we now try to find a functional l that corresponds to our expression for $2LR_T$ in (5.31). Then we examine the distribution of $l(\{S(\gamma) : \gamma \in \Gamma\})$ and use this for our asymptotic distribution of $2LR_T$.

It is relatively easy to find an appropriate functional for l. Let us define $V(\gamma) = \mathbb{E}[s_t(\gamma)s_t(\gamma)']$, the $3d \times 3d$ covariance matrix appearing in (5.38), $V^{2d}(\gamma)$,

the $2d \times 2d$ upper left corner of $V(\gamma)$, and V^d the $d \times d$ lower right corner of $V(\gamma)$. These last two matrices correspond to the asymptotic covariance matrices of

$$\frac{1}{\sqrt{T}} \sum \left. \frac{\partial \log g(y_t \mid \mathcal{G}_{t-1}; \gamma, \theta)}{\partial \theta'} \right|_{\theta^*} \quad \text{and} \quad \begin{bmatrix} \frac{1}{\sqrt{T}} \sum \left. \frac{\partial \log f(y_t | \mathcal{G}_{t-1}; \beta, \phi)}{\partial \beta'} \right|_{\beta^*, \phi^*} \\ \frac{1}{\sqrt{T}} \sum \left. \frac{\partial \log f(y_t | \mathcal{G}_{t-1}; \beta, \phi)}{\partial \phi} \right|_{\beta^*, \phi^*} \end{bmatrix}.$$

Inspection of $s_t(\gamma)$ in (5.37) shows that V^d does not depend upon γ. Now let $h(\gamma)$ be an element in $\mathcal{C}^{3d}[\Gamma]$. We may consider $h(\gamma)$ a $3d$ dimensional vector and partition $h(\gamma)$ as $h(\gamma) = (h^{2d}(\gamma)', h^d(\gamma)')'$ where the superscripts denote the associated length of the vectors. Now we define

$$l(h) = \sup_{\gamma \in \Gamma} \left[h^{2d}(\gamma)' \left(V^{2d}(\gamma) \right)^{-1} h^{2d}(\gamma) - h^d(\gamma)' \left(V^d \right)^{-1} h^d(\gamma) \right]. \tag{5.41}$$

When we examine equations (5.25), (5.33), and (5.31) we see that

$$2LR_T = 2 \left(\sup_{\gamma} L_T^g(\gamma, \hat{\theta}(\gamma)) - L_T^f(\hat{\beta}, \hat{\phi}) \right) = l \left(\left\{ \frac{1}{\sqrt{T}} \sum_{t=1}^{T} s_t(\gamma) \colon \gamma \in \Gamma \right\} \right) + o_p(1). \tag{5.42}$$

It is a subtle but important point that the $o_p(1)$ term above arises from the condition that the $o_p(1)$ term in (5.25) is uniform for $\gamma \in \Gamma$. From the relation in (5.40) we conclude

$$2LR_T \xrightarrow{D} l\left(\{ S(\gamma) \colon \gamma \in \Gamma \} \right) \tag{5.43}$$

where $\frac{1}{\sqrt{T}} \sum s_t(\cdot)$ converges weakly to $S(\cdot)$. To determine the process $S(\cdot)$ we know it is marginally Gaussian, i.e. $S(\gamma) \sim N\left(\underline{0}, V(\gamma)\right)$. We obtain this by considering the functional $l_\gamma(h(\cdot)) = h(\gamma)$ and noting that

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} s_t(\gamma) \xrightarrow{D} N\left(\underline{0}, V(\gamma)\right).$$

As a Gaussian process is completely determined by its mean and covariance structure we only need find the right covariance structure. Though we do not present the details it is clear that by considering different types of functionals it must be the case that $\mathbb{E}[S(\gamma^1)S(\gamma^2)'] = \mathbb{E}[s_t(\gamma^1)s_t(\gamma^2)']$. For each pair (γ^1, γ^2) let us define $K(\gamma^1, \gamma^2) = \mathbb{E}[s_t(\gamma^1)s_t(\gamma^2)']$. Then the mean zero Gaussian process with covariance kernel $K(\gamma^1, \gamma^2)$ must be the unique process $S(\gamma)$ satisfying (5.43).

Now we are in a position to explore the distribution of $l(\{S(\gamma) : \gamma \in \Gamma\})$. For the moment suppose we know the elements of $K(\gamma^1, \gamma^2)$ for every (γ^1, γ^2). Then we could use a random number generator to create independent realizations of the $S(\cdot)$ process and compute $l(S(\cdot))$ for the different realizations. From this sample we can find an empirical distribution of $l(S(\cdot))$ to use for approximating the distribution of $2LR_T$.

Of course in practice we do not know $K(\gamma^1, \gamma^2)$ but we may estimate it from the data in an obvious manner. We define

$$\hat{K}_T\left(\gamma^1, \gamma^2\right) = \frac{1}{T}\sum_{t=1}^{T} s_t\left(\gamma^1; \hat{\beta}, \hat{\phi}\right) s_t\left(\gamma^2; \hat{\beta}, \hat{\phi}\right)$$

and from Condition 5.2.A and Theorem 3.18 it is easy to show $\hat{K}_T\left(\gamma^1, \gamma^2\right)$ converges uniformly, with probability 1, to $K\left(\gamma^1, \gamma^2\right)$. From $\hat{K}_T\left(\gamma^1, \gamma^2\right)$ we can generate realizations of $\hat{S}_T(\gamma)$ processes where the $\hat{S}_T(\gamma)$ processes are defined so that

1. $\hat{S}_T(\gamma)$ has a $N\left(\underline{0}, \hat{K}_T(\gamma, \gamma)\right)$ distribution and,
2. $\mathbb{E}\left[\hat{S}_T\left(\gamma^1\right)\hat{S}_T\left(\gamma^2\right)\right] = \hat{K}_T\left(\gamma^1, \gamma^2\right)$.

Next we define a new functional $\hat{l}_T$ (because the l functional depends on unknown parameters as well):

$$\hat{l}_T(h(\cdot)) = \sup_{\gamma\in\Gamma}\left[h^{2d}(\gamma)'\left(\hat{K}_T^{2d}(\gamma,\gamma)\right)^{-1}h^{2d}(\gamma) - h^{d}(\gamma)'\left(\hat{K}_T^{d}(\gamma,\gamma)\right)^{-1}h^{d}(\gamma)\right]$$

where $h(\gamma) = (h^{2d}(\gamma)', h^d(\gamma)')'$ is an element of $C^{3d}[\Gamma]$. With these definitions we approximate the functional

$$l(S(\cdot)) = \sup_{\gamma\in\Gamma}\left\{S^{2d}(\gamma)'\left(K^{2d}(\gamma,\gamma)\right)^{-1}S^{2d}(\gamma) - S^{d}(\gamma)'\left(K^{d}(\gamma,\gamma)\right)^{-1}S^{d}(\gamma)\right\} \text{ by}$$
$$\hat{l}_T\left(\hat{S}_T(\cdot)\right) = \sup_{\gamma\in\Gamma}\left\{\hat{S}_T^{2d}(\gamma)'\left(\hat{K}_T^{2d}(\gamma,\gamma)\right)^{-1}\hat{S}_T^{2d}(\gamma) - \hat{S}_T^{d}(\gamma)'\left(\hat{K}_T^{d}(\gamma,\gamma)\right)^{-1}\hat{S}_T^{d}(\gamma)\right\}.$$

As our test relies on this approximation we must show $\hat{l}_T\left(\hat{S}_T(\cdot)\right) \xrightarrow{\mathcal{D}} l(S(\cdot))$.

Theorem 5.5. *Let $\hat{S}_T(\gamma)$ be a stochastically equicontinuous sequence in $C^{3d}[\Gamma]$ with $\hat{S}_T(\cdot)$ converging weakly to $S(\cdot)$. Also, suppose $\hat{K}_T^{2d}(\gamma,\gamma) \xrightarrow{a.s.} K^{2d}(\gamma,\gamma)$ and $\hat{K}_T^{d}(\gamma,\gamma) \xrightarrow{a.s.} K^{d}(\gamma,\gamma)$ where convergence is uniform in Γ and both $K^{2d}(\gamma,\gamma)$ and $K^d(\gamma,\gamma)$ are nonstochastic invertible matrices. Then $\hat{l}_T\left(\hat{S}_T(\cdot)\right) \xrightarrow{\mathcal{D}} l(S(\cdot))$.*

Proof. We sketch the proof as follows. Let $h(\gamma)$ be an element in $\mathcal{C}^{3d}[\Gamma]$ and partition $h(\gamma)$ as $h(\gamma) = (h^{2d}(\gamma)', h^d(\gamma)')'$. Now we define the transformations

$$\begin{aligned} Q(h(\gamma)) =& h^{2d}(\gamma)'\left(K^{2d}(\gamma,\gamma)\right)^{-1}h^{2d}(\gamma) - h^{d}(\gamma)'\left(K^{d}(\gamma,\gamma)\right)^{-1}h^{d}(\gamma) \text{ and} \\ \hat{Q}_T(h(\gamma)) =& h^{2d}(\gamma)'\left(\hat{K}_T^{2d}(\gamma,\gamma)\right)^{-1}h^{2d}(\gamma) - h^{d}(\gamma)'\left(\hat{K}_T^{d}(\gamma,\gamma)\right)^{-1}h^{d}(\gamma). \end{aligned}$$

Then we may write

$$\hat{Q}_T(\hat{S}_T(\gamma)) = Q(\hat{S}_T(\gamma)) + \left[Q_T(\hat{S}_T(\gamma)) - Q(\hat{S}_T(\gamma))\right]. \tag{5.44}$$

Now because $\hat{S}_T(\cdot)$ converges weakly to $S(\cdot)$ and $K^{2d}(\gamma,\gamma)$ and $K^d(\gamma,\gamma)$ are nonstochastic and invertible it is clear that $Q(\hat{S}_T(\cdot))$ converges weakly to $Q(S(\cdot))$

and thus

$$\sup_{\gamma} Q(\hat{S}_T(\gamma)) \xrightarrow{D} \sup_{\gamma} Q(S(\gamma)). \tag{5.45}$$

Next, we want to show the term in brackets in (5.44) is $o_p(1)$, uniformly for $\gamma \in \Gamma$. To prove this we see that

$$\sup_{\gamma} \left| \left[Q_T(\hat{S}_T(\gamma)) - Q(\hat{S}_T(\gamma)) \right] \right| \leq \left(\sup_{\gamma} \left\| \hat{K}_T^{2d}(\gamma,\gamma) - K^{2d}(\gamma,\gamma) \right\| + \right.$$
$$\left. \sup_{\gamma} \left\| \hat{K}_T^{d}(\gamma,\gamma) - K^{d}(\gamma,\gamma) \right\| \right) \cdot \sup_{\gamma} \hat{S}_T(\gamma)' \hat{S}_T(\gamma).$$

From our uniform convergence conditions regarding $\hat{K}_T(\cdot,\cdot)$ and $K(\cdot,\cdot)$ we see the term in parentheses is $o_p(1)$. Because $\hat{S}_T(\gamma))$ is stochastically equicontinuous we may conclude that $\sup_{\gamma} \hat{S}_T(\gamma)' \hat{S}_T(\gamma)$ is $O_p(1)$ if $\hat{S}_T(\tilde{\gamma})' \hat{S}_T(\tilde{\gamma})$ is integrable for some $\tilde{\gamma} \in \Gamma$. The proof of this last assertion is similar to the proof of Lemma 5.6 in the next section so we omit it here. The integrability condition is clearly met for all γ because $\hat{S}_T(\gamma)$ has a multivariate normal distribution. Consequently we obtain

$$\sup_{\gamma} \left| \left[Q_T(\hat{S}_T(\gamma)) - Q(\hat{S}_T(\gamma)) \right] \right| < O_p(1) \cdot o_p(1) = o_p(1).$$

In light of these results we can rewrite 5.44 to say

$$\begin{aligned} \sup_{\gamma} \hat{Q}_T(\hat{S}_T(\gamma)) &= \sup_{\gamma} \{ Q(\hat{S}_T(\gamma)) + o_p(1,\gamma) \} \\ &= \sup_{\gamma} \{ Q(\hat{S}_T(\gamma)) \} + o_p(1). \end{aligned} \tag{5.46}$$

where $o_p(1,\gamma)$ corresponds to the bracketed term in (5.44). This implies

$$\hat{l}_T(\hat{S}_T(\cdot)) \equiv \sup_{\gamma} \hat{Q}_T(\hat{S}_T(\gamma)) = \sup_{\gamma} Q(\hat{S}_T(\gamma)) + o_p(1) \xrightarrow{\mathcal{D}} \sup_{\gamma} Q(S(\gamma)) \equiv l(S(\cdot))$$

and our proof is complete. □

We will illustrate this procedure in the last chapter but first we examine the conditions of our test in the case of logistic mixtures of normally distributed AR(1) processes.

5.3 Examination of Conditions for Mixtures of Normal AR(1) Processes

We reintroduce the notation for our test case of logistic mixtures of normally distributed AR(1) processes:

$$\begin{aligned} f(Y_t \mid \mathcal{G}_{t-1}; \beta, \phi) =& f(Y_t \mid Y_{t-1}; \beta, \phi) \\ =& \exp\left(-\frac{(Y_t - Y_{t-1}\beta)^2}{2\phi} - \frac{1}{2}\log 2\pi\phi\right) \end{aligned} \tag{5.47}$$

$$\mathbb{P}\left[I_t = 1 \mid \mathcal{G}_{t-1}; \gamma\right] = \mathbb{P}\left[I_t = 1 \mid Y_{t-1}; \gamma\right] = \frac{\exp(\gamma_0 + Y_{t-1}\gamma_1)}{1 + \exp(\gamma_0 + Y_{t-1}\gamma_1)}. \tag{5.48}$$

From these distributions we construct our general model of a logistic mixture of this type as

$$\begin{aligned} g(Y_t \mid \mathcal{G}_{t-1}; \psi) = g(Y_t \mid Y_{t-1}; \psi) =& f(Y_t \mid Y_{t-1}; \beta_1, \phi_1)\mathbb{P}\left[I_t = 1 \mid Y_{t-1}; \gamma\right] + \end{aligned} \tag{5.49}$$

$$f(Y_t \mid Y_{t-1}; \beta_0, \phi_0)\mathbb{P}\left[I_t = 0 \mid Y_{t-1}; \gamma\right], \tag{5.50}$$

$$\text{where } \psi = (\beta_0, \beta_1, \phi_0, \phi_1, \gamma_0, \gamma_1)'. \tag{5.51}$$

Our mixtures will have some restrictions on the parameter space – the β_0 and β_1 terms are assumed to lie in $[-1+\epsilon, 1-\epsilon]$ and there exists $0 < \phi_{min} \leq \phi_1, \phi_0 \leq \phi_{max}$ and (γ_0, γ_1) are assumed to lie in Γ, a compact subset of $\mathbb{R}^2$ excluding all points of the form $\{(\gamma_0, 0)\colon \gamma_0 \in \mathbb{R}\}$. These points are excluded to ensure the regime probabilities vary. We denote by Ψ the set of $(\beta_0, \beta_1, \phi_0, \phi_1, \gamma_0, \gamma_1)'$

satisfying these restrictions. In Chapter 3 we spent some effort showing the ergodic and stationary behavior of $\{Y_t, X_t, Z_t\} = \{Y_t, Y_{t-1}, Y_{t-1}\}$ when the logistic mixture is correctly specified. Here we examine the case when the mixture is incorrectly specified (i.e. the null hypothesis of no mixture is true). In this case $Y_t \mid Y_{t-1}$ has a $N(\beta^* Y_{t-1}, \phi^*)$ distribution and we can easily find the stationary marginal distribution for Y_t as

$$Y_t \sim N\left(0, \frac{\phi^*}{1-(\beta^*)^2}\right).$$

Furthermore, $\{Y_t, X_t, Z_t\} = \{Y_t, Y_{t-1}, Y_{t-1}\}$ has an easily derivable trivariate normal stationary distribution. Thus Condition 5.2.A is met.

Conditions 5.2.B, 5.2.D, and 5.2.H are easily satisfied by finding dominating functions that are integrable. For example, suppose we wish to verify that $\mathbb{E}[g(W;\psi)]$ is continuous with respect to ψ. Given the continuity of the integrand we need only find an integrable function that bounds $g(W;\psi)$ for any $\psi \in \Psi$. Let W_1, W_0 denote r.v.'s that have the stationary distribution associated with Y_t, Y_{t-1}. Then for any $\psi \in \Psi$ we have

$$|\log g(W;\psi)| < |\log f(W_1 \mid W_0; \beta_1, \phi_1)| + |\log f(W_1 \mid W_0; \beta_0, \phi_0)| \text{ where}$$

$$|\log f(W_1 \mid W_0; \beta, \phi)| = \left| -\frac{(W_1 - W_0\beta^*)^2}{2\phi} - \frac{1}{2}\log 2\pi\phi \right| \tag{5.52}$$

$$< c + \frac{1}{2}\log 2\pi\phi_{max} + \frac{W_1^2 + W_0^2 + 2|W_1 W_0|}{2\phi_{min}}. \tag{5.53}$$

Because W_1 and W_0 are bivariate normal the right-hand side of (5.53) is clearly integrable so continuity is established. The other claims in these conditions can be similarly established.

Condition 5.2.C may be verified using the techniques in Section 3.2.3.

The difficulty in obtaining a well behaved information matrix is the crux of

the trouble associated with tests for mixtures. As described in the introductory section to this chapter it is this trouble which leads to an empirical processes approach of Ghosh and Sen (1985) and Dacunha-Castelle and Gassiat (1997). Unfortunately, neither of these approaches can be directly used as the restriction to (β, ϕ) and (β^*, ϕ^*) combinations satisfying

$$\mathbb{E}\left[\left(\left|\frac{f(W;\beta,\phi)}{f(W;\beta^*,\phi^*)}\right|^k\right)\right] < \infty$$

is too restrictive in the case of Gaussian AR processes (this was discussed at length in Section 5.1). In our approach we consider a different restriction of the parameter space that allows us to avoid ratios of this form.

To check that our information matrices in Condition 5.2.E are well behaved we check to see that the conditions of Lemma 5.1 are met. To use the notation of the lemma we take

$$p = \frac{\exp(\gamma_0 + \gamma_1 W_0)}{1 + \exp(\gamma_0 + \gamma_1 W_0)} \text{ and } V = \begin{bmatrix} \frac{(W_1 - W_0\beta^*)W_0}{\phi^*} \\ \frac{(W_1 - W_0\beta^*)^2}{2(\phi^*)^2} - \frac{1}{2\phi^*} \end{bmatrix}.$$

We want to show

$$\mathbb{E}\left[\begin{pmatrix} pV \\ qV \end{pmatrix}\begin{pmatrix} pV \\ qV \end{pmatrix}'\right] \text{ is positive definite.}$$

In Section 5.2 we claimed $\mathbb{P}[\Omega] = 1$ for our logistic mixture of Gaussian AR(1) processes, where

$$\Omega = \{p\colon \mathbb{E}[VV' \mid p] \text{ is positive definite }\}.$$

To see this note that a given value of p uniquely determines a given value of W_0 so we want to show $\mathbb{E}[VV' \mid W_0]$ is positive definite for almost all W_0. To show

this we first note that

$$V = \left.\frac{\partial \log f(W_1 \mid W_0; \beta, \phi)}{\partial (\beta, \phi)'}\right|_{\beta=\beta^*, \phi=\phi^*} \triangleq \begin{bmatrix} \left.\frac{\partial \log f(W_1|W_0;\beta,\phi)}{\partial \beta'}\right|_{\beta=\beta^*,\phi=\phi^*} \\ \left.\frac{\partial \log f(W_1|W_0;\beta,\phi)}{\partial \phi}\right|_{\beta=\beta^*,\phi=\phi^*} \end{bmatrix}.$$

Then it follows that

$$\begin{aligned} \mathbb{E}[VV' \mid W_0] &= -\mathbb{E}\left[\left.\frac{\partial^2 \log f(W_1 \mid W_0; \beta, \phi)}{\partial (\beta, \phi) \partial (\beta, \phi)'}\right|_{\beta=\beta^*, \phi=\phi^*} \middle| W_0 \right] \\ &= \begin{bmatrix} \frac{W_0^2}{\phi^*} & 0 \\ 0 & \frac{1}{2\phi^{*2}} \end{bmatrix} \end{aligned}$$

which is clearly positive definite if $W_0 \neq 0$. As W_0 is marginally Gaussian this exceptional set has measure 0. Thus the conditions for Lemma 5.1 are met and Condition 5.2.E is justified.

The next condition we check is 5.2.F which concerns uniform convergence. In section 3.4.3 we showed how to prove uniform convergence of the Hessian matrix when the logistic mixture was correctly specified by finding integrable functions that bounded the third derivatives. This same approach works here and we consider this condition justified.

Establishing Condition 5.2.G requires a little effort. Though there may be other methods of proof we will use a theorem for stochastic equicontinuity of martingales.

Lemma 5.6. *Suppose*

$$\left\{ \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left.\frac{\partial \log g(y_t \mid Y_{t-1}, \gamma, \theta)}{\partial \theta'}\right|_{\theta^*} \right\}_{T \in \mathbb{N}} \tag{5.54}$$

is a stochastically equicontinuous sequence of functions in $\mathcal{C}^{2d}[\Gamma]$, the space of $\mathbb{R}^{2d}$-valued continuous functions with domain Γ. In addition suppose that, for

some $\tilde{\gamma}$ we have

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left. \frac{\partial \log g(y_t \mid Y_{t-1}, \tilde{\gamma}, \theta)}{\partial \theta'} \right|_{\theta^*} \xrightarrow{\mathcal{D}} X \tag{5.55}$$

where X denotes an integrable random variable. Then

$$\sup_{\gamma \in \Gamma} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left. \frac{\partial \log g(y_t \mid Y_{t-1}, \gamma, \theta)}{\partial \theta'} \right|_{\theta^*} = O_p(1). \tag{5.56}$$

Proof. Let $\epsilon > 0$ be given and define

$$R_T(\gamma) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left. \frac{\partial \log g(y_t \mid Y_{t-1}, \gamma, \theta)}{\partial \theta'} \right|_{\theta^*}.$$

From (5.55) we know there exists a T_ϵ and M_ϵ such that

$$\mathbb{P}\left[|R_T(\tilde{\gamma})| > M_\epsilon\right] < \epsilon/2 \text{ for all } T > T_\epsilon. \tag{5.57}$$

From our definition of stochastic equicontinuity let δ satisfy

$$\overline{\lim_{T \to \infty}} \mathbb{P}\left[\sup_{||\gamma^1 - \gamma^2|| < \delta} \left|R_T(\gamma^1) - R_T(\gamma^2)\right| > \epsilon/2\right] < \epsilon/2, \tag{5.58}$$

and let $K = \inf\{N \in \mathbb{N}\colon N\delta > \sup_{\gamma^1, \gamma^2} ||\gamma^1 - \gamma^2||\}$ where throughout this proof γ^1 and γ^2 are arbitrary elements in Γ. K is guaranteed finite since Γ is compact. Then for arbitrary $\gamma \in \Gamma$, $||\gamma - \tilde{\gamma}|| < K\delta$, and

$$\begin{aligned}
\left[\sup_{\gamma \in \Gamma} |R_T(\gamma) - R_T(\tilde{\gamma})| > K\epsilon/2\right] &\subset \left[\sup_{||\gamma^1 - \gamma^2|| < K\delta} \left|R_T(\gamma^1) - R_T(\gamma^2)\right| > K\epsilon/2\right] \\
&\subset \left[\sup_{||\gamma^1 - \gamma^2|| < \delta} \left|R_T(\gamma^1) - R_T(\gamma^2)\right| > \epsilon/2\right]
\end{aligned}$$

which implies

$$\mathbb{P}\left[\sup_{\gamma \in \Gamma} |R_T(\gamma) - R_T(\tilde{\gamma})| > K\epsilon/2\right] \leq \mathbb{P}\left[\sup_{||\gamma^1 - \gamma^2|| < \delta} \left|R_T(\gamma^1) - R_T(\gamma^2)\right| > \epsilon/2\right] \leq \epsilon/2. \tag{5.59}$$

From (5.57) and (5.59) we see

$$\overline{\lim_{T\to\infty}}\,\mathbb{P}\left[\sup_{\gamma\in\Gamma}|R_T(\gamma)| > M_\epsilon + K\epsilon/2\right] \tag{5.60}$$

$$\leq \overline{\lim_{T\to\infty}}\,\mathbb{P}\left[|R_T(\tilde{\gamma})| + \sup_{\gamma\in\Gamma}|R_T(\gamma) - R_T(\tilde{\gamma})| > M_\epsilon + K\epsilon/2\right] \tag{5.61}$$

$$\leq \overline{\lim_{T\to\infty}}\,\mathbb{P}\left[|R_T(\tilde{\gamma})| > M_\epsilon\right] + \overline{\lim_{T\to\infty}}\,\mathbb{P}\left[\sup_{\gamma\in\Gamma}|R_T(\gamma) - R_T(\tilde{\gamma})| > K\epsilon/2\right] < \epsilon. \tag{5.62}$$

So

$$\overline{\lim_{T\to\infty}}\,\mathbb{P}\left[\sup_{\gamma\in\Gamma}|R_T(\gamma)| > M_\epsilon + K\epsilon/2\right] < \epsilon.$$

From our definition of $R_T(\gamma)$ the proof is complete. □

Remark: As mentioned in the previous section, this proof works to show that if $\hat{S}_T(\gamma)$ is stochastically equicontinuous and $\mathbb{E}\left[\hat{S}_T(\tilde{\gamma})'\hat{S}_T(\tilde{\gamma})\right] < \infty$ for some $\tilde{\gamma} \in \Gamma$ then $\sup_\gamma \hat{S}_T(\gamma)'\hat{S}_T(\gamma)$ is $O_p(1)$. In this case we define $R_T(\gamma) = \hat{S}_T(\gamma)$ and the proof follows as above.

At this point we note that any $\gamma \in \Gamma$ will suffice as the $\tilde{\gamma}$ term in the statement of the Lemma 5.6 since we know that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left.\frac{\log g(y_t \mid \mathcal{G}_{t-1};\gamma,\theta)}{\partial\theta'}\right|_{\theta^*} \xrightarrow{\mathcal{D}} N\left(\underline{0}, \mathbb{E}\left[s_t(\gamma)s_t(\gamma)'\right]\right).$$

As both this lemma and Condition 5.2.I are concerned with stochastic equicontinuity it seems the appropriate time to discuss a theorem that tells us when a martingale may be stochastically equicontinuous. The theorem below is drawn from Hansen (1996b) and is particularly suited for Lipshitz smooth functions of our parameters, γ. Let W_t be R^p valued random vector on $(\Omega, \mathcal{F}, \mathbb{P})$ and $h(w,\gamma)$ be a parametric class of random functions from $R^p \times \Gamma \xrightarrow{\mathbb{R}}$ where Γ is compact set (parameter space) in R^a . The conditions for this theorem are

A.1: The function h satisfies a Lipshitz condition

$$\left|h(w,\gamma^1) - h(w,\gamma^2)\right| \leq b(w)\left|\left|\gamma^1 - \gamma^2\right|\right| \tag{5.63}$$

for all $\gamma^1, \gamma^2 \in \Gamma$ where $b : \mathbb{R}^p \to \mathbb{R}$,

A.2: For some $q > \max(2, a)$ (where a is the dimension of γ) $||b(w)||_q < \infty$,

A.3: For all $\gamma \in \Gamma, ||h(W_t, \gamma)||_q < \infty$, and

A.4: $\{h(W_t, \gamma), \sigma(W_{t-1}, W_{t-2}, ...)\}$ is a stationary and ergodic martingale difference sequence,

where $||\cdot||_q$ in A.2 and A.3 denotes the usual L^q norm.

Theorem 5.7. *Under conditions* A.1 - A.4 we have that for every $\epsilon > 0$ there exists a $\delta > 0$ such that

$$\overline{\lim_{T\to\infty}} \left|\left| \sup_{\substack{\gamma^1,\gamma^2: \\ ||h(W,\gamma^1)-h(W,\gamma^2)||_q<\delta}} \frac{1}{\sqrt{T}} \left| \sum_{t=1}^{T} (h(W_t, \gamma^1) - h(W_t, \gamma^2)) \right| \right|\right|_q < \epsilon. \quad (5.64)$$

The W in $||h(W, \gamma^1) - h(W, \gamma^2)||_q$ has the common distribution of the stationary W_t's. We should note that the result above is stated in terms of L^q equicontinuity instead of the more familiar L^0 result and uses the L^q norm in defining the modulus of continuity,

$$\sup_{\substack{\gamma^1,\gamma^2: \\ ||h(W,\gamma^1)-h(W,\gamma^2)||_q<\delta}} \frac{1}{\sqrt{T}} \left| \sum_{t=1}^{T} (h(W_t, \gamma^1) - h(W_t, \gamma^2)) \right|.$$

An application of the Markov inequality and the Lipshitz continuity condition A.1 let us rewrite this in a more familiar way:

Corollary 5.8. *Under conditions* A.1 - A.4 we have for all $\epsilon > 0$ there exists a $\delta > 0$ such that

$$\overline{\lim_{T\to\infty}} \mathbb{P}\left[\sup_{\substack{\gamma^1,\gamma^2: \\ ||\gamma^1-\gamma^2||<\delta}} \frac{1}{\sqrt{T}} \left| \sum_{t=1}^{T} (h(W_t, \gamma^1) - h(W_t, \gamma^2)) \right| > \epsilon \right] < \epsilon.$$

Remark: As defined above, $h(W,\gamma)$ takes values in $\mathbb{R}$ though we have in mind showing stochastic equicontinuity (s.e.) of $\frac{1}{\sqrt{T}}\sum s_t(\gamma)$ which is a vector. However, if each component of $\frac{1}{\sqrt{T}}\sum s_t(\gamma)$ is s.e. then it follows that the vector $\frac{1}{\sqrt{T}}\sum s_t(\gamma)$ is also s.e. Thus it is sufficient to use the theorem to prove each element of the vector is stochastically equicontinuous.

So as γ is two dimensional we wish to apply this theorem twice with $W_t = (Y_t, Y_{t-1})$ and $h(W_t,\gamma) = s_t(\gamma)_i$, the i^{th} component of $s_t(\gamma)$ as defined in (5.37) with

$$b(X_t'\beta^*) = \frac{(Y_{t-1}\beta^*)^2}{2}, c(y_t,\phi) = -\frac{\log 2\pi\phi}{2} - \frac{-y_t^2}{2\phi}, \text{ and}$$
$$p_t(\gamma) = \frac{\exp(\gamma_0 + Y_{t-1}\gamma_1)}{1+\exp(\gamma_0 + Y_{t-1}\gamma_1)} \text{ and } i = 0,1.$$

We have already demonstrated that $s_t(\gamma)$ is a martingale and so we need only demonstrate the existence of a function that uniformly bounds the derivatives (with respect to γ) and the existence of qth moments for $s_t(\gamma)$. Techniques for showing the integrability of such functions has been demonstrated previously so we omit it and consider this condition justified.

With this demonstration we have shown how our sample model fits the conditions of the previous section and will therefore have a likelihood ratio statistic with the appropriate asymptotic distribution – see equation (5.43). The next chapter is concerned with the performance of our test and includes a brief summary of the underlying theory we presented in this chapter.

Chapter 6

Applications of the Likelihood Ratio Test

In this chapter we examine the performance of our likelihood ratio test using simulations and the GATE dataset. We begin by summarizing the theory underlying our test and then discuss a general algorithm for producing realizations of Gaussian fields that have the required covariance structure. There are at least two straightforward methods for producing such realizations and because they have different computational burdens it is worthwhile to examine the methods in some detail. After describing our algorithm we perform some simulations to check that our test works well when the null hypothesis of a one regime (no mixture) model is true. Our results support the theory developed in the last chapter. Next we examine the power of our test by comparing it to an alternative test procedure in simulations of logistic mixtures of AR(1) processes. In this chapter we refer to our test as the empirical process, or EP, test and the alternative test as the Monte Carlo, or MC, test. The simulations suggest our test works well in comparison to the alternative Monte Carlo method. Finally, we apply our test to the GATE data discussed in Chapter 4 and conclude that

there is strong evidence of a mixture of log-normal densities.

6.1 Implementing the Test

We begin with an observed set of data $\{y_t, X_t, Z_t\}_{t=1}^T$. Under the hypothesis that a logistic mixture is present we assume the conditional density of the $\{y_t\}$ is given by

$$\begin{aligned}
g(y_t \mid \mathcal{G}_{t-1}; \psi^*) =& g(y_t \mid X_t, Z_t; \beta_1^*, \phi_1^*, \beta_0^*, \phi_0^*, \gamma^*) \\
=& \mathbb{P}\left[I_t = 1 \mid Z_t; \gamma^*\right] \cdot f(y_t \mid X_t; \beta_1^*, \phi_1^*) + \\
& \left(1 - \mathbb{P}\left[I_t = 1 \mid Z_t; \gamma^*\right]\right) \cdot f(y_t \mid X_t; \beta_0^*, \phi_0^*), \\
& \text{where } \mathcal{G}_{t-1} = \sigma(X_t, Z_t), \psi^* = (\beta_1^{*\prime}, \phi_1^*, \beta_0^{*\prime}, \phi_0^*, \gamma^{*\prime})' \\
f(y_t \mid X_t; \beta, \phi) =& \exp\left(\frac{y_t X_t'\beta - b\left(X_t'\beta\right)}{\phi} + c(y_t, \phi)\right) \text{ and} \\
\mathbb{P}\left[I_t = 1 \mid Z_t; \gamma\right] =& \frac{\exp\left(Z_t'\gamma\right)}{1 + \exp\left(Z_t'\gamma\right)}.
\end{aligned}$$

ψ^* is assumed to be some unknown element of Ψ, a compact subset of Euclidean space. The γ component of ψ is assumed to lie in Γ a compact subset of $\mathbb{R}^{r+1}$ and it is assumed that for any γ in Γ, $Z_t'\gamma = \gamma_0 + \gamma_1 Z_{t1} + \cdots + \gamma_r Z_{tr}$ is not constant (i.e. $Z_t'\gamma$ varies with t).

Here we will summarize our results from Chapter 5. Using the EM algorithm approach of Chapter 2, or some other maximization procedure, one finds the maximum likelihood estimates of the mixture parameters, $\hat{\psi} = (\hat{\beta}_1', \hat{\phi}_1, \hat{\beta}_0', \hat{\phi}_0, \hat{\gamma}')'$. One must ensure that these estimates lie in the compact region Γ – perhaps by using a constrained optimization procedure. Once these parameters are found we obtain the value of the mixture likelihood associated with this particular set.

We denote this by

$$L_T^g(\hat{\psi}) \stackrel{\triangle}{=} \text{Arg} \max_{\psi \in \Psi} \sum_{t=1}^{T} \log g(y_t \mid \mathcal{G}_{t-1}; \hat{\psi}). \tag{6.1}$$

To get the log likelihood ratio we need the corresponding maximum likelihood parameters for a single regime model. We obtain these through some generalized linear models fitting package and evaluate the log likelihood at these parameters. This value we denote as

$$L_T^f(\hat{\beta}, \hat{\phi}) \stackrel{\triangle}{=} \text{Arg} \max_{\beta, \phi} \sum_{t=1}^{T} \log f(y_t \mid \mathcal{G}_{t-1}; \beta, \phi). \tag{6.2}$$

Under the null hypothesis that there is no mixture and the true conditional density is given by $f(y_t \mid X_t; \beta^*, \phi^*)$, we showed in the last chapter that

$$2 * \left(L_T^g(\hat{\psi}) - L_T^f(\hat{\beta}, \hat{\phi}) \right)$$

converges (in distribution) to the supremum of a transformation of a Gaussian random field – elements of the field are denoted by $S(\gamma)$ for $\gamma \in \Gamma$. For any $\gamma \in \Gamma$, $S(\gamma)$ has a $3d$ dimensional mean zero multivariate normal distribution with variance matrix $K(\gamma, \gamma) \stackrel{\triangle}{=} \mathbb{E}\left[s_t(\gamma) s_t(\gamma)'\right]$ where $s_t(\gamma)$ is defined in (5.37) and d is the dimension of $(\beta', \phi)'$. For any two elements in Γ, the covariance matrix is given by $K(\gamma^1, \gamma^2) = \mathbb{E}\left[s_t(\gamma^1) s_t(\gamma^2)'\right]$.

The transformation we apply to this field is given by

$$S(\gamma)^{2d'} \left[K(\gamma, \gamma)^{2d}\right]^{-1} S(\gamma)^{2d} - S(\gamma)^{d'} \left[K(\gamma, \gamma)^{d}\right]^{-1} S(\gamma)^{d}$$

where $S(\gamma)^{2d}$ denotes the first $2d$ elements in the vector $S(\gamma)$, $S(\gamma)^d$ denote the last d elements, and $K(\gamma, \gamma)^{2d}$ is the $2d \times 2d$ upper left hand corner of the $K(\gamma, \gamma)$ matrix, and $K(\gamma, \gamma)^d$ is the $d \times d$ lower right hand corner of the same matrix.

Showing $K(\gamma,\gamma)^{2d}$ is invertible was an important part of our discussion. Our primary result from Chapter 5 was showing that

$$2 * \left(L_T^g(\hat{\psi}) - L_T^f(\hat{\beta},\hat{\phi}) \right) \xrightarrow{\mathcal{D}}$$
$$\sup_{\gamma\in\Gamma} \left\{ S(\gamma)^{2d'} \left[K(\gamma,\gamma)^{2d} \right]^{-1} S(\gamma)^{2d} - S(\gamma)^{d'} \left[K(\gamma,\gamma)^{d} \right]^{-1} S(\gamma)^{d} \right\}. \tag{6.3}$$

Because we do not know $K(\gamma^1,\gamma^2)$ we must estimate it from our sample data by

$$\hat{K}_T(\gamma^1,\gamma^2) \stackrel{\triangle}{=} \frac{1}{T}\sum_{t=1}^{T} s_t(\gamma^1;\hat{\beta},\hat{\phi}) s_t(\gamma^2;\hat{\beta},\hat{\phi})' \tag{6.4}$$

where $s_t(\gamma;\hat{\beta},\hat{\phi})$ corresponds to our definition of $s_t(\gamma)$ in (5.37) except with $(\hat{\beta},\hat{\phi})$ in place of (β^*,ϕ^*). With this change we approximate the original functional with a new one:

$$\sup_{\gamma\in\Gamma} \left\{ S(\gamma)^{2d'} \left[K(\gamma,\gamma)^{2d} \right]^{-1} S(\gamma)^{2d} - S(\gamma)^{d'} \left[K(\gamma,\gamma)^{d} \right]^{-1} S(\gamma)^{d} \right\} \approx$$
$$\sup_{\gamma\in\Gamma} \left\{ \hat{S}_T(\gamma)^{2d'} \left[\hat{K}_T(\gamma,\gamma)^{2d} \right]^{-1} \hat{S}_T(\gamma)^{2d} - \hat{S}_T(\gamma)^{d'} \left[\hat{K}_T(\gamma,\gamma)^{d} \right]^{-1} \hat{S}_T(\gamma)^{d} \right\} \tag{6.5}$$
$$\text{where } \hat{S}_T(\gamma) = \left(\hat{S}_T(\gamma)^{2d'}, \hat{S}_T(\gamma)^{d'} \right)' \text{ has a } N\left(0, \hat{K}_T(\gamma,\gamma)\right)$$

distribution with covariance kernel $\hat{K}_T(\gamma^1,\gamma^2)$ for $\gamma^1,\gamma^2 \in \Gamma$. In Theorem 5.5 we showed the two functionals above have the same asymptotic distribution.

We now address the question of how best to find the distribution of this approximation. The simplest way is to create a large number, say L, independent Gaussian random fields that have the required covariance function $\hat{K}_T(\cdot,\cdot)$ and directly compute

$$\sup_{\gamma\in\{\gamma^1\ldots\gamma^N\}} \left\{ \hat{S}_T(\gamma)^{2d'} \left[\hat{K}_T(\gamma,\gamma)^{2d} \right]^{-1} \hat{S}_T(\gamma)^{2d} - \hat{S}_T(\gamma)^{d'} \left[\hat{K}_T(\gamma,\gamma)^{d} \right]^{-1} \hat{S}_T(\gamma)^{d} \right\}$$

for each realization where $\{\gamma^1 \ldots \gamma^N\}$ is a grid of points in Γ. The maximum we obtain over the grid is then our proxy value for the maximum value over Γ.

The L maxima thus obtained will form an empirical distribution whose quantiles approximate the quantiles of our unknown functional

$$\sup_{\gamma \in \Gamma} \left\{ S(\gamma)^{2d'} \left[K(\gamma,\gamma)^{2d} \right]^{-1} S(\gamma)^{2d} - S(\gamma)^{d'} \left[K(\gamma,\gamma)^{d} \right]^{-1} S(\gamma)^{d} \right\}.$$

Hence, we are now interested in finding good ways to generate Gaussian random fields with the appropriate covariance structure. The first way that might come to mind requires the construction of a matrix with entries for each $\hat{K}_T(\gamma^i, \gamma^j)$ submatrix where γ^i and γ^j are in $\{\gamma^1, \ldots, \gamma^N\}$. With this approach we construct a matrix with dimensions $3d \cdot N \times 3d \cdot N$ consisting of N^2 blocks of size $3d \times 3d$:

$$\hat{Q}_T(\gamma^1, ..., \gamma^N) = \begin{bmatrix} \hat{K}_T(\gamma^1, \gamma^1), & \ldots & \hat{K}_T(\gamma^1, \gamma^N) \\ \ldots & \hat{K}_T(\gamma^i, \gamma^j) & \\ \hat{K}_T(\gamma^N, \gamma^1), & \ldots & \hat{K}_T(\gamma^N, \gamma^N) \end{bmatrix}.$$

Each $\hat{K}_T(\gamma^i, \gamma^j)$ block is constructed from the data as in equation (6.4). From $\hat{Q}_T(\gamma^1, ..., \gamma^N)$ we may obtain a Cholesky factorization, $\hat{M}_T(\gamma^1, \ldots, \gamma^N)$ satisfying

$$\hat{M}_T(\gamma^1, \ldots, \gamma^N)\hat{M}_T(\gamma^1, \ldots, \gamma^N)' = \hat{Q}_T(\gamma^1, \ldots, \gamma^N)$$

where $\hat{M}_T(\gamma^1, \ldots, \gamma^N)$ is a lower triangular matrix and $\hat{M}_T(\gamma^1, \ldots, \gamma^N)'$ its upper triangular transpose. If Z is a vector of i.i.d. $N(0,1)$ random variables with length $3d \cdot N$ then it is clear that

$$\hat{M}_T(\gamma^1, \ldots, \gamma^N)Z \text{ has a } N(0, \hat{Q}_T(\gamma^1, \ldots, \gamma^N)) \text{ distribution.}$$

By combining the appropriate elements of $\hat{M}_T(\gamma^1, \ldots, \gamma^N)Z$ with $\left[\hat{K}_T(\gamma^i, \gamma^i)^{2d}\right]^{-1}$ and $\left[\hat{K}_T(\gamma^i, \gamma^i)^{d}\right]^{-1}$ one can find the supremum for this particular realization of the field. Different realizations of Z give different realizations of the field. While this method is conceptually clear there is a significant drawback associated

with it. If γ is multidimensional then the number of gridpoints can be quite large and can make construction of this $Q_T(\gamma^1, \ldots, \gamma^N)$ matrix difficult from a computational perspective.

As an example let $\gamma \in \Gamma \subset \mathbb{R}^3$. Suppose that for each of the three dimensions of Γ we sample 10 points so that we end up with $1000 = 10^3$ points in Γ, i.e. N = 1000. Now, d is the dimension of (β, ϕ) and is at least 2. So our matrix $\hat{Q}_T(\gamma^1, \ldots, \gamma^N)$ has at least 36 million entries. If each entry requires 8 bytes this means we need approximately 288 megabytes of space to merely store the $\hat{Q}_T(\gamma^1, \ldots, \gamma^N)$ matrix. This does not take into account the resources and time necessary to compute the Cholesky factor. Alternatively, if Γ is 2 dimensional and we sample at 10 points for each dimension we require $8 * 36 * 100 * 100$ bytes of space, less than 3 megabytes. This example indicates the Cholesky method is quite sensitive to the dimension of Γ. Increasing the number of gridpoints by a factor of 10 increases the computational burden by a factor of 100.

We present an alternative method that is less computationally demanding. Here we define $Z = (Z_1, \ldots, Z_t, \ldots, Z_T)$ to be a vector of i.i.d. $N(0, 1)$ random variables of length T, the sample size. For each $\gamma^i \in \{\gamma^1, \ldots, \gamma^N\}$ define

$$\hat{S}_T\left(\gamma^i\right) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} Z_t \cdot s_t(\gamma^i, \hat{\beta}, \hat{\phi}).$$

Then a bit of algebra shows $\left\{ \hat{S}_T\left(\gamma^i\right) : \gamma^i \in \{\gamma^1, \ldots, \gamma^N\} \right\}$ has the right marginal and joint distributions. As before, different realizations of Z lead to different realizations of the Gaussian field. In this case we need not generate the $\hat{K}_T\left(\gamma^i, \gamma^j\right)$ terms (for $i \neq j$) which required so much time and space in the Cholesky method. While this method is still sensitive to the dimension of Γ, the increased burden from increasing the number of gridpoints is linear with this method, not

quadratic. We will use this method in the simulations and application below.

6.2 Simulation Results

In this section we present some simulation results. The first set of simulations examines our claim that under the null hypothesis

$$2 * \left(L_T^g(\hat{\psi}) - L_T^f(\hat{\beta}, \hat{\phi}) \right) \approx \tag{6.6}$$

$$\sup_{\gamma \in \{\gamma^1 \ldots \gamma^N\}} \left\{ \hat{S}_T(\gamma)^{2d'} \left[\hat{K}_T(\gamma,\gamma)^{2d} \right]^{-1} \hat{S}_T(\gamma)^{2d} - \hat{S}_T(\gamma)^{d'} \left[\hat{K}_T(\gamma,\gamma)^{d} \right]^{-1} \hat{S}_T(\gamma)^{d} \right\} \tag{6.7}$$

for some suitably chosen grid of values in Γ. In a second set of simulations we see how well our test detects a difference when the data is generated by a mixture – i.e. we examine the power of our test for a fixed alternative and compare it to that of another test.

Simulation Under the Null Hypothesis

We begin by generating 500 observations from a normal AR(1) process with mean .6 and standard deviation .5. That is

$$Y_t \,|\, Y_{t-1} \sim N\left(.6Y_{t-1}, .25\right).$$

Our alternative mixtures are characterized by the conditional density

$$\begin{aligned} g(y_t \mid Y_{t-1}; \psi) =& f(y_t \mid Y_{t-1}; \beta_1, \phi_1)\mathbb{P}\left[I_t = 1 \mid Y_{t-1}; \gamma\right] + \\ & f(y_t \mid Y_{t-1}; \beta_0, \phi_0)\mathbb{P}\left[I_t = 0 \mid Y_{t-1}; \gamma\right], \\ \text{where } \psi =& (\beta_0, \beta_1, \phi_0, \phi_1, \gamma)' \text{ and} \\ f(y_t \mid Y_{t-1}; \beta, \phi) =& \exp\left(-\frac{(y_t - Y_{t-1}\beta)^2}{2\phi} - \frac{1}{2}\log 2\pi\phi\right) \\ \mathbb{P}\left[I_t = 1 \mid Y_{t-1}; \gamma\right] =& \frac{\exp(\gamma Y_{t-1})}{1 + \exp(\gamma Y_{t-1})}. \end{aligned}$$

To ensure the probabilities $\mathbb{P}\left[I_t = 1 \mid Y_{t-1}; \gamma\right]$ are not constant, γ is restricted to lie in a compact set of the real line that excludes 0.

From our set of 500 observations we fit a one regime (or no mixture) model and generate estimates $(\hat{\beta}, \hat{\phi})$. These are the estimates that would be obtained through ordinary maximum likelihood fitting of AR(1) data. We also fit a logistic mixture model and obtain estimates of $(\beta_1, \phi_1, \beta_0, \phi_0, \gamma)$. From these estimates we obtain a log likelihood ratio that is one-half the term in (6.6).

Next we select a grid of points at which to evaluate our random field – in this case the grid is very coarse, consisting of the points $\{-3.5, -2.5, -2, -1.5, -1, -.5, -.2, .2, .5, 1, 1.5, 2, 2.5, 3.5\}$. Were we not performing simulations a finer grid would be chosen but to save time we used this one. We also chose γ to be one dimensional for the same reason. From examining realizations of our field we believe the maxima would not be much greater if the grid were made more fine. Thus we feel comfortable with using a relatively coarse grid.

Following the procedure we outlined in the previous section, for our initial sequence of 500 observations we generated $L = 100$ random fields with the covariance structure indicated by the $s_t(\gamma; \hat{\beta}, \hat{\phi})$ terms. From these 100 realizations of the random field we obtained 100 suprema corresponding to the supremum in

equation (6.7). To construct a test with approximate size of .05 we reject our hypothesis of no mixture if

$$2 * \left(L_T^g(\hat{\psi}) - L_T^f(\hat{\beta}, \hat{\phi}) \right) > 95^{th} \text{ ordered value of the 100 suprema.}$$

More generally, if we seek a test with size $\alpha/100$ we reject the hypothesis of no mixture if

$$2 * \left(L_T^g(\hat{\psi}) - L_T^f(\hat{\beta}, \hat{\phi}) \right) > (100 - \alpha)^{th} \text{ ordered value of the 100 suprema.}$$

Also we generate a p-value for each simulation by defining

$$\text{p-value} = \% \text{ of the 100 suprema that exceed } 2 * \left(L_T^g(\hat{\psi}) - L_T^f(\hat{\beta}, \hat{\phi}) \right).$$

We performed these steps for each simulation (a simulation corresponds to a single realization of 500 observations). We duplicated this process 100 times with new sequences of 500 observations generated by an AR(1) process with mean = .6, standard deviation = .5. From these 100 simulations we produced 100 .95 quantiles to which we compared our 100 likelihood ratio statistics (as in (6.6)). We also generated .75, .85, .90, and .99 quantiles as well. The table below shows how our empirical quantiles corresponded to the theoretical results. The empirical frequencies refer to the number of trials (out of 100) in which the likelihood ratio statistic exceeded the empirical quantile. The results indicate the empirical distribution fits quite well. The mean quantiles in the Table 6.1 are the averages of the 100 quantiles generated. To further investigate the fit we constructed a quantile plot of the p-values which we defined above. If the empirical distribution is a good fit to the log-likelihood ratio we should see the plot of the p-values lying close the the diagonal line that corresponds to the quantiles of a uniform random variable.

Figure 6.1 confirms that the fit is good over the entire distribution, not just at the selected quantiles presented in Table 6.1. A Kolmogorov-Smirnov two-sided test yields a p-value of .27 for the null hypothesis that the empirical p-values pictured as dots on the graph come from a uniform distribution. We believe these results support our view that the asymptotic distribution of the log likelihood ratio is given by the distribution of our proposed functional.

Part of what is not addressed in this simulation is how well the asymptotic distribution characterizes smaller sample sizes. We chose a large sample size of 500 to have some confidence that the results should reflect asymptotic behavior. We did not explore the performance of the test with fewer observation. Future work on logistic mixtures should examine this question.

Simulations Under the Alternative Hypothesis

Here we investigate how well our test detects the presence of a mixture. As we wanted to contrast our results with a test that is already in use, we had at least three tests from which to choose. Two such tests were discussed in the Chapter 1 – tests by Hansen (1992, 1996b) and by Gong and Mariano (1997). Both of

Probability	Mean Quantile	Empirical Frequency	Theoretical Frequency
.75	3.20	26	25
.85	4.22	17	15
.90	5.03	12	10
.95	6.40	7	5
.99	9.17	4	1

Table 6.1: 100 Simulations of 500 Observations

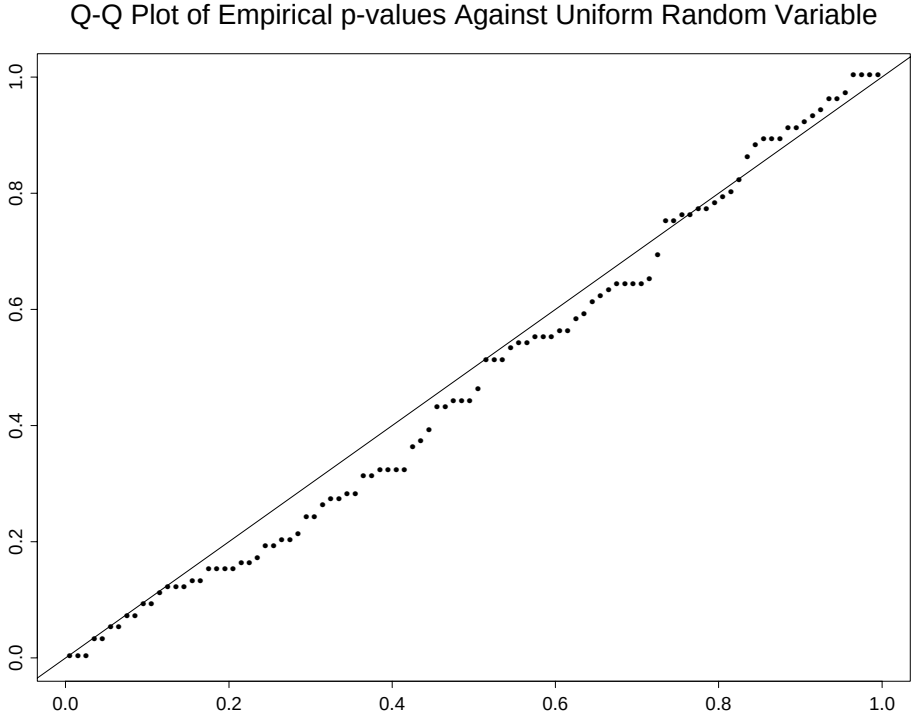

Figure 6.1: Quantile Plot When the Null Hypothesis is True

these tests were developed in the context of hidden Markov model regression and we explored the possibilities for adapting them to our logistic mixture model.

In the case of Hansen's test we developed an analogous procedure (Hansen's test is quite general with respect to model specification) for logistic mixtures, yet the test's power was poor. This is not surprising as the critical values for the test are derived not for the distribution of the log-likelihood ratio, but rather for a variable that bounds (from above) the likelihood ratio. The observed likelihood ratio is then compared to the critical value based on this bound. In short, this bound is too generous to have much power. We did perform some simulations (not presented here) in which both our test and the alternative test we chose performed much better than Hansen's. Originally we had planned to use a Hansen-like approach to testing the likelihood ratio but its poor power

suggested that we develop another means.

Gong and Mariano (1997) presented a test statistic with an exact asymptotic distribution, not a bound. The test is different from any other considered in that its test statistic is drawn from the spectral representation of the $\{Y_t\}$. As with Hansen's test this procedure was developed for a two regime hidden Markov models with regime probabilities determined by a fixed transition matrix. However, this test requires an analytic expression for the spectral distribution of Y_t under the alternative hypothesis and it is not clear to us how to derive this in the context of logistic mixtures with covariates. Consequently we are unable to use this test as an alternative.

Both of these proposed test and our test use the idea of performing profiled maximum likelihood holding fixed some parameter, say γ, and then considering the result a process that varies with γ. A potentially more appealing approach is to use a Monte Carlo or bootstrap approach to the problem. In the context of i.i.d. data with constant regime probabilities this approach has been examined by a number of authors – among them McLachlan (1987), McLachlan, Green, and Basford (1993), and Feng and McCulloch (1997). The idea is that given a sequence of observed data we obtain the log likelihood ratio by estimating the model under both the single regime and the mixture hypotheses. Then, using the results from the single regime estimation we generate independent datasets according to the distribution given by the single regime parameter estimates. To each of these datasets we fit both one and two regime models that give us a likelihood ratio statistic. Thus each independent dataset generates a likelihood ratio statistic that is derived when the data was generated by our original sample's one regime estimates. We then compare the original likelihood ratio statistic to

quantiles derived from our empirical sample of likelihood ratio statistics. We reject the hypothesis of no mixture if the original likelihood ratio statistic exceeds some pre-specified quantile of the empirical sample.

There are at least two drawbacks to such a procedure. First (as pointed out by Hansen (1992)), while the design is intuitively appealing the approach lacks a theoretical basis for claiming the empirical sample should provide a good estimate of the likelihood ratio when the null hypothesis is true. It may be that there is an asymptotic equality but we are not aware of a demonstration to this effect. Second, as also pointed out by Hansen (1992) and Hamilton (1990), there is some difficulty in finding maximum likelihood estimates of a mixture when the data are generated by a one regime model. In such cases the m.l.e.s are difficult to find as the likelihood surface is likely to be relatively flat with many local maxima. Yet this Monte Carlo method depends upon finding the mixture m.l.e.s for each of several independently generated datasets. In practice the search is likely to result in an underestimated maximized likelihood values. This will lead to rejecting the hypothesis of no mixture more often than is correct (under the assumption that the empirical distribution of the independent datasets is a good estimate of the likelihood ratio statistic under the null hypothesis). The more complicated the parameter space (or more dimensions) the more likely one is to underestimate the maximized likelihood under the mixture hypothesis. On the other hand many authors report this method works well in simulations and even performs adequately in small samples (Feng and McCulloch (1997)).

To examine the two tests we used data generated by the following logistic

mixture model:

$$\begin{aligned} g(y_t \mid Y_{t-1}; \psi) =& f(y_t \mid Y_{t-1}; \beta_1, \phi_1)\mathbb{P}\left[I_t = 1 \mid Y_{t-1}; \gamma\right] + \\ & f(y_t \mid Y_{t-1}; \beta_0, \phi_0)\mathbb{P}\left[I_t = 0 \mid Y_{t-1}; \gamma\right], \\ \text{where } \psi =& (\beta_0, \beta_1, \phi_0, \phi_1, \gamma)' \text{ and} \\ f(y_t \mid Y_{t-1}; \beta, \phi) =& \exp\left(-\frac{(y_t - Y_{t-1}\beta)^2}{2\phi} - \frac{1}{2}\log 2\pi\phi\right) \\ \mathbb{P}\left[I_t = 1 \mid Y_{t-1}; \gamma\right] =& \frac{\exp(\gamma Y_{t-1})}{1 + \exp(\gamma Y_{t-1})}. \end{aligned}$$

In the simulations $(\beta_1, \phi_1) = (.6, .25)$, $(\beta_0, \phi_0) = (.2, .49)$ and $\gamma = -1$ and the sample size was 200. These model parameters were chosen because they seemed to generate data for which the tests had quite variable results as opposed to different parameter choices for which the tests nearly always rejected or nearly always accepted the null hypothesis. Fifty samples of independent data were generated according to the model above. For each of the fifty datasets we determined the likelihood ratio statistic. Also for each of the fifty datasets, 100 independent Gaussian fields were generated which our test (which we refer to as the empirical process, or EP test) used to create a p-value for the the hypothesis that the dataset was generated by a one regime model (the grid points used in the earlier simulations was used here as well). In addition, for each of the fifty datasets 100 Monte Carlo simulations were produced to obtain a p-value for the Monte Carlo (MC) test. We could have generated more than fifty sets of data but the trends were clear.

In Figure 6.2 we show box-plots of the 50 p-values we obtained under the two test procedures. The figure indicates that on average the EP test performed somewhat better than the MC procedure – this despite the likelihood that the

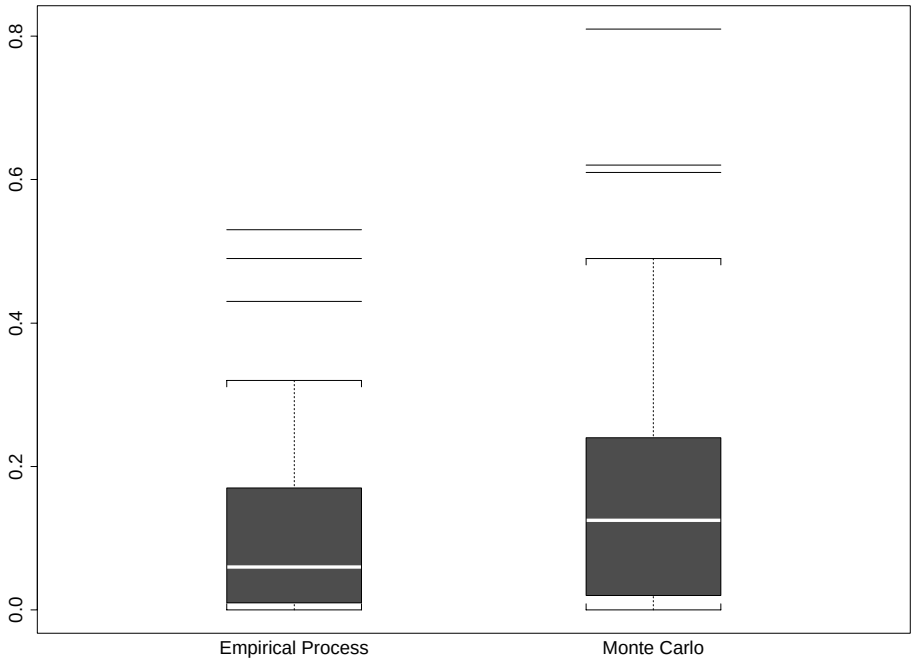

Figure 6.2: Comparison of Empirical Process and Monte Carlo Tests via Boxplot

MC test probably underestimated the maximized likelihood under the mixture hypothesis. (It is also true that the EP test underestimates the quantiles of true distribution because our search for the maxima is restricted to a finite grid of points. However, as mentioned above, our impression was that the realizations of the chi-square processes were relatively flat and that more points would not have greatly increased the maximum values.) The mean p-value for the EP test was .11 and that for the MC test was .18. The corresponding standard deviations were .13 and .19.

In Figure 6.3 we use a scatterplot to show how the two tests fared on each trial. For points above the diagonal 45 degree line the EP p-value was lower than the MC test and the reverse was true for points above the line. From this we see

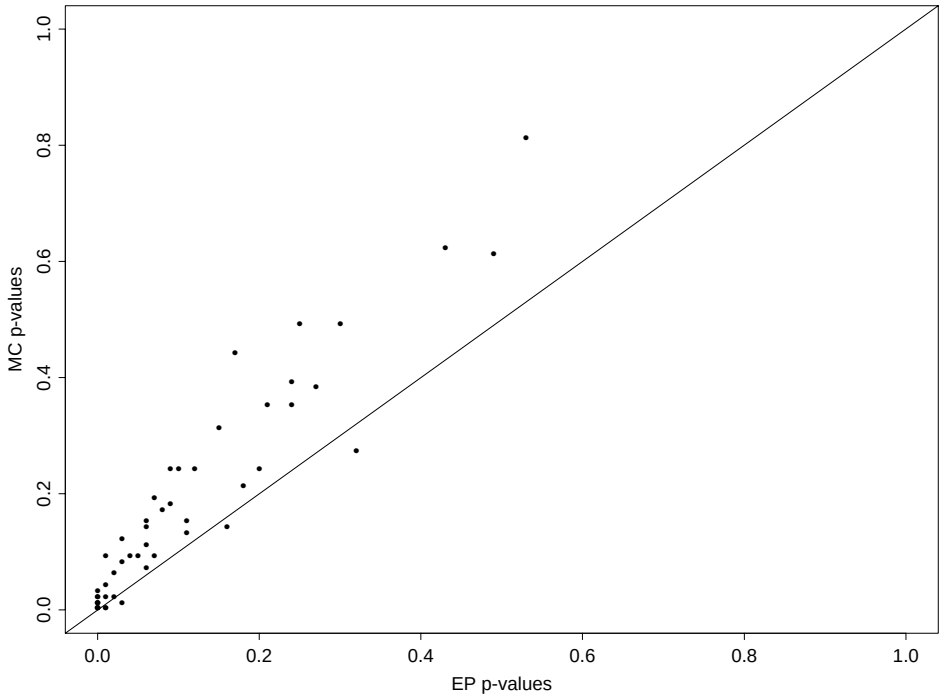

Figure 6.3: Comparison of Empirical Process and Monte Carlo Tests via Scatterplot

that our empirical process test was more discriminating in the great majority of the cases and we feel comfortable concluding that this suggests our empirical process test is more powerful than the Monte-Carlo test – at least for this choice of parameter values.

6.3 Application to Rain Data

In this section we apply our test to the rain rate data described in Chapter 4. We recall from Table 4.5 that the log-likelihood associated with the one regime (1R) model was -1430.09 and that of the logistic mixture (LM) was -1381 so our

test statistic is given by

$$2 * (-1380.97 + 1430.09) = 98.24.$$

Let $\gamma = (a, b, d)$ where $a, b,$ and d are as defined in section 4.3. Now define

$$s_t(\gamma; \hat{\mu}, \hat{\phi}) = \begin{bmatrix} p_t^s(h_t;\gamma) \sum \frac{y_t - \hat{\mu}}{\hat{\phi}} \\ p_t^s(h_t;\gamma) \sum \frac{1}{2\hat{\phi}} \left(\frac{(y_t - \hat{\mu})^2}{\hat{\phi}} - 1 \right) \\ p_t^c(h_t;\gamma) \sum \frac{y_t - \hat{\mu}}{\hat{\phi}} \\ p_t^c(h_t;\gamma) \sum \frac{1}{2\hat{\phi}} \left(\frac{(y_t - \hat{\mu})^2}{\hat{\phi}} - 1 \right) \\ \sum \frac{y_t - \hat{\mu}}{\hat{\phi}} \\ \sum \frac{1}{2\hat{\phi}} \left(\frac{(y_t - \hat{\mu})^2}{\hat{\phi}} - 1 \right) \end{bmatrix} \text{ and}$$

$$\hat{K}_T(\gamma^1, \gamma^2) = \frac{1}{T} \sum s_t(\gamma^1; \hat{\mu}, \hat{\phi}) s_t(\gamma^2; \hat{\mu}, \hat{\phi})$$

where $\hat{\mu} = .497$ and $\hat{\phi} = 1.63$ correspond to the one regime model estimates in Table 4.5 and we abuse notation by writing $p_t^s(h_t; a, b, d)$ and $p_t^c(h_t; a, b, d)$ in section 4.3 as $p_t^s(h_t;\gamma)$ and $p_t^c(h_t;\gamma)$ here. If we denote by $\hat{S}_T(\gamma)$ a normally distributed, six dimensional mean zero random vector such that

$$\mathbb{E}\left[\hat{S}_T(\gamma^1) \hat{S}_T(\gamma^2) \right] = \hat{K}_T(\gamma^1, \gamma^2)$$

then we want to compare our test statistic's value of 98.24 to quantiles of

$$\sup_{\gamma \in \Gamma} \left\{ \hat{S}_T(\gamma)^{4'} \left[\hat{K}_T(\gamma, \gamma)^4 \right]^{-1} \hat{S}_T(\gamma)^4 - \hat{S}_T(\gamma)^{2'} \left[\hat{K}_T(\gamma, \gamma)^2 \right]^{-1} \hat{S}_T(\gamma)^2 \right\} \quad (6.8)$$

where the superscripts denote the partitioned components of the associated vectors and matrices as described in Section 6.1 and Γ corresponds to a three dimensional parameter space for (a, b, d). For any fixed γ we know the expected value of

$$\hat{Q}_T(\gamma) \stackrel{\triangle}{=} \hat{S}_T(\gamma)^{4'} \left[\hat{K}_T(\gamma, \gamma)^4 \right]^{-1} \hat{S}_T(\gamma)^4 - \hat{S}_T(\gamma)^{2'} \left[\hat{K}_T(\gamma, \gamma)^2 \right]^{-1} \hat{S}_T(\gamma)^2 \quad (6.9)$$

is $4-2=2$ because both quadratic forms have a marginal chi-square distribution. While we cannot easily determine a closed form solution for the variance of $\hat{Q}_T(\gamma)$ (because of the correlation between the chi-square terms) it is hard to imagine the variance would be large enough so that the supremum of the $\hat{Q}_T(\gamma)$ process might approach our test statistic's value of 98.24. However, in the interests of completeness we do perform our test procedure. To implement our test we need to specify a grid of points over which to search for suprema. For our grid we take

$$\begin{aligned} a \in A = &\quad \{-2.1, -1.9, \ldots, 1.9, 2.1\} \\ b \in B = &\quad \{0, \pi/10, 2\pi/10, \ldots, 2\pi\} \\ d \in D = &\quad \{-2, -1.6, -1.2, \ldots, 1.6, 2.0\} \\ \Gamma = &\quad A \times B \times D \text{ contains } 5082 \text{ points.} \end{aligned}$$

We created 100 simulations of our field and the empirical distribution of the maxima is given below. The first two figures in the table are the empirical

Mean	Variance	25%	50%	75%	90%	95%	99%	Max
4.44	6.23	2.65	4.19	5.78	7.56	9.45	11.3	13.9

Table 6.2: Empirical Suprema

distribution's mean and variance – the other figures correspond to quantiles. It is clear from these data that our test statistic of 98.24 far exceeds all the empirical maxima. From this comparison we would reject the null hypothesis that the rain data is produced by a single log-normal distribution in favor of the alternative of a 2 regime logistic mixture model with non-constant regime probabilities.

Another test we might wish to consider would examine the null hypothesis of

a mixture with constant probabilities (the 2R model in Chapter 4) against the LM model. As discussed in Chapter 4, a likelihood ratio test in this situation also suffers from identifiability problems that might be eliminated with an empirical process approach. Future work on the GATE dataset might include such analysis.

Chapter 7

Main Results and Future Work

7.1 Main Results

In this dissertation we introduced a broad class of time series mixture models. Our results were for mixtures with only two component densities. The component densities were assumed to have a GLM form and the mixture probabilities varied according to a logistic regression model. We think these models an important addition to modeling choices as they allow the analyst to include factors that may make one regime more likely than another. Threshold models have this flavor but seem to us somewhat rigid. In Chapter 2 we defined an EM algorithm approach to estimation and next showed that the estimates are consistent and asymptotically normal under a set of general conditions. In Chapter 4 we used simulations to suggest that these logistic mixture models may be superior to conventional threshold autoregressive models that yield biased estimates if the threshold variable is measured with noise.

In Chapter 5 we dealt with likelihood ratio tests for determining the presence of a logistic mixture versus the null hypothesis that the data is generated by a

single regime (i.e. no mixture). We found that because the regime probabilities are not constant it is possible to obtain the asymptotic distribution of the likelihood ratio statistic. This test necessarily excludes mixtures with constant regime probabilities from the set of mixtures under the alternative hypothesis. In Chapter 6 we used simulations to see the test had good performance under both the null and alternative hypotheses.

7.2 Future Work

We encountered several interesting questions that we think worthy of more consideration. We think there should be not much difficulty in extending the results to more than two component densities – at least this should be true for the estimation, consistency, and asymptotic normality results in Chapters 2 and 3. These models have a potentially wide range of applications. Situations in which threshold and hidden Markov models have been used should be appropriate for investigation via logistic mixtures.

We think the most interesting future work might involve the testing questions addressed in Chapters 5 and 6. As we saw in Chapter 5, the likelihood ratio test's distribution depends crucially upon the parameter space under the alternative hypothesis. Mixtures models with more than two component densities may have more complicated restrictions on the region of the parameter space that may be considered under the alternative hypothesis. For models with two components it would be useful to conduct a more thorough analysis of the test's power – particularly against other testing methods like the Monte Carlo approach discussed in Chapter 6. Of particular interest might be the tests' performance for smaller

sample sizes.

Also, we have the unfortunate caveat of excluding an interesting part of the parameter space under the alternative hypotheses (that which corresponds to mixtures with constant probabilities). We suspect it may be possible to remove this restriction and obtain a more general test. Techniques used by Dacunha-Castelle and Gassiat (1997) may be useful in this respect.

We are eager to examine our test in the context of hidden Markov model regression. In these models the regime probabilities change according to the value of an unobserved Markov process. This randomness suggests that our test may work for these models as well.

Bibliography

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B.N. Petrov and F. Czáki, editors, *Second International Symposium on Information Theory*, pages 267–281, Budapest, 1973. Akademiai Kiadó.

[2] D.W.K. Andrews. Consistency in nonlinear econometric models: A generic uniform law of large numbers. *Econometrica*, 55:1465–1471, 1987.

[3] D.W.K. Andrews. An introduction to econometric applications of functional limit theory for dependent random variables. *Econometric Reviews*, pages 183–216, 1993.

[4] D.W.K. Andrews and W. Ploberger. Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica*, 62(6):1383–1414, 1994.

[5] T.L. Bell and R. Suhasini. Principal modes of variation of rain-rate probability distributions. *Journal of Applied Meteorology*, 33(9):1067–1078, 1994.

[6] P. Billingsley. *Statistical Inference for Markov Processes.* University of Chicago Press, Chicago, 1961.

[7] P. Billingsley. *Convergence of Probability Measures.* Wiley, New York, 1968.

[8] P. Billingsley. *Probability and Measure*. Wiley, New York, second edition, 1986.

[9] K.S. Chan. A review of some limit theorems of Markov chains and their applications. In H. Tong, editor, *Dimension Estimation and Models*. World Scientific Publishing Co. Pte. Ltd, Singapore and River Edge, NJ, USA, 1993.

[10] H. Chernoff. On the distribution of the likelihood ratio. *The Annals of Mathematical Statistics*, 25:573–578, 1954.

[11] D.R. Cox. Partial likelihood. *Biometrika*, 64:269–276, 1975.

[12] H. Cramér. *Mathematical Methods of Statistics*. Princeton University Press, Princeton, 1946.

[13] D. Dacunha-Castelle and E. Gassiat. Testing in locally conic models, and application to mixture models. *ESAIM: Probability and Statistics*, 1:285–317, July 1997.

[14] D. A. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. Roy. Statist. Soc. Ser. B*, 39:1–22, 1977.

[15] W.S. Desarbo and M. Wedel. A mixture likelihood approach for generalized linear models. *Journal of Classification*, 12(1):21–56, 1995.

[16] F.X. Diebold, J. Lee, and G.C. Weinbach. Regime switching with time-varying transition probabilities. In C. Hargreaves, editor, *Nonstationary Time Series and Cointegration*. Oxford University Press, Oxford, 1994.

[17] C. Engel and J.D. Hamilton. Long swings in the dollar: Are they in the data and do markets know it? *American Economic Review*, 80:689–713, 1990.

[18] B.S. Everitt and D. J. Hand. *Finite Mixture Distributions*. Chapman and Hall, London, 1981.

[19] L. Fahrmeir and H. Kaufmann. Consistency and Asymptotic Normality of the Maximum Likelihood Estimator in Generalized Linear Models. *The Annals of Statistics*, 13:342–368, 1985.

[20] L. Fahrmeir and G. Tutz. *Multivariate Statistical Modeling Based on Generalized Linear Models*. Springer-Verlag, New York, 1994.

[21] Z.D. Feng and C.E. McCulloch. Using bootstrap likelihood ratios in finite mixture models. *Journal of the Royal Statistical Society, Series B*, 58(3):609–617, 1996.

[22] A.J. Filardo. Business cycle phases and their transitional dynamics. *Journal of Business and Economic Statistics*, 9:299–308, 1994.

[23] K. Fokianos. *Categorical Time Series: Prediction and Control*. PhD thesis, University of Maryland, College Park, 1996.

[24] M. Fridman. A two state capital asset pricing model. Technical report, Institute for Mathematics and its Applications, University of Minnesota, 1994.

[25] B. Garel. Asymptotic theory of the log-likelihood ratio test for mixtures with two components. *Comptes Rendus des Séances de l'Académie des Sciences. Série I. Mathématique*, 323(2):199–202, 1996. in French.

[26] J.K. Ghosh and P.K. Sen. On the asymptotic performance of the log-likelihood ratio statistic for the mixture model and related results. In Lucien M. Le Cam and Richard A. Olshen, editors, *Proceedings of the Berkeley Symposium in honour of Jerzy Neyman and Jack Kiefer*, volume II, pages 789 – 806, Belmont, CA, 1985. Wadsworth.

[27] F. Gong and R.S. Mariano. Testing under non-standard conditions in frequency domain: With applications to Markov regime switching models of exchange rates and the federal funds rate. Staff report, Federal Reserve Bank of New York, 1997.

[28] J.D. Hamilton. Analysis of time series subject to changes in regime. *Journal of Econometrics*, 45:39–70, 1990.

[29] J.D. Hamilton. *Time Series Analysis*. Princeton University Press, Princeton, 1994.

[30] J.D. Hamilton. Specification testing in Markov-switching time-series models. *Journal of Econometrics*, 70:127–57, 1996.

[31] B. E. Hansen. The likelihood ratio test under nonstandard conditions: testing the Markov switching model of gnp. *Journal of Applied Econometrics*, 7:S61–S82, 1992.

[32] B.E. Hansen. Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica*, 64(2):413–430, 1996a.

[33] B.E. Hansen. Stochastic equicontinuity for unbounded dependent heterogeneous arrays. *Econometric Theory*, 12:347–359, 1996b.

[34] B.E. Hansen. Erratum: The likelihood ratio test under nonstandard conditions: Testing the Markov switching model of gnp. *Journal of Applied Econometrics*, 11(2):195–198, 1996c.

[35] L.P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50:1029–1054, 1982.

[36] R.A. Houze. Structures of atmospheric precipitation: A global survey. *Radio Science*, 16:671–689, 1981.

[37] M.D. Hudlow and V.L. Patterson. *GATE Radar Rainfall Atlas*. Washington, 1979. NOAA Special Report.

[38] H. Kaufmann. Regression models for nonstationary categorical time series: Asymptotic estimation theory. *Annals of Statistics*, 15(1):79–98, 1987.

[39] B. Kedem. *Time Series Analysis by Higher Order Crossings*. Institute of Electrical and Electronics Engineers, Inc., New York, 1994.

[40] B. Kedem, L.S. Chiu, and G.R. North. Estimation of mean rain rate: Application to satellite observations. *Journal of Geophysical Research*, 95(2):1965–1972, 1990.

[41] B. Kedem, R. Pfeiffer, and D. A. Short. Variability of space-time mean rain rate. *Journal of Applied Meteorology*, 36(5):443–451, May 1997.

[42] N. M. Kiefer. Discrete parameter variation: Efficient estimation of a switching regression model. *Econometrica*, 46:427–434, 1978.

[43] A.Y.C. Kuk and C.H. Chen. A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, 79(3):531–41, 1992.

[44] M.G. Larson and G.E. Dinse. A mixture model for the regression analysis of competing risks data. *Applied Statistics*, pages 201–211, 1985.

[45] M. Lemdani and O. Pons. Likelihood ratio tests in mixture models. *Comptes Rendus des Séances de l'Académie des Sciences. Série I. Mathématique*, 322(4):399–404, 1996.

[46] B.G. Leroux. Maximum-likelihood estimation for hidden Markov models. *Stochastic Processes and their Applications*, 40:127–143, 1992.

[47] G. Lindgren. Markov regime models for mixed distribution and switching regressions. *Scandinavian Journal of Statistics*, 5:81–91, 1978.

[48] P. McCullagh. *Tensor Methods in Statistics*. Chapman and Hall, London, 1987.

[49] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. University Press, Cambridge, second edition, 1989.

[50] G. J. McLachlan and K. E. Basford. *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, Inc., New York, NY, 1988.

[51] G.J. McLachlan, K.E. Basford, and M. Green. On inferring the number of components in normal mixture models. Research report #9, Department of Mathematics, The University of Queensland Australia, June 1993.

[52] G.J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, New York, 1997.

[53] D.L. McLeish. Dependent central limit theorems and invariance principles. *Annals of Probability*, 2(4):620–628, 1974.

[54] E. Nummelin. *General Irreducible Markov Chains and Non-negative Operators*. Cambridge University Press, Cambridge, 1984.

[55] D. Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, New York, 1984.

[56] D. Pollard. *Empirical Processes: Theory and Applications*. Institute of Mathematical Statistics, Hayward, 1990.

[57] R. Quandt. The estimation of parameters of linear regression system obeying two separate regimes. *Journal of the American Statistical Association*, 55:873–880, 1958.

[58] R. Quandt and S. Goldfeld. A Markov model for switching regressions. *Journal of Econometrics*, 1:3–16, 1973.

[59] R.E. Quandt and J.B. Ramsey. Estimating mixtures of normal distributions and switching regression. *J. Amer. Statist. Assoc.*, 73:730–738, 1978.

[60] R. Redner. Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *Annals of Statistics*, 9:225–228, 1981.

[61] J. Sansom and P.J. Thomson. Rainfall classification using breakpoint pluviograph data. *Journal of Climate*, 5:755–764, 1992.

[62] P.K. Sen and J.M. Singer. *Large Sample Methods in Statistics: An Introduction with Applications*. Chapman & Hall, London, 1993.

[63] E. Slud and B. Kedem. Partial likelihood analysis of logistic regression and autoregression. *Statistica Sinica*, 4:89–106, 1994.

[64] H. Teicher. Identifiability of finite mixtures. *Annals of Mathematical Statistics*, pages 1265–1269, 1963.

[65] D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York, 1985.

[66] H. Tong. Threshold models in non-linear time series analysis. In *Lecture Notes in Statistics, 21*. Springer, Berlin, 1983.

[67] H. Tong. *Non-Linear Time Series: A Dynamical System Approach*. Oxford University Press, Oxford, 1990.

[68] T. Tweedie. Sufficient conditions for ergoditicity and recurrence of Markov chains on a general state space. *Stochastic Processes and their Applications*, 3:385–403, 1975.

[69] A. Wald. Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, 20:595–601, 1949.

[70] S.S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics*, 9:60–62, 1937.

[71] W. H. Wong. Theory of partial likelihood. *The Annals of Statistics*, 14:86–123, 1986.

[72] C. F. Wu. On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, 11:95–103, 1983.