Variable Mixtures as Alternatives to Threshold Models

Neal Jeffries *　　　Benjamin Kedem †

September 1, 1998

Key Words: Variable mixtures, mixture models, threshold models, EM algorithm

Abstract

A mixture model is presented in which the regime probabilities are allowed to change over time and are modeled with a binary regression structure. We call these variable mixture models and they may be used as alternatives to simple mixture, threshold, or hidden Markov models. They are particularly useful as alternatives to threshold models when the threshold is measured with noise. It is shown that under some circumstances such noise may induce inconsistency in the the threshold models while the variable mixture models remain consistent.

1 Introduction

In this paper we discuss 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

* National Institute of Neurological Disorders and Stroke, Bethesda, MD.

† Department of Mathematics and Institute for Systems Research, University of Maryland, College Park, MD. Research Supported by NASA grant NAG52783, NSF grant EEC-940-2384.

component distributions and modeling the probability of which distribution is applicable (the regime probability) through a regression structure. We call these variable mixture 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 process. 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. Our two component variable mixture model is characterized by the following:

- 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

$$\mathbb{P}\left[I_t = 1 \mid \mathcal{P}_{t-1}; \gamma\right] = F(Z_t'\gamma) \tag{1.1}$$

  and γ represents an unknown vector of regression parameters. For this paper we consider $F(s)$

of the form

$$F(s) = \frac{\exp(s)}{1+\exp(s)} \text{ (leading to logistic mixtures)} \tag{1.2}$$

$$F(s) = \Phi(s) \text{ where } \Phi \text{ is the cdf of a } N(0,1) \text{ variable.} \tag{1.3}$$

This second form of $F(s)$ leads to what we will call probit mixtures. In other words, I_t is the dependent variable in a binary regression model with covariate vector Z_t. It is important to note that the I_t's are *not* observed. (Throughout this work, vectors such as Z_t will be assumed column vectors with transposes denoted by Z_t'.)

- We denote the component densities of Y_t given the indicator I_t, the past $\mathcal{P}_{t-1}$ and regime specific covariates, α_i by

$$f(y_t|x_{t1};\alpha_1) \text{ if } I_t = 1,\ f(y_t|x_{t0};\alpha_0) \text{ if } I_t = 0 \tag{1.4}$$

$$\text{with } x_{t1} \text{ and } x_{t0} \text{ realizations of } X_{t1}, X_{t0} \in \mathcal{P}_{t-1}. \tag{1.5}$$

The functions $f(y_t \mid x_{ti};\alpha_i), i \in \{0,1\}$ are considered densities with respect to some σ-finite measure, ν on the real line.

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 $F(\cdot), \gamma$ and z_t, compute $\mathbb{P}\left[I_t = 1 \mid W_t, Y_{t-1}, W_{t-1}, ..., Y_{t-p}, W_{t-p}\right] = F(z_t'\gamma)$.
- Generate a Bernoulli random variable, I_t, with mean given by the probability above.
- If $I_t = 1$ generate Y_t from $f(y_t|x_{t1};\alpha_1)$; if $I_t = 0$ generate Y_t from $f(y_t|x_{t0};\alpha_0)$.

The variable probabilities in (1.1) allow us to capture important effects covariates may have on the regime probability. This is an important and useful extension of simple 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.

This modeling of variable mixtures is not entirely new – Filardo (1994) and Diebold et. al. (1994) have employed logistic regression models within the transition matrix of a hidden Markov model and Kuk and Chen (1992), and Larson and Dinse (1985) have employed logistic mixtures in the context of mixtures of hazard rates. However, we believe that neither asymptotic results nor testing for whether a mixture is present has been addressed in the context of variable mixtures. Both of these issues have been developed at length in Jeffries (1998) for logistic mixtures. In this paper we focus upon a link between variable mixtures and two state threshold models that has not yet been detailed.

Threshold Models

As a basis of comparison to variable mixtures we consider 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. Once values of $\widehat{p}, \widehat{d}, \widehat{\tau}$, $\widehat{\xi}_{01}, \ldots, \widehat{\xi}_{p0}$ have been determined, estimates of $\hat{\sigma}_1^2$ and $\hat{\sigma}_0^2$ are derived in an obvious manner as the associated OLS or MLE estimates. 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.

Partial Likelihood

Throughout this paper 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 in turn based 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

$$\prod_{t=1}^{T} g(y_t, w_t \mid y_{t-1} w_{t-1}, ..., y_{-p}, w_{-p}) = \prod_{t=1}^{T} g(y_t \mid d_t) \prod_{t=1}^{T} g(w_t \mid c_t) \tag{1.6}$$

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 (1.6) is the partial likelihood. Slud and Kedem (1994) provides a more formal definition which includes parameters for a conditional density.

Definition *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 $g(z_t \mid \mathcal{G}_{t-1}; \theta)$ where θ represents a vector of parameters. The partial likelihood function relative to $\theta, \mathcal{G}_{t-1}$, and the data $z_1, z_2, ..., z_T$ is given by*

$$\prod_{t=1}^{T} g(z_t \mid \mathcal{G}_{t-1}; \theta). \tag{1.7}$$

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 (1.6). 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. The vector $\hat{\theta}$ that maximizes (1.7) 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.

Overview of Presentation

In Section 2 we describe an EM algorithm approach to finding maximum likelihood estimates (or more precisely, maximum partial likelihood estimates) for the parameters in the two component distributions as well as those parameters in the variable regression that describe the regime probabilities. Furthermore, we state the large sample properties of a correctly specified logistic mixture model. As might be expected is is possible to demonstrate consistency and asymptotic normality under some very general assumptions. In Section 3 we illustrate estimation with some simulations and show the variable mixture models 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. In Section 4 we consider a specific class of models and prove the inconsistency of threshold model estimation and consistency of probit mixtures.

2 Estimation of the Variable Mixture Model

We define $\theta = (\alpha_1, \alpha_0, \gamma)$. From our representations of the conditional densities and the regime probabilities we may write the conditional distribution of Y_t given the observed past, $\mathcal{P}_{t-1}$, as

$$g(y_t \mid \mathcal{P}_{t-1}; \theta) = f(y_t \mid x_{t1}; \alpha_1)\mathbb{P}\left[I_t = 1 \mid z_t; \gamma\right] + f(y_t \mid x_{t0}; \alpha_0)\mathbb{P}\left[I_t = 0 \mid z_t; \gamma\right]. \tag{2.1}$$

We call these variables mixtures because $\mathbb{P}[I_t = 1 \mid z_t; \gamma]$ changes as z_t changes. It is the product of such terms that we define as our partial likelihood evaluated at θ, $\prod g(y_t \mid \mathcal{P}_{t-1}; \theta)$

Our goal is to estimate θ from our incomplete knowledge of the process (i.e. $\{I_t\}$ is not observed). Here we discuss the EM algorithm popularized by Dempster, Laird, and Rubin (1977) – see also Wu (1983) or McLachlan and Krishnan (1997). It should be noted that an EM approach is not necessary – the parameters could be estimated by software that has minimization or maximization routines. We want to point out though, that $L_T(\theta) = \sum_t \log g(y_t | x_{t1}, x_{t0}, z_t; \theta)$ will not be a concave function because it is a mixture likelihood, and thus the Newton-Raphson based optimizations may not work well. The advantage of using an EM algorithm lies in the fact that the likelihood values increase (weakly) with each iteration – thus ensuring the algorithm will converge to a local maximum in almost all cases. This means that if θ^k denotes the EM algorithm's estimate of the true parameters θ^* after k iterations, then the EM-estimates satisfy $L_T(\theta^{k+1}) \geq L_T(\theta^k)$. This monotone property of the EM-estimates, that is clearly not true for estimates obtained by Newton-Raphson, becomes more important as the number of estimated parameters increases (either because the model may have more than two component densities, or the component densities are extended to have more parameters). As more parameters are added it is more likely that standard maximization routines (e.g. Newton-Raphson based methods) will fail. There are pathological constructions in which the EM-estimates may converge to a critical point other than a local maximum, or other difficulties may occur (see Wu (1983) and McLachlan and Basford (1988)), but such aberrations are usually overcome by changing the starting values of the algorithm.

The implementation and interpretation of the EM algorithm in the context of mixture models has been discussed by several authors including Titterington et. al. (1985), McLachlan and

Basford (1988), and McLachlan and Krishnan (1997). The derivation of the algorithm for logistic mixtures (i.e. variable probabilities of the form $\exp(z_t'\gamma)/(1+\exp(z_t'\gamma)))$ is straightforward, detailed in Jeffries (1998), and will not be presented here. The procedure for general variable mixtures may be summarized as follows: given observed outcomes $y_1, y_2, \ldots, y_T$ and covariates $z_1, z_2, \ldots, z_T, x_{11}, \ldots, x_{T1}, x_{10}, \ldots, x_{T0}$,

1. Choose initial parameter values $\theta^1 = (\alpha_0^1, \alpha_1^1, \gamma^1)$.

2. Given $\theta^k = (\alpha_0^k, \alpha_1^k, \gamma^k)$ (the estimate after k iterations of the algorithm) calculate

$$\tilde{p}_t^k = \frac{f\left(y_t \mid x_{t1}; \alpha_1^k\right) F(z_t'\gamma^k)}{f\left(y_t \mid x_{t1}; \alpha_1^k\right) F(z_t'\gamma^k) + f\left(y_t \mid x_{t0}; \alpha_0^k\right)(1 - F(z_t'\gamma^k))}, t = 1...T$$

   where $F(\cdot)$ corresponds to either the logistic, probit, or some other cdf for binary data.

3. Given $\{\tilde{p}_t^k\}$ find $\alpha_0^{k+1}, \alpha_1^{k+1}$, and γ^{k+1} where

$$\alpha_1^{k+1} = \arg\max_{\alpha} \sum_t \tilde{p}_t^k \log f(y_t \mid x_{t1}; \alpha) \tag{2.2}$$

$$\alpha_0^{k+1} = \arg\max_{\alpha} \sum_t (1 - \tilde{p}_t^k) \log f(y_t \mid x_{t0}; \alpha) \text{ and} \tag{2.3}$$

$$\gamma^{k+1} = \arg\max_{\gamma} \sum_t \tilde{p}_t^k \log F(z_t'\gamma) + (1 - \tilde{p}_t^k) \log(1 - F(z_t'\gamma)). \tag{2.4}$$

4. Repeat steps 2) and 3) until $\|\theta^{k+1} - \theta^k\|$ and

$$L_T(\theta^{k+1}) - L_T(\theta^k) = \sum_t \log g(y_t \mid z_t, x_{t1}, x_{t0}; \theta^{k+1}) - \sum_t \log g(y_t \mid z_t, x_{t1}, x_{t0}; \theta^k)$$

   are smaller than some prespecified tolerance level.

Several different starting values should be used, and the maximum likelihood estimate of θ will correspond to that associated with the largest value of $\lim_k L_T(\theta^k)$ that was obtained from the different starting values.

We note that the maximizations in equations (2.2) and (2.3) are generally easy to perform. In particular, if the $f(y_t \mid x_{ti}; \alpha)$ are of a generalized linear model family (e.g. normal, poisson, beta, binomial) many statistical software packages will perform the maximizations though a routine generalized linear regression command that accepts $(\tilde{p}_1^k, \ldots, \tilde{p}_T^k)$ and $(1-\tilde{p}_1^k, \ldots, 1-\tilde{p}_T^k)$ as vectors of weights. Similarly, the estimates of γ^{k+1} could be mechanically obtained by performing logistic regression with $\tilde{p}_1^k, \ldots, \tilde{p}_T^k$ as the dependent variables and $z_1, \ldots, z_T$ as the covariates.

Alternatively, a Newton-Raphson procedure or Fisher scoring algorithm could be used to perform the maximizations in (2.2–2.4). In many cases (e.g. if $f(y_t \mid x_{ti}; \alpha)$ are of a generalized linear model family – see Fahrmeir and Tutz (1994)) the objective functions will be concave and therefore the Newton-Raphson based algorithms work well in finding the optima.

Precision of the Estimates

Thus far we have only considered finding the maximum likelihood estimates, i.e.

$$\hat{\theta}_T = \arg\max_{\theta} \sum_{t}^{T} \log g(y_t | x_{t1}, x_{t0}, z_t; \theta). \tag{2.5}$$

Whether these are obtained through the EM or other algorithms, we now consider how to obtain standard errors for $\hat{\theta}_T$. Given that $(Y_t, X_{t1}, X_{t0}, Z_t)$ are geometrically ergodic (see Tweedie (1975), Nummelin (1984), Chan (1993), or Chan (1990) for discussion) then under general conditions it may be shown that

$$\sqrt{T}\left(\hat{\theta}_T - \theta^*\right) \xrightarrow{\mathcal{D}} N\left(\underline{0}, Q^{-1}\right) \tag{2.6}$$

where $\theta^* = (\alpha_0^*, \alpha_1^*, \gamma^*)$ denotes the true parameters and

$$[Q]_{ij} = \mathbb{E}\left[\frac{\partial \log g\left(Y_t \mid X_{t1}, X_{t0}, Z_t; \theta\right)}{\partial \theta_i} \frac{\partial \log g\left(Y_t \mid X_{t1}, X_{t0}, Z_t; \theta\right)}{\partial \theta_j} \bigg|_{\theta^*} \right].$$

Such conditions are detailed in Jeffries (1998). From the result in (2.6) we use the following relationship for determining standard errors of our estimates:

$$\sqrt{T}\left(\hat{\theta}_T-\theta^*\right)\sim N\left(\underline{0},\hat{Q}_T^{-1}\right) \tag{2.7}$$

where

$$\left[\hat{Q}_T\right]_{ij}=\frac{1}{T}\sum_t\left[\frac{\partial\log g\left(y_t\mid x_{t1},x_{t0},z_t;\theta\right)}{\partial\theta_i}\frac{\partial\log g\left(y_t\mid x_{t1},x_{t0},z_t;\theta\right)}{\partial\theta_j}\Bigg|_{\hat{\theta}_T}\right]$$

and "$\sim$" means "approximately distributed". The basis for this approximation is the fact that $\hat{Q}_T\xrightarrow{a.s.}Q$ under general conditions.

3 Comparison to Threshold Model

We have suggested above that simple threshold models may yield inconsistent results if the threshold variable is measured with error (or equivalently, the threshold is not fixed but randomly fluctuates about some level in an unobservable manner). To illustrate this point we present a simulation of the following model: Let

$$Y_t=\begin{cases}.6Y_{t-1}+.5*\epsilon_t \text{ if } W_{t-1}<.1\\ \\ .1Y_{t-1}+.3*\epsilon_t \text{ if } W_{t-1}\geq .1\end{cases}$$

where $W_t=Y_t+.8\eta_t$ and both η_t and ϵ_t are i.i.d. $N(0,1),\epsilon_t$ independent of η_s for all s.

We view this as a two state threshold model with threshold variable W_t and threshold .1. For each trial in our simulation, 500 Y_t were generated according to the process above with $Y_0=0$. A total of 100 trials were conducted. In modeling this process, suppose that W_t is not observed, and only Y_t is available. Consequently the analyst might fit a TAR model:

$$Y_t=\begin{cases}\beta_1Y_{t-1}+\sigma_1*\epsilon_t \text{ if } Y_{t-1}<\tau\\ \\ \beta_0Y_{t-1}+\sigma_0*\epsilon_t \text{ if } Y_{t-1}\geq\tau.\end{cases}$$

As an alternative to fitting a TAR model one may also fit logistic and probit mixture models to the same data. In these instances the variable mixtures correspond to the model:

$$Y_t = \begin{cases} \beta_1 Y_{t-1} + \sigma_1 * \epsilon_t \text{ if } I_t = 1 \\ \\ \beta_0 Y_{t-1} + \sigma_0 * \epsilon_t \text{ otherwise, and} \end{cases}$$

$$\mathbb{P}\left[I_t = 1 \mid Y_{t-1}\right] = F(\gamma_0 + \gamma_1 Y_{t-1}) \text{ with}$$

$$F(s) = \frac{\exp(s)}{1+\exp(s)} \text{ for the logistic}$$

$$F(s) = \Phi(s) \text{ for the probit mixture.}$$

We emphasize that these variable mixture models also use Y_{t-1} as a proxy for the unobserved W_{t-1} to discriminate between the two states. In Table 1 we see the estimates of the TAR model (labeled TAR(Y_{t-1})) appear biased and possibly inconsistent, while those derived under the probit (PM(Y_{t-1})) and logistic (LM(Y_{t-1})) seem more accurate. The standard errors reported in parentheses underneath the parameter estimates were empirically derived as the sample standard deviation for the 100 trials.

To explain the relatively poor results of the TAR, let τ be given and define C_1^τ to be the set of observations for which $Y_{t-1} < \tau$ and C_0^τ be the set for which $Y_{t-1} \geq \tau$. We suspect that because the threshold variable W_{t-1} is not directly observed, its imperfect proxy, Y_{t-1} occasionally misclassifies observations into C_1^τ and C_0^τ. That is, if we denote by f_1 the conditional density associated with $W_{t-1} < .1$ and f_0 the density associated with $W_{t-1} \geq .1$ the class C_1^τ incorrectly contains some observations that were generated by f_0. Consequently, the estimates of β_1 and σ_1 will be biased because they are not based on the appropriate set of observations. As the rate of misclassification will not decrease with sample size the estimates are inconsistent as well. A similar consequence holds

Parameter	True	TAR(Y_{t-1})	PM(Y_{t-1})	LM(Y_{t-1})
β_1	.6	.532	.585	.586
		(.06)	(.08)	(.08)
σ_1^2	.25	.209	.252	.250
		(.02)	(.04)	(.04)
β_0	.1	.200	.101	.094
		(.06)	(.08)	(.09)
σ_0^2	.09	.154	.086	.084
		(.03)	(.03)	(.03)

Table 1: 100 Simulations of 500 Observations Using Y_{t-1} as Threshold Variable

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 W_{t-1} instead of Y_{t-1} in fitting the various models. For the TAR model this corresponds to the occurrence that the analyst observers the threshold variable, W_{t-1}, without error. The purpose of this table is to illustrate the degree to which the variable mixture models are inferior to a correctly specified threshold model. In computing the variable mixture estimates, W_{t-1} was used in fitting the logistic and probit regressions. The results suggest little precision is lost in estimating the parameters

Parameter	True	TAR(W_{t-1})	PM(W_{t-1})	LM(W_{t-1})
β_1	.6	.603	.599	.597
		(.05)	(.06)	(.06)
σ_1^2	.25	.252	.254	.255
		(.02)	(.02)	(.03)
β_0	.1	.101	.100	.102
		(.04)	(.05)	(.05)
σ_0^2	.09	.096	.088	.089
		(.01)	(.01)	(.01)

Table 2: 100 Simulations of 500 Observations Using W_{t-1} as Threshold Variable

of the component densities through a variable mixture – while much may be gained with variable mixtures if the TAR model is misspecified.

4 Analytic Comparison of Threshold and Variable Mixture Models

Here we examine a threshold model analytically, to explain the empirical results described above. Throughout this section our data is generated by the following model

$$
\begin{aligned}
Y_t \sim & \begin{cases} f(y_t \mid y_{t-1}; \alpha_1^*) \text{ if } I_t = 1 \\ f(y_t \mid y_{t-1}; \alpha_0^*) \text{ if } I_t = 0, \text{ and} \end{cases} \\
I_t =& 1 \text{ if } Y_{t-1} + \phi^* V_{t-1} < \tau^* \\
I_t =& 0 \text{ if } Y_{t-1} + \phi^* V_{t-1} \geq \tau^*.
\end{aligned}
\tag{4.1}
$$

It is assumed that α_1^* and α_0^* are not equal. Here V_t is i.i.d. $N(0,1)$, ϕ^* is a constant controlling variance of the noise, $f(y_t \mid y_{t-1}; \alpha), \alpha \in A \in \mathbb{R}^m$ denote a class of parametric densities with support equal to $\mathbb{R}$ when viewed as a function of y_t. V_t is unobservable and is independent of Y_s for all $s \leq t$. As an example, we may take $f(y_t \mid y_{t-1}; \alpha)$ corresponding to an AR(1) model, i.e.

$$
Y_t = \beta Y_{t-1} + \sigma \epsilon_t,\ \alpha = (\beta, \sigma),\ \text{ and } \epsilon_t \text{ i.i.d. } N(0,1). \tag{4.2}
$$

We assume the analyst is most interested in estimating the parameters of the component densities, α_1^* and α_0^*. Two models are considered in fitting data generated as above:

1. A threshold model

$$
Y_t \sim \begin{cases} f(y_t \mid y_{t-1}; \alpha_1) \text{ if } I_t = 1 \\ f(y_t \mid y_{t-1}; \alpha_0) \text{ if } I_t = 0, \text{ and } I_t = 1 \textit{ iff } Y_{t-1} < \tau. \end{cases} \tag{4.3}
$$

2. A probit mixture

$$
Y_t \sim \begin{cases} f(y_t \mid y_{t-1}; \alpha_1) \text{ if } I_t = 1 \\ f(y_t \mid y_{t-1}; \alpha_0) \text{ if } I_t = 0, \text{ and } \mathbb{P}\left[I_t = 1 \mid Y_{t-1}\right] = \Phi\left(\gamma_0 + \gamma_1 Y_{t-1}\right). \end{cases} \tag{4.4}
$$

In the threshold model we estimate α_1, α_0, and τ, while in the probit mixture α_1, α_0, and $\gamma = (\gamma_0, \gamma_1)$ are estimated. We are interested in determining when the maximum likelihood estimates of α_1 and α_0 in our two models converge to α_1^* and α_0^*, the true values of our data generation process.

Theorem 1. *Assume* (Y_t, V_t) *evolve as described in* (4.1). *Given* α_1, α_0, *and* τ *define*

$$g(y_t, y_{t-1}; \alpha_1, \alpha_0, \tau) = f(y_t \mid y_{t-1}; \alpha_1)^{I_{[y_{t-1}<\tau]}} f(y_t \mid y_{t-1}; \alpha_0)^{1-I_{[y_{t-1}<\tau]}} \text{ and} \tag{4.5}$$

$$\hat{\alpha}_1, \hat{\alpha}_0, \hat{\tau} = \arg \max_{\alpha_1, \alpha_0, \tau} \sum_{t=1}^{T} \log g(y_t, y_{t-1}; \alpha_1, \alpha_0, \tau). \tag{4.6}$$

Then under general conditions there exist constants α_1° *and* α_0° *such that* $\hat{\alpha}_1 \xrightarrow{P} \alpha_1^\circ$, $\hat{\alpha}_0 \xrightarrow{P} \alpha_0^\circ$, *but* $\alpha_1^\circ \neq \alpha_1^*$ *and* $\alpha_0^\circ \neq \alpha_0^*$. *In other words, the estimates of the parameters in the component densities are inconsistent under the misspecified threshold model.*

Proof. The family of $g(y_t, y_{t-1}; \alpha_1, \alpha_0, \tau)$ functions obtained from different values of α_1, α_0, and τ correspond to different conditional distributions of Y_t given Y_{t-1} under the threshold model as described in (4.3). The proof of this theorem is based on making rigorous the argument in the previous section regarding the misclassification of observations. If we define $W_t = (Y_t, Y_{t-1}, V_{t-1})$ then it follows that W_t is a Markov chain in $\mathbb{R}^3$. Further, we assume that Y_t, V_t are such that W_t is aperiodic and irreducible with respect to Lebesgue measure on $\mathbb{R}^3$. Chan (1993) and Chan (1990) provide theorems based on Nummelin (1984) and Tweedie (1975) which give conditions whereby W_t is geometrically ergodic. This ergodicity implies there exists a random vector $W = (Y, Y_{-1}, V_{-1})$ such that

1. $W_t \xrightarrow{D} W$, and
2. if $l(W)$ is an integrable function of W then $1/T \sum l(W_t) \xrightarrow{a.s.} \mathbb{E}[l(W)]$.

We will write $\mathbb{E}[l(W)]$ as $\int\int\int l(y, y_{-1}, v_{-1})\mathbb{P}(dy, dy_{-1}, dv_{-1})$ to indicate integration with respect to this stationary distribution. We will abuse notation somewhat and write $\int l^*(y_{-1})\mathbb{P}(dy_{-1})$ as shorthand for $\int\int\int l^*(y_{-1})\mathbb{P}(dy, dy_{-1}, dv_{-1})$ to indicate integration with respect to the marginal distribution of a component of W, in this case Y_{-1}.

Sufficient conditions for the existence of the stationary distribution and geometric ergodicity are given in the aforementioned works – particularly in the Chan references. Here we make the assumption that such conditions are met. Now, recalling our definition of the estimators in (4.6) we can use the strong law of large numbers above and an adaptation of Wald's approach to consistency (see Wald (1949) and Jeffries (1998)) to show that under very general conditions $\hat{\alpha}_1 \xrightarrow{P} \alpha_1^\circ$, $\hat{\alpha}_0 \xrightarrow{P} \alpha_0^\circ$, and $\hat{\tau} \xrightarrow{P} \tau^\circ$ where $(\alpha_1^\circ, \alpha_0^\circ, \tau^\circ)$ are parameters that satisfy the following:

$$\alpha_1^\circ, \alpha_0^\circ, \tau^\circ = \arg \max_{\alpha_1, \alpha_0, \tau} \mathbb{E}[\log g(Y, Y_{-1}; \alpha_1, \alpha_0, \tau)] \tag{4.7}$$

$$= \arg \max_{\alpha_1, \alpha_0, \tau} \int \log g(y, y_{-1}; \alpha_1, \alpha_0, \tau)\mathbb{P}(dy, dy_{-1}, dv_{-1}). \tag{4.8}$$

(We assume $\mathbb{E}[\log g(Y, Y_{-1}; \alpha_1, \alpha_0, \tau)]$ is finite for all α_1, α_0, τ under consideration.) Consequently, to prove our theorem we need only show that $\alpha_1^\circ \neq \alpha_1^*$ and $\alpha_0^\circ \neq \alpha_0^*$.

Toward this end we first find an analytic expression for $\mathbb{E}[\log g(Y, Y_{-1}; \alpha_1, \alpha_0, \tau)]$. Define

$$\begin{aligned} h(y_t, y_{t-1}, w_t; \alpha_1^*, \alpha_0^*, \tau^*, \phi^*) &= f(y_t \mid y_{t-1}; \alpha_1^*)^{I_{[y_{t-1}+\phi^* v_{t-1}<\tau^*]}} f(y_t \mid y_{t-1}; \alpha_0^*)^{1-I_{[y_{t-1}+\phi^* v_{t-1}<\tau^*]}} \\ &= I_{[y_{t-1}+\phi^* v_{t-1}<\tau^*]} f(y_t \mid y_{t-1}; \alpha_1^*) + \\ &\quad \left(1 - I_{[y_{t-1}+\phi^* v_{t-1}<\tau^*]}\right) f(y_t \mid y_{t-1}; \alpha_0^*). \end{aligned} \tag{4.9}$$

This corresponds to the true conditional density of Y_t given Y_{t-1} and V_{t-1}. As (Y, Y_{-1}, V_{-1}) has the stationary distribution for the Markov chain, this function also describes the conditional density of

Y given Y_{-1} and V_{-1}. Then for any α_1, α_0, and τ we have

$$\mathbb{E}[\log g(Y, Y_{-1}; \alpha_1, \alpha_0, \tau)] = \mathbb{E}[\mathbb{E}[\log g(Y, Y_{-1}; \alpha_1, \alpha_0, \tau) \mid Y_{-1}, V_{-1}]] = \tag{4.10}$$

$$\int\int\int \log g(y, y_{-1}; \alpha_1, \alpha_0, \tau) h(y, y_{-1}, v_{-1}; \alpha_1^*, \alpha_0^*, \tau^*, \phi^*) dy \mathbb{P}(dy_{-1}, dv_{-1}) = \tag{4.11}$$

$$\int\int\int I_{[y_{-1}+\phi^* v_{-1} < \tau^*]} \log g(y, y_{-1}; \alpha_1, \alpha_0, \tau) f(y \mid y_{-1}; \alpha_1^*) dy \mathbb{P}(dy_{-1}, dv_{-1}) + \tag{4.12}$$

$$\int\int\int I_{[y_{-1}+\phi^* v_{-1} \geq \tau^*]} \log g(y, y_{-1}; \alpha_1, \alpha_0, \tau) f(y \mid y_{-1}; \alpha_0^*) dy \mathbb{P}(dy_{-1}, dv_{-1}). \tag{4.13}$$

Now, from our definition of $g(y, y_{-1}; \alpha_1, \alpha_0, \tau)$ we have

$$\log g(y, y_{-1}; \alpha_1, \alpha_0, \tau) = I_{[y_{-1} < \tau]} \log f(y \mid y_{-1}; \alpha_1) + I_{[y_{-1} \geq \tau]} \log f(y \mid y_{-1}; \alpha_0). \tag{4.14}$$

So we may rewrite the term in (4.12) as

$$\int\int\int I_{[y_{-1}+\phi^* v_{-1} < \tau^*, y_{-1} < \tau]} \log f(y \mid y_{-1}; \alpha_1) f(y \mid y_{-1}; \alpha_1^*) dy \mathbb{P}(dy_{-1}, dv_{-1}) + \tag{4.15}$$

$$\int\int\int I_{[y_{-1}+\phi^* v_{-1} < \tau^*, y_{-1} \geq \tau]} \log f(y \mid y_{-1}; \alpha_0) f(y \mid y_{-1}; \alpha_1^*) dy \mathbb{P}(dy_{-1}, dv_{-1}) \tag{4.16}$$

We may write a similar expression for (4.13). After integrating with respect to dy we obtain

$$\mathbb{E}[\log g(y, y_{-1}; \alpha_1, \alpha_0, \tau)] = \tag{4.17}$$

$$\int\int I_{[y_{-1}+\phi^* v_{-1} < \tau^*, y_{-1} < \tau]} D(y_{-1}, \alpha_1, \alpha_1^*) \mathbb{P}(dy_{-1}, dv_{-1}) + \tag{4.18}$$

$$\int\int I_{[y_{-1}+\phi^* v_{-1} < \tau^*, y_{-1} \geq \tau]} D(y_{-1}, \alpha_0, \alpha_1^*) \mathbb{P}(dy_{-1}, dv_{-1}) + \tag{4.19}$$

$$\int\int I_{[y_{-1}+\phi^* v_{-1} \geq \tau^*, y_{-1} < \tau]} D(y_{-1}, \alpha_1, \alpha_0^*) \mathbb{P}(dy_{-1}, dv_{-1}) + \tag{4.20}$$

$$\int\int I_{[y_{-1}+\phi^* v_{-1} \geq \tau^*, y_{-1} \geq \tau]} D(y_{-1}, \alpha_0, \alpha_0^*) \mathbb{P}(dy_{-1}, dv_{-1}), \text{ where} \tag{4.21}$$

$$D(y_{-1}, \alpha, \alpha^*) = \int \log f(y \mid y_{-1}; \alpha) f(y \mid y_{-1}; \alpha^*) dy. \tag{4.22}$$

This is our expression for $\mathbb{E}[\log g(Y, Y_{-1}; \alpha_1, \alpha_0, \tau)]$ where the indicator functions $I_{[y_{-1}+\phi^* v_{-1}<\tau^*, y_{-1}\geq\tau]}$ and $I_{[y_{-1}+\phi^* v_{-1}\geq\tau^*, y_{-1}<\tau]}$ are the misclassification events that were informally discussed in the previous section. As we turn to the question of what values of $(\alpha_1, \alpha_0, \tau)$ maximize this expectation we address some properties of the $I_{[y_{-1}+\phi^* v_{-1}<\tau^*, y_{-1}\geq\tau]}$ and $D(y_{-1}, \alpha, \alpha^*)$ terms that simplify analysis.

Now, because Y_{-1} and V_{-1} are assumed to have support equal to the whole real line we have that the misclassification probabilities defined as

$$\begin{aligned} P_0(\tau) &= \int\int I_{[y_{-1}+\phi^* v_{-1}<\tau^*, y_{-1}\geq\tau]} \mathbb{P}(dy_{-1}, dv_{-1}) \text{ and} \\ P_1(\tau) &= \int\int I_{[y_{-1}+\phi^* v_{-1}\geq\tau^*, y_{-1}<\tau]} \mathbb{P}(dy_{-1}, dv_{-1}) \end{aligned}$$

are non-zero continuous function of τ. We will assume τ and hence τ° lie in some compact set Γ, thus implying there exists a constant c such that

$$P_0(\tau) \geq c > 0, \text{ and } P_1(\tau) \geq c > 0$$

for all $\tau \in \Gamma$. The point is to see that $P_0(\tau)$ and $P_1(\tau)$ are strictly positive for all τ including τ°.

Now we examine $D(y_{-1}, \alpha, \alpha^*)$ as defined in (4.22). We assume the Kullback-Leibler information inequality holds for almost all y_{-1}, i.e.

$$\begin{aligned} D(y_{-1}, \alpha, \alpha^*) < D(y_{-1}, \alpha^*, \alpha^*) \text{ for almost all } y_{-1} \text{ if } \alpha \neq \alpha^*, \text{ and} \\ \left.\frac{\partial D(y_{-1}, \alpha, \alpha^*)}{\partial \alpha}\right|_{\alpha=\alpha^*} = \underline{0}. \end{aligned} \tag{4.23}$$

The relevant distribution for almost all y_{-1} is that described by $\mathbb{P}(dy_{-1})$. This strict inequality corresponds to an identifiability condition on $f(y \mid y_{-1}; \alpha)$ and is easy to verify in many cases – in practical applications, the exceptional set will often correspond to the case of $y_{-1} = 0$.

Next, we assume that the first order conditions for a maximum are met with $\alpha_1 = \alpha_1^\circ$, $\alpha_0 = \alpha_0^\circ$, and $\tau = \tau^\circ$. In the case of α_1° this means

$$\left.\frac{\partial \mathbb{E}\left[\log g(Y, Y_{-1}; \alpha_1, \alpha_0, \tau)\right]}{\partial \alpha_1}\right|_{\alpha_1^\circ} = \tag{4.24}$$

$$\left.\frac{\partial\left\{\int\int I_{[y_{-1}+\phi^* v_{-1}<\tau^*, y_{-1}<\tau]} D(y_{-1}, \alpha_1, \alpha_1^*)\mathbb{P}(dy_{-1}, dv_{-1})\right\}}{\partial \alpha_1}\right|_{\alpha_1^\circ} + \tag{4.25}$$

$$\left.\frac{\partial\left\{\int\int I_{[y_{-1}+\phi^* v_{-1}\geq\tau^*, y_{-1}<\tau]} D(y_{-1}, \alpha_1, \alpha_0^*)\mathbb{P}(dy_{-1}, dv_{-1})\right\}}{\partial \alpha_1}\right|_{\alpha_1^\circ} = \underline{0}. \tag{4.26}$$

If we evaluate this same derivative at α_1^* we obtain

$$\left.\frac{\partial \mathbb{E}\left[\log g(Y, Y_{-1}; \alpha_1, \alpha_0, \tau)\right]}{\partial \alpha_1}\right|_{\alpha_1^*} = \left.\frac{\partial\left\{\int\int I_{[y_{-1}+\phi^* v_{-1}\geq\tau^*, y_{-1}<\tau]} D(y_{-1}, \alpha_1, \alpha_0^*)\mathbb{P}(dy_{-1}, dv_{-1})\right\}}{\partial \alpha_1}\right|_{\alpha_1^*} \tag{4.27}$$

because the term corresponding to that in (4.25) is zero due to the second assumption in (4.23) (and the assumption we may freely exchange differentiation and integration). We would like that the right hand side of (4.27) is non-zero but we need be a little careful. The pair of conditions in assumption (4.23) are not sufficient to conclude that

$$\int\int I_{[y_{-1}+\phi^* v_{-1}\geq\tau^*, y_{-1}<\tau]} \left.\frac{\partial D(y_{-1}, \alpha_1, \alpha_0^*)}{\partial \alpha_1}\right|_{\alpha_1^*} \mathbb{P}(dy_{-1}, dv_{-1}) \neq \underline{0}. \tag{4.28}$$

In many instances this last relationship can be easily verified for specific forms of $f(y \mid y_{-1}; \alpha)$. For example, if $f(y \mid y_{-1}; \alpha)$ corresponds to an AR(1) model with autoregressive parameter α (.i.e $Y_t = \alpha * Y_{t-1} + \epsilon_t$ with ϵ_t normally distributed with mean 0 and known variance σ^2) then

$$\frac{\partial D(y_{-1}, \alpha, \alpha^*)}{\partial \alpha} = \frac{(\alpha^* - \alpha)y_{-1}^2}{\sigma^2}.$$

So if $\alpha_1 < \alpha_1^*$ we have that the left hand side of (4.28) is necessarily positive if $y_{-1} \neq 0$. Extending this example to account for situations in which α contains scale parameters is straightforward.

So assuming the condition in (4.28) is met we are able to conclude that

$$\left.\frac{\partial \mathbb{E}[\log g(Y, Y_{-1}; \alpha_1, \alpha_0, \tau)]}{\partial \alpha_1}\right|_{\alpha_1^*} \neq \underline{0}$$

and thus α_1° is not equal to α_1^* and $\hat{\alpha}_1$ has been shown to be an inconsistent estimator of α_1^*. A similar development may be performed for $\hat{\alpha}_0$ and α_0°. □

This proof helps explain the inconsistent estimates we observed in Table 1. To explain the relatively good performance of the probit mixture model we give the following theorem.

Theorem 2. *Let (Y_t, V_t) evolve as described in (4.1). Given α_1, α_0, and $\gamma = (\gamma_0, \gamma_1)$ define*

$$k(y_t, y_{t-1}; \alpha_1, \alpha_0, \gamma) = f(y_t \mid y_{t-1}; \alpha_1)\Phi(\gamma_0 + \gamma_1 y_{t-1}) + f(y_t \mid y_{t-1}; \alpha_0)\left(1 - \Phi(\gamma_0 + \gamma_1 y_{t-1})\right),$$

where $\Phi(\cdot)$ is the cdf for a standard normal random variable. Then $\hat{\alpha}_1 \xrightarrow{P} \alpha_1^*$ and $\hat{\alpha}_0 \xrightarrow{P} \alpha_0^*$ where

$$\hat{\alpha}_1, \hat{\alpha}_0, \hat{\gamma} = \arg\max_{\alpha_1, \alpha_0, \gamma} \sum k(y_t, y_{t-1}; \alpha_1, \alpha_0, \gamma).$$

Proof. Here the family of $k(y_t, y_{t-1}; \alpha_1, \alpha_0, \gamma)$ functions obtained from different values of α_1, α_0, and γ correspond to different conditional distributions of Y_t given Y_{t-1} under the probit mixture model as described in (4.4). As before, we take as given that $\hat{\alpha}_1 \xrightarrow{P} \alpha_1^\circ$, $\hat{\alpha}_0 \xrightarrow{P} \alpha_0^\circ$, $\hat{\gamma} \xrightarrow{P} \gamma^\circ$ where

$$\alpha_1^\circ, \alpha_0^\circ, \gamma^\circ = \arg\max_{\alpha_1, \alpha_0, \gamma} \mathbb{E}[\log k(Y, Y_{-1}; \alpha_1, \alpha_0, \gamma)]. \tag{4.29}$$

We will show $\alpha_1^\circ = \alpha_1^*$, $\alpha_0^\circ = \alpha_0^*$, and $\gamma^\circ = \gamma^*$ where $\gamma^* = (\tau^*/\phi^*, -1/\phi^*)$. We assume $\alpha_1^\circ, \alpha_0^\circ$ and γ° are unique. We note that for any $\tilde{\alpha}_1, \tilde{\alpha}_0$, and $\tilde{\gamma}$,

$$\log k(y, y_{-1}; \tilde{\alpha}_1, \tilde{\alpha}_0, \tilde{\gamma}) = \log k(y, y_{-1}; \tilde{\alpha}_0, \tilde{\alpha}_1, -\tilde{\gamma})$$

so the uniqueness assumption requires a restriction of the parameter space to exclude such 'mirror images' of the parameters. There are more substantive examples when identifiability (and hence uniqueness) is not guaranteed in mixture structures (see Chapter 3 in Titterington et. al. (1985)). Discussions of the identifiability of mixtures are in Teicher (1963), Titterington et. al. (1985), and Jeffries (1998). We will not pursue these questions further in this work other than to reiterate our assumption that $\alpha_1^\circ, \alpha_0^\circ$ and γ° are unique.

In order to show α_1^*, α_0^* and γ^* maximize $\mathbb{E}[\log k(Y, Y_{-1}; \alpha_1, \alpha_0, \gamma)]$ we first show

$$\mathbb{E}[\log k(Y, Y_{-1}; \alpha_1, \alpha_0, \gamma) \mid Y_{-1}] \leq \mathbb{E}[\log k(Y, Y_{-1}; \alpha_1^*, \alpha_0^*, \gamma^*) \mid Y_{-1}] \tag{4.30}$$

for all Y_{-1} and all $\alpha_1, \alpha_0, \gamma$. To see this we note that $h(y, y_{-1}, v_{-1}; \alpha_1^*, \alpha_0^*, \tau^*)$ as defined in (4.9) is the true conditional distribution of Y given Y_{-1} and V_{-1}. From the independence of Y_{-1} and V_{-1} we may integrate out the V_{-1} and conclude the conditional distribution of Y given Y_{-1} is

$$\int h(y, y_{-1}, v_{-1}; \alpha_1^*, \alpha_0^*, \tau^*) \varphi(v_{-1}) dv_{-1} = \tag{4.31}$$

$$\int \left\{ I_{[v_{-1} < \tau^*/\phi^* - y_{-1}/\phi^*]} f(y \mid y_{-1}; \alpha_1^*) + \left(1 - I_{[v_{-1} < \tau^*/\phi^* - y_{-1}/\phi^*]}\right) f(y \mid y_{-1}; \alpha_0^*) \right\} \varphi(v_{-1}) dv_{-1} =$$

$$\Phi\left(\frac{\tau^* - y_{-1}}{\phi^*}\right) f(y \mid y_{-1}; \alpha_1^*) + \left(1 - \Phi\left(\frac{\tau^* - y_{-1}}{\phi^*}\right)\right) f(y \mid y_{-1}; \alpha_0^*) = \tag{4.32}$$

$$k(y, y_{-1}; \alpha_1^*, \alpha_0^*, \gamma^*) \tag{4.33}$$

where $\varphi(\cdot)$ denotes the pdf of a standard normal random variable. From an application of the Kullback-Leibler information inequality to this conditional distribution we obtain our desired result in (4.30), i.e.

$$\mathbb{E}[\log k(Y, Y_{-1}; \alpha_1, \alpha_0, \gamma) \mid Y_{-1}] = \int \log k(y, Y_{-1}; \alpha_1, \alpha_0, \gamma) k(y, Y_{-1}; \alpha_1^*, \alpha_0^*, \gamma^*) dy \tag{4.34}$$

$$\leq \int \log k(y, Y_{-1}; \alpha_1^*, \alpha_0^*, \gamma^*) k(y, Y_{-1}; \alpha_1^*, \alpha_0^*, \gamma^*) dy \tag{4.35}$$

$$= \mathbb{E}[\log k(Y, Y_{-1}; \alpha_1^*, \alpha_0^*, \gamma^*) \mid Y_{-1}]. \tag{4.36}$$

Integrating this result over Y_{-1} and applying our uniqueness/identifiability assumption allows us to conclude

$$\mathbb{E}[\log k(Y, Y_{-1}; \alpha_1, \alpha_0, \gamma)] < \mathbb{E}[\log k(Y, Y_{-1}; \alpha_1^*, \alpha_0^*, \gamma^*)] \text{ for all } \alpha_1, \alpha_0, \gamma \neq \alpha_1^*, \alpha_0^*, \gamma^* \tag{4.37}$$

which implies

$$\alpha_1^*, \alpha_0^*, \gamma^* = \arg \max_{\alpha_1, \alpha_0, \gamma} \mathbb{E}[\log k(Y, Y_{-1}; \alpha_1, \alpha_0, \gamma)] = \alpha_1^\circ, \alpha_0^\circ, \gamma^\circ, \tag{4.38}$$

thus completing the proof. □

In Table 1 and Table 2 we saw that the logistic mixture performed about as well as the probit mixture, though the proof of Theorem 2 shows that it is the probit mixture that correctly specifies the conditional distribution of Y_t given Y_{t-1} because the noise in the threshold is normally distributed. Consequently, the logistic mixture is misspecified. On the other hand we should not be very suprised that the logistic model performs about as well because the logistic and probit cdf's are both similar in shape.

Conclusion

The import of these theorems and simulations should be apparent. First, it is clear that threshold models can produce inconsistent results when the threshold mechanism is noisy. We believe this has been overlooked in the application of such models. This leads to the idea of using variable mixtures as diagnostic tools for assessing the goodness of fit of threshold models. If the two models yield quite different results then one may be in a situation where the threshold model fails to adequately describe the data generation. As an illustration we revisit the data presented in Table 2. Table 3 indicates that there is very good agreement between the estimates of the threshold and probit

Parameter Difference	Mean	Empirical SD of Difference
$\hat{\beta}_1^{TAR} - \hat{\beta}_1^{PM}$	.004	.015
$\hat{\sigma}_1^{2\ TAR} - \hat{\sigma}_1^{2\ PM}$	-.002	.006
$\hat{\beta}_0^{TAR} - \hat{\beta}_0^{PM}$	.001	.022
$\hat{\sigma}_0^{2\ TAR} - \hat{\sigma}_0^{2\ PM}$	.008	.013

Table 3: Differences of Estimates in Correctly Specified Model

models when the threshold model is correctly specified (without noise). When the threshold model is incorrectly specified (noise is present), one obtains much worse agreement as evidenced in Table 4. The magnitude of the difference is 12 to 71 times larger under the misspecified model.

Parameter Difference	Mean	Empirical SD of Difference
$\hat{\beta}_1^{TAR} - \hat{\beta}_1^{PM}$	-.053	.052
$\hat{\sigma}_1^{2\ TAR} - \hat{\sigma}_1^{2\ PM}$	-.043	.028
$\hat{\beta}_0^{TAR} - \hat{\beta}_0^{PM}$	.099	.077
$\hat{\sigma}_0^{2\ TAR} - \hat{\sigma}_0^{2\ PM}$	.068	.035

Table 4: Differences of Estimates in Misspecified Model

Alternatively, an analyst may wish to consider variable mixture models as the true data generation model. In this paper we have considered how well variable mixtures fit data generated by threshold methods. One could also consider how well or poorly threshold methods fit data generated by variable mixtures (we believe the same type of inconsistency would arise because of misclassification problems). In choosing variable mixture models the analyst presumes the question of which component distribution generates a given observation is influenced by the magnitude of the threshold variable, but not completely determined by it. If this seems a reasonable supposition then variable mixture models should be seriously considered as descriptions of the true process.

Finally, while we have presented the misspecification in the models as arising from an unobserved normally distributed error, we think it likely similar results would be obtained under a broad range of other types of misspecification (at least the inconsistency of the threshold models). For example, it appears that the logistic mixture models may also be consistent for data generated via a TAR model with normally distributed noise in the threshold variable. This is despite the fact that we know it is the probit mixture that correctly models the conditional distribution of Y_t given Y_{t-1}. It may be that both types of variable mixtures presented here are consistent for α_1^* and α_0^* if the noise in the threshold is symmetrically distributed. Future work should investigate how robust variable mixtures may be to different types of misspecification.

References

[1] CHAN, K.S. (1990) Appendix 1: Deterministic stability, stochastic stability, and ergodicity. In H. Tong, *Non-Linear Time Series: A Dynamical System Approach.* Oxford University Press, Oxford.

[2] CHAN, K.S. (1993) 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.

[3] COX, D.R. (1975) Partial likelihood. *Biometrika*, 64:269–276.

[4] DEMPSTER, D.A. and LAIRD, N.M. and RUBIN, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39:1–22.

[5] DIEBOLD, F.X. and LEE, J. and WEINBACH, G.C. (1994) Regime switching with time-varying transition probabilities. In C. Hargreaves, editor, *Nonstationary Time Series and Cointegration.* Oxford University Press, Oxford.

[6] EVERITT, B.S. and HAND, D.J. (1981) *Finite Mixture Distributions.* Chapman and Hall, London.

[7] FAHRMEIR, L. and TUTZ, G. (1994) *Multivariate Statistical Modeling Based on Generalized Linear Models.* Springer-Verlag, New York.

[8] FILARDO, A.J. (1994) Business cycle phases and their transitional dynamics. *Journal of Business and Economic Statistics*, 9:299–308.

[9] FOKIANOS, K. (1996) *Categorical Time Series: Prediction and Control.* PhD thesis, University of Maryland, College Park.

[10] JEFFRIES, N.O. (1998) *Logistic Mixtures of Generalized Linear Model Time Series.* PhD thesis, University of Maryland, College Park.

[11] KUK, A.Y.C. and CHEN, C.H. (1992) A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, 79(3):531–41.

[12] LARSON, M.G. and DINSE, G.E. (1985) A mixture model for the regression analysis of competing risks data. *Applied Statistics*, 201–211.

[13] MCLACHLAN, G.J. and BASFORD, K.E. (1988) *Mixture Models: Inference and Applications to Clustering.* Marcel Dekker, Inc., New York, NY.

[14] MCLACHLAN, G.J. and KRISHNAN, T. (1997) *The EM Algorithm and Extensions.* Wiley, New York.

[15] NUMMELIN, E. (1984) *General Irreducible Markov Chains and Non-negative Operators.* Cambridge University Press, Cambridge.

[16] SLUD, E. and KEDEM, B. (1994) Partial likelihood analysis of logistic regression and autoregression. *Statistica Sinica*, 4:89–106.

[17] TEICHER, H. (1963) Identifiability of finite mixtures. *Annals of Mathematical Statistics*, 1265–1269.

[18] TITTERINGTON, D.M. and SMITH, A.F.M. and MAKOV, U.E. (1985) *Statistical Analysis of Finite Mixture Distributions.* Wiley, New York.

[19] TONG, H. (1983) Threshold models in non-linear time series analysis. In *Lecture Notes in Statistics, 21.* Springer, Berlin.

[20] TONG, H. (1990) *Non-Linear Time Series: A Dynamical System Approach.* Oxford University Press, Oxford.

[21] TWEEDIE, T. (1975) Sufficient conditions for ergodicity and recurrence of Markov chains on a general state space. *Stochastic Processes and their Applications*, 3:385–403.

[22] WALD, A. (1949) Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, 20:595–601.

[23] WONG, W.H. (1986) Theory of partial likelihood. *The Annals of Statistics*, 14:86–123.

[24] WU, C.F. (1983) On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11:95–103.