A New Stochastic Regime Switching Model with Time-varying Regression Coefficients and Error Variances

A Dissertation presented

by

Xiaojin Dong

to

The Graduate School in Partial Fulfillment of the Requirements for the

Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Concentration - Statistics)

Stony Brook University

May 2016

Stony Brook University

The Graduate School

Xiaojin Dong

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Haipeng Xing – Dissertation Advisor
Associate Professor, Department of Applied Mathematics and Statistics

Peifen Kuan – Chairperson of Defense
Assistant Professor, Department of Applied Mathematics and Statistics

Wei Zhu
Professor & Deputy Chair, Department of Applied Mathematics and Statistics

Keli Xiao
Assistant Professor, College of Business, Stony Brook University

This dissertation is accepted by the Graduate School.

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

A New Stochastic Regime Switching Model with Time-varying Regression Coefficients and Error Variances

by

Xiaojin Dong

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Concentration - Statistics)

Stony Brook University

2016

Since the publication of Hamilton's (1989) seminal work on Markov switching model, a large number of applications have been found in economics and finance. The classical Markov switching models characterize the estimation of parameters in finite state, limited by the pre-specified number of regimes, thus it is restrictive in empirical studies. In this thesis, we develop a stochastic regime switching model, where the model parameters are both categorical and continuous. By assuming conjugate priors and defining stochastic regime switching variables, we derive recursive filtering and smoothing algorithms to estimate the regimes and develop closed-form recursive Bayes estimates of the regression parameters. Moreover, bounded complexity mixture (BCMIX), an approximation scheme, is derived to increase the computation efficiency substantially and yet this method is comparable to the Bayes estimates in statistical efficiency. Hyperparameters are estimated via expectation and maximization procedure and presented in closed form solutions. Intensive simulation studies show that lower order of bounded complexity mixture procedure is as efficient as Bayes estimates and that estimation performs well on moderate large transition probability scenarios. A comparative simulation study shows that classical Markov switching models have a tendency to overestimate the transition probabilities. We used our model to analyze several US economic data, such as unemployment rate, industrial production and manufacturing and trade inventory, to show our model is more suitable than classical regime switching models in analyzing business cycles of economic time series data.

To my parents, Jingzhi Zhao and Caiwen Dong

and

To Douglas Fisk

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my advisor, Prof. Haipeng Xing, for proposing this interesting and exciting topic and for his guidance and continuous support.

I would also want to express my immense gratitude to Prof. Wei Zhu and Prof. Estie Arkin for giving me the opportunity to teach courses as a mean to fund my doctoral studies.

I am also indebted to many of my fellow group members and friends, especially to Yuan Yao and Lu Zhao for their help and valuable suggestions.

Chapter 1

Introduction

Many economic and financial time series display prominent features such as heavy tails, skewness, excess kurtosis and multimodality. Outside of time series domains, observations that possess the aforementioned features are often generated from different sub-populations. Suppose that a random variable y given the k^{th} sub group follows a distribution $p(\cdot|\boldsymbol{\theta}_k)$ indexed by a parameter set $\boldsymbol{\theta}_k$ and if the probability of being in the k^{th} sub group is denoted by η_k, the marginal distribution of Y takes the form of the mixture density

$$p(y) = \eta_1 p(y|\boldsymbol{\theta}_1) + \cdots + \eta_K p(y|\boldsymbol{\theta}_K) \tag{1.1}$$

K is the number of sub groups. Equation (1.1) is a basic construction of a standard finite mixture model considered by Everitt and Hand (1981) and Titterington et al. (1985). The important assumptions for finite mixture models are that $y_1, \ldots, y_N$ are independent and the sub group indicator $(S_1, \ldots, S_K)$ follows a standard discrete distribution.

It is widely acknowledged that a normal economy often experiences disruptive events such as economic contraction, economic expansion, or governmental policy change. These disruption may generate different dynamics for a time series that is of interest to econometricians. Traditional linear time series models fail to capture the structural changes in the inherent data generating process; nonlinearity shown in mixture model (1.1) seems to explain such a process reasonably better. For example, Quandt (1958, 1972) introduced a regime switching regression allowing the observations to be generated by distinct regression equations with probability $\lambda_1, \lambda_2, \ldots$. This model assumes the regime indicators at time t are independent of what the system was in the past and also assumes that the observations are regressed on exogenous variables only. Unfortunately, these models are not able to make an inference on regime status for the observation at each time point. The estimation of regime status is often of great use and interest to econometricians.

To deal with broader time series underpinning structural change, one may relax two conditions in the standard mixture model. First, allow y_t to be dependent on past values, enabling the model to capture autocorrelation. Second, specify a Markov probability law about the regime indicator, making the inference of regime status feasible. Such a model, called Markov switching model, was introduced into econometrics by Goldfeld and Quandt (1973). As an extension of the work of Quandt (1972), their model allows explicit dependency between the regime indicator S_t to be a two-state Markov chain, whereas Quandt (1972) assumes that S_t is an i.i.d. random sequence.

Markov switching models have many applications in engineering time series under the name "hidden Markov model". It was an important tool in speech and pattern recognition developed in 1980's; see a good survey by Rabiner and Juang (1986) and Ghahramani (2001). The term "hidden Markov" originates from the fact that the observation y_t was generated by a process where the state S_t of y_t is unobserved and the fact that the unobserved state satisfies the Markov property. The hidden Markov model is also known by other names, such as Markov mixture model, which is preferred by biologists (P. Albert, 1991). Frühwirth-Schnatter (2006) gives a comprehensive overview about Markov mixture model. The term Markov switching model or regime-switching model is preferred by economists to analyze economic time series; see Neftçi (1984), J. Hamilton (1990) and Krolzig (1997). I will follow economics convention in this thesis.

1.1 The Basic Markov Switching Concept

Markov switching models form a very flexible class of nonlinear time series models and are able to capture many features of marginal distributions of practical time series such as asymmetry, nonnormaility with fat tails or multimodality to a greater extent than non-Markov mixture models. By introducing Markov properties into hidden state S_t in equation (1.1), Markov switching model allows more data dependence. In its simplest structure, although the observed process Y_t is uncorrelated when hidden state S_t is known, the autocorrelation in the marginal distribution of Y_t and also the squared process Y_t^2 exist. The mathematical details about the moments of Markov switching models are explored thoroughly by Timmermann (2000).

A time series $y_1, y_2, \ldots, y_T$ is a realization of a stochastic process $\{Y_t\}_{t\geq 1}$ and each Y_t is governed by an unobserved state S_t where S_t has K states and follows a Markov distribution with transition matrix P. An example distribution of S_t would be a K-state first order Markov chain whose transition matrix is $\{p_{ij}\},\ i, j = 1, 2, \ldots, K$ where $p_{ij} = P(S_t = j|S_{t-1} = i)$ with $\sum_{j=1}^{K} p_{ij} = 1$. Let $\boldsymbol{\theta_k}$ be the parameter associated with $S_t = k,\ k = 1, 2, \ldots, K$. Given S_t, conditional distribution of Y_t is generated from a specific distribution family:

$$Y_t|S_t = k \sim f_\theta(\cdot|\boldsymbol{\theta_k}) \tag{1.2}$$

There are many variations of this model with respect to the nature of stochastic process $\{Y_t\}_{t\geq 1}$ and properties of Markov chain S_t. As far as Markov chain S_t is concerned, commonly seen properties include order, reducibility, periodicity and homogeneity which directly affect the definition of the transition matrix and the initial distribution of the chain. For example, the transition probabilities of a r^{th} order K state Markov chain are denoted as

$$P(S_t = k_t|S_{t-1} = k_{t-1}, \cdots, S_{t-r} = k_{t-r}),$$

where $k_t = 1, 2, \cdots, K$ and $\sum_{i=1}^{K} P(S_t = i | S_{t-1} = k_{t-1}, \cdots, S_{t-r} = k_{t-r}) = 1$. Reducibility is related to whether S_t has the ability to leave the current state in the chain. A two-state and first order transition matrix such as

$$\begin{bmatrix} p_{11} & 1 - p_{11} \\ 0 & 1 \end{bmatrix}$$

is an example of a reducible Markov chain, i.e. once the chain reaches state 2, it remains there with no possibility of returning to state 1 again. A periodic Markov chain switches between states in a periodic manner given an initial state S_0. In a first order Markov chain with transition matrix

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

given initial state $S_0 = 1$, $S_t = 1$ at time $t = 3, 5, 7 \ldots$, with probability 1 and $S_t = 2$ at time $t = 2, 4, 6, \ldots$, with probability 1 and so on. Such a Markov chain has period 2 and does not converge.

The aforementioned transition probabilities depend only on the order of the states in the Markov chain which is named homogeneous (time invariant) Markov chain. The inhomogeneous (time-varying) Markov chain arises from the fact that conditional distribution of S_t depends not only on recent values of $S_{t-1}, S_{t-2}, \ldots$, but also on exogenous variables or history of Y_t. Let

$$\Omega_t = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_t, \boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_t\} \tag{1.3}$$

where $\boldsymbol{z}_i$'s are exogenous variables. The transition probability of an inhomogeneous Markov chain can be defined as

$$P(S_t = k|S_{t-1}, S_{t-2}, \ldots, \Omega_{t-1}, \boldsymbol{z}_t)$$

The Markov switching models with time-varying transition probabilities are discussed by Diebold et al. (1994), Filardo (1998) and Peria (2002).

In business cycle analysis, numerous research discovers that a business cycle is asymmetric and exhibits a pattern of recession, rapid growth after recession and normal growth and back to recession. See references in Acemoglu and Scott (1997); Kim and Nelson (1998); J. Hamilton and Raj (2002); Chauvet et al. (2002). This research has also indicated that the durations of the states are not the same. Therefore, certain Markov properties would not be appropriate for business cycle analysis. The recurrent nature of economic activity invalidates a reducible chain; asymmetry disqualifies a periodic chain. The widely used assumption for hidden Markov chain in economic time series analysis is defined in the following:

Assumption 1.1. *S_t is an irreducible, aperiodic, first order homogeneous Markov chain starting from its ergodic distribution $\boldsymbol{\pi} = (\pi_1, \pi_2, \cdots, \pi_K)$.*

$Y_1, Y_2, \ldots, Y_T$ can be discrete-valued or continuous-valued time series. Given the state S_t, Y_t may come from any parametric family. This thesis mainly focuses on reviewing and modelling the continuous-valued time series with the Markov chain defined in assumption 1.1, which appears in section

1.2. Readers interested in discrete-valued time series generated by a hidden Markov chain may refer to MacDonald and Zucchini (1997). Equation (1.2) indicates that Y_t's are independent if S_t is known. However, in practice, Y_t may be also correlated with its own lags, with exogenous variables, or even with history of S_t. The exact relation is determined by the model specification of Y_t discussed in section 1.2. The relation also determines the complexity of statistical inference discussed in section 1.3.

1.2 Major Markov Switching Models

It is true that the autocorrelation of marginal distribution of Y_t can be captured under basic Markov switching model (1.2) via only specifying a Markov chain of the hidden state. The conditional distribution of Y_t is not necessarily always uncorrelated. To relax the assumption in equation (1.2), Let

$$P(y_t|S_t, S_{t-1}, \ldots, \Omega_{t-1}, \boldsymbol{z}_t,\ \boldsymbol{\Theta}) \tag{1.4}$$

be the conditional density of Y_t depending on endogenous or exogenous variables or more lagged hidden states, where $\boldsymbol{\Theta}$ includes all the model parameters and Ω_t is defined in (1.3).

It is widely acknowledged that J. Hamilton (1989)'s seminal paper popularizes the application of Markov switching model in economic time series analysis. He introduced a two-state Markov chain in the mean level in an autoregressive model in the analysis of log difference of GNP data:

$$Y_t - \mu_{S_t} = \phi_1(Y_{t-1} - \mu_{S_{t-1}}) + \cdots + \phi_p(Y_{t-p} - \mu_{S_{t-p}}) + \epsilon_t, \tag{1.5}$$

where $\epsilon_t \sim N(0, \sigma^2)$. Later this model was extended to allow changing states of autoregressive coefficients and error variance in the form of

$$Y_t = \phi_{S_t,0} + \phi_{S_t,1} Y_{t-1} + \cdots + \phi_{S_t,p} Y_{t-p} + \epsilon_t, \tag{1.6}$$

where $\epsilon_t \sim N(0, \sigma^2_{S_t})$, by Holst et al. (1994) and McCulloch and Tsay (1994). When S_t takes K states, equation (1.6) is usually denoted as MS(K)-AR(p) which describes a Markov switching autoregressive model with K states and autoregressive order p.

Equation (1.5) and (1.6) relies on only endogenous variables, i.e. the history of Y_t. Other regression models may only depend on exogenous variables $\boldsymbol{z}_t$ such as

$$Y_t = \boldsymbol{z}_t' \boldsymbol{\beta}_{S_t} + \epsilon_t, \tag{1.7}$$

where variance of ϵ_t can be homoscedastic as σ^2 or heteroscedastic as $\sigma^2_{S_t}$ and coefficient $\boldsymbol{\beta}_t$'s are state dependent. The most general form of Markov regression model may allow both endogenous and exogenous regressors and the coefficients of the regressors can be partly state dependent and partly state independent and the variance of error term can be either homoscedastic or heteroscedastic. Some authors name this general form as a Markov switching dynamic regression model (Frühwirth-Schnatter, 2006). However, I do not distinguish the terms in the following discussion.

Markov switching concept has also been utilized in autoregressive conditional heteroscedastic (ARCH) model introduced by Engle (1982) and in the generalized autoregressive conditionally heteroscedastic (GARCH) model by Bollerslev (1986) to capture the volatility clustering in financial time series analysis. Numerous researchers have proposed Markov switching ARCH models. The basic idea is

$$\begin{aligned} Y_t &= \sqrt{\gamma_{S_t}} h_t \epsilon_t, \\ h_t^2 &= 1 + \frac{\alpha_1}{\gamma_{S_{t-1}}} Y_{t-1}^2 + \cdots + \frac{\alpha_m}{\gamma_{S_{t-m}}} Y_{t-m}^2, \end{aligned} \tag{1.8}$$

formulated by J. Hamilton and Susmel (1994). Other formulations and variation in model details can be seen in Cai (1994), Gray (1996), Wong and Li (2001) and Kaufmann and Frühwirth-Schnatter (2002). Francq et al. (2001) consider a GARCH(m, n) model where all coefficients are allowed to switch:

$$\begin{aligned} Y_t &= \sigma_t \epsilon_t, \quad \epsilon_t \sim N(0,1) \\ \sigma_t^2 &= \gamma_{S_t} + \alpha_{S_t,1} y_{t-1}^2 + \cdots + \alpha_{S_t,m} y_{t-m}^2 + \delta_{S_t,1} \sigma_{t-1}^2 + \cdots + \delta_{S_t,n} \sigma_{t-n}^2 \end{aligned} \tag{1.9}$$

In such a model, conditional density of Y_t depends on the whole history of S_t. See application of Markov switching GARCH in stock market returns in Dueker (1997) and exchange rate time series data in Klaasen (2002).

There is always challenge in statistical inference when Y_t relies on the whole history of hidden state. State-space representation is flexible and allows reconstructing a complex Markov switching model into two equations: transition and observation equations. The fundamental purpose of writing in such a form is to make Y_t conditionally independent given unobserved state variable S_t, i.e.

$$p(y_1, \ldots, y_T | s_1, \ldots, s_T) = \prod_{t=1}^{T} p(y_t | s_t)$$

and to make state variable a first-order Markov chain, i.e.

$$p(s_1, \ldots, s_T) = \prod_{t=2}^{T} p(s_t | s_{t-1}) p(s_1)$$

Kim (1994) developed simple filter and smoother to make the estimation of Markov switching state-space model easier and widely applicable. See also Kim and Nelson (1999).

1.3 Statistical Inference of Markov Switching Models

Apart from the model specification, there are additional three important issues in the statistical inference of a Markov switching model. First, modeling requires the knowledge of the number of state K in the hidden chain. Second, state-specific parameters $\boldsymbol{\theta_k},\ k = 1, \cdots, K$ and probabilities in transition matrix are unknown and need to be estimated from the data. Finally, the estimation of the probability of the state at a time point, $P(S_t = k)$ for $t = 1, \cdots, T$ tells which hidden state Y_t belongs to.

Although the Markov structure and state dependent parameters in the regression model bring more flexibility in feature of Y_t, the challenge is that the more complex the model structure, the more difficult the statistical inference. The most commonly used methods to estimate parameters are maximum likelihood estimation and Bayesian estimation, both of which rely on conditional likelihood. For example

$$f(Y_t|\Omega_{t-1}, \boldsymbol{z}_t; \boldsymbol{\Theta}) = \sum_{i=1}^{K} f(Y_t|S_t = i, \Omega_{t-1}, \boldsymbol{z}_t; \boldsymbol{\Theta}) P(S_t = i|\Omega_{t-1}, \boldsymbol{z}_t; \boldsymbol{\Theta}) \quad (1.10)$$

where Ω_t is defined in (1.3) and $\boldsymbol{\Theta}$ includes all the model parameters. Equation (1.10) is also called one-step ahead forecast of y_t which is a mixture of a density family weighed by one-step forecast of probability of the hidden states. The second piece of right hand side of (1.10) is an intermediate step of filtering estimation of probability of S_t, the answer to the third issue discussed in detail below.

Hamilton did a great contribution to the statistical inference of Markov switching regression model. In his 1989 paper, he developed an iterative method to estimate probabilities of hidden states at time t given all the information available at time t. The byproduct of this procedure is a likelihood function which can be used to estimate other model parameters. Here I present a short summary of his algorithm along with the reference to Kim and Nelson (1999).

Let $\boldsymbol{y}_t$ be a univariate or a vector-valued observed variable and $\boldsymbol{z}_t$ be an exogenous variable, either univariate or multidimensional for $t = 1, 2, \ldots, T$. $\boldsymbol{y}_t$ is governed by a finite discrete hidden state S_t as in assumption 1.1 with K states. S_t is uncorrelated with $\boldsymbol{z}_t$, notationally,

$$P(S_t = j|S_{t-1} = i, S_{t-2} = i_{t-1}, \cdots, \boldsymbol{z}_t, \Omega_{t-1}) = P(S_t = j|S_{t-1} = i) = p_{ij} \tag{1.11}$$

where Ω_t is defined in (1.3) and $\sum_{j=1}^{K} p_{ij} = 1$ for $i, j = 1, 2, \ldots, K$. Conditional distribution of $\boldsymbol{y}_t$ depends on the specification of regression model and error distribution and has the general form of

$$f(\boldsymbol{y}_t|S_t, S_{t-1}, \cdots, \boldsymbol{z}_t, \Omega_{t-1}; \boldsymbol{\Theta}) \tag{1.12}$$

if the regression model parameters are collected into $\boldsymbol{\alpha}$, and transition probabilities p_{ij}'s into $\boldsymbol{\lambda}$ and $\boldsymbol{\Theta} = (\boldsymbol{\alpha}, \boldsymbol{\lambda})$. For example, let y_t be univariate and

$$y_t = \beta_{S_t,0} + \beta_{S_t,1} y_{t-1} + \beta_{S_t,2} z_t + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2_{S_t})$$

Then (1.12) is reduced to

$$f(y_t|S_t = j, z_t, \Omega_{t-1}; \boldsymbol{\Theta}) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left\{ - \frac{(y_t - \beta_{j,0} - \beta_{j,1} y_{t-1} - \beta_{j,2} z_t)^2}{2\sigma_j^2} \right\}$$

for $j = 1, 2, \ldots, K$.

The ultimate goal of Hamilton's iterative algorithm is to estimate $P(S_t|\Omega_t;\mathbf{\Theta})$. For simplicity and the ease of comprehension, I present this algorithm in terms of the regression model where y_t is uncorrelated with lags of S_t except S_t and S_{t-1} in the following five steps.

STEP 1: Compute the joint distribution of $(S_t,\ S_{t-1})$ given all the information available at $t-1$ and also by (1.11):

$$\begin{aligned} P(S_t, S_{t-1}|\Omega_{t-1};\mathbf{\Theta}) &= P(S_t|S_{t-1},\Omega_{t-1};\mathbf{\Theta})P(S_{t-1}|\Omega_{t-1};\mathbf{\Theta}) \\ &= P(S_t|S_{t-1})P(S_{t-1}|\Omega_{t-1};\mathbf{\Theta}) \end{aligned} \tag{1.13}$$

STEP 2: Compute the joint distribution of y_t, S_t, S_{t-1} given all the information available at time $t-1$.

$$f(y_t, S_t, S_{t-1}|\Omega_{t-1};\mathbf{\Theta}) = f(y_t|S_t, S_{t-1},\Omega_{t-1};\mathbf{\Theta})P(S_t, S_{t-1}|\Omega_{t-1};\mathbf{\Theta}) \tag{1.14}$$

The last two terms come from (1.12) and step 1 respectively.

STEP 3: We then have

$$f(y_t|\Omega_{t-1};\mathbf{\Theta}) = \sum_{i=1}^{K}\sum_{j=1}^{K} f(y_t, S_t = j, S_{t-1} = i|\Omega_{t-1};\mathbf{\Theta}) \tag{1.15}$$

the conditional distribution of y_t given all information available at $t-1$ and all the model parameters.

STEP 4: By Bayes' theorem and with the information in step 2 and 3, we can update joint distribution of S_t, S_{t-1} by adding y_t to the available information.

$$\begin{aligned} P(S_t, S_{t-1}|\Omega_t; \boldsymbol{\Theta}) &= \frac{f(y_t, S_t, S_{t-1}|\Omega_{t-1}; \boldsymbol{\Theta})}{f(y_t|\Omega_{t-1}; \boldsymbol{\Theta})} \\ &= \frac{f(y_t|S_t, S_{t-1}, \Omega_{t-1}; \boldsymbol{\Theta}) P(S_t, S_{t-1}|\Omega_{t-1}; \boldsymbol{\Theta})}{f(y_t|\Omega_{t-1}; \boldsymbol{\Theta})} \end{aligned} \tag{1.16}$$

STEP 5: Marginalize S_t in previous step.

$$P(S_t|\Omega_t; \boldsymbol{\Theta}) = \sum_{i=1}^{K} P(S_t, S_{t-1} = i|\Omega_t; \boldsymbol{\Theta}) \tag{1.17}$$

Equation (1.17) is called filtered probability of S_t. Filtered estimation refers to the estimation of S_t conditional on the information up to time t. An alternative important way to estimate probability of S_t is by smoothing which refers to the estimation of S_t conditional on all the information in the sample i.e.

$$P(S_t|\Omega_T, \boldsymbol{\Theta}) \text{ for } t \leq T.$$

To obtain this smoothed probability distribution, one needs to know the filtered probability as in (1.17) and the smoothed probability distribution $P(S_{t+1}|\Omega_T, \boldsymbol{\Theta})$. A comprehensive derivation of full sample smoother is given in J. Hamilton 1989's paper, as well as in other good references such as J. D. Hamilton (1994), Kim (1994), Kim and Nelson (1999) and Scott (2002).

To start the algorithm, we need to provide an initial value $P(S_0|\Omega_0, \boldsymbol{\Theta})$

which is the ergodic distribution of S_t if the Markov chain follows assumption 1.1. For a first order irreducible and aperiodic Markov chain, the ergodic distribution is a function of transition probabilities in (1.11). The general form of ergodic distribution is derived in detail in J. D. Hamilton (1994, p. 684). In case of permanent regime change (e.g. reducible chain) , one may consider an independent initial distribution such as discrete uniform distribution, i.e. $P(S_t = k) = \frac{1}{K}$, for all $k = 1, 2, \ldots, K$.

The byproduct of this iterative algorithm is the convenience of writing the likelihood function by (1.15):

$$L(\boldsymbol{\Theta}; \Omega_t) = \prod_{t=1}^{T} f(y_t|\Omega_{t-1}; \boldsymbol{\Theta})$$

whose log likelihood function is

$$l(\boldsymbol{\Theta}; \Omega_t) = \sum_{t=1}^{T} \ln\big(f(y_t|\Omega_{t-1}; \boldsymbol{\Theta})\big). \tag{1.18}$$

The estimation of model parameters can be made based on (1.18). Apparently, calculating likelihood is inevitable in this iterative algorithm. A succinct matrix representation of this algorithm is given in J. D. Hamilton (1994, p. 692 – 696). The algorithm documented above is convenient for computational purpose.

Typically, we may apply Maximum likelihood estimation (MLE) method to make an inference about model parameters based on (1.17) and (1.18). It has been proven that MLE is consistent, asymptotically normal with variance equal to the inverse of Fisher information matrix: See Casella and Berger (2001). An excellent review of the asymptotic properties of ML estimator for hidden Markov models can be seen in Cappé et al. (2005, Chapter, 12). However, the likelihood function of Markov switching models is often featured with multiple local optima, essential singularities and even occasionally unbounded functional values, which causes trouble in numerical optimization. Popular numerical methods such as steepest ascent, Newton-Raphson or Davidon-Fletcher-Powell may suffer from numerical instability and fail to produce valid sample Hessian in the case of non-concave objective functions. Therefore under MLE method, research in practice has often been limited to simpler models, lower dimension and a small number of regimes.

An alternative method to estimate model parameters is the Expectation-Maximization (EM) algorithm, originally motivated by Dempster et al. (1977) and extended and specialized by J. Hamilton (1990) for Markov switching model. EM algorithm has the principal advantage over MLE methods for its numerical robustness despite the ill chosen starting values. Good initial values are preferred and experimenting multiple initial values is recommended among practitioners. J. Hamilton (1990, section 3) proved that using complete data density in expectation step and maximizing such an expectation yield exactly the maximum likelihood estimates. Further benefit of EM algorithm includes a potential application to a large vector system. Ideally, after EM algorithm, the solution of parameter estimation has an analytical form, whereas in many case, maximization of the function in E-step ends up with the numerical optimization where the same trouble as in MLE is likely to occur.

In no conflict with the nondifferentiability of the likelihood function, Bayesian methods have become attractive to estimate Markov switching models since they can allow flexible relationship between parameters in various hidden states, are easier to implement and computationally feasible. Besides, exploration of posterior distributions of the parameters can provide more information than first and second moments that MLE can only provide. Conjugate priors exist for very few model specifications and most analysis of posterior distributions relies on Markov chain Monte Carlo (MCMC) method. Gibbs sampling by J. H. Albert and Chib (1993) is rather straightforward to use in data augmentation and conditioning in a Markov switching autoregressive model. Chib (1996) extended their method to general Markov switching models. See other references in McCulloch and Tsay (1994), Kaufmann (2000) and Chauvet et al. (2002). Care must be taken to choose prior distributions to avoid improper posterior. Inappropriate prior may also cause biased posterior estimates. MCMC can be very expensive for large dimensional complex models.

So far, Hamilton's iterative algorithm turned out to be a prevalent solution to the third issue in statistical inference of Markov switching models. The advent of various techniques in MLE, EM algorithm and Bayesian estimation in current literature provides possible solutions to the second issue.

Inevitably no method is universally superior to the others. To choose a good estimation method is always a daunting task because the practitioners need to consider the aspects of model complexity, data size, and the available computational resource as a whole. In a large vector system, the estimation of parameters may include a combination of different methods at different stage of the estimation. In this thesis, the proposed model uses Bayesian method to estimate the model parameters and EM algorithm to estimate the prior parameters as in section 2.6.

The first and foremost inferential issue of Markov switching model is the estimation of number of regimes that has to be addressed the last simply because the current research of testing the number of regimes relies on the conditional log-likelihood which can be only computed after the model parameters are estimated. J. Hamilton (1996) pointed out the challenge of testing on Markov switching models. He proposed tests on other aspects of model misspecification but did not test on number of regimes. Testing on number of regimes involves inference for an overfitting mixture model which presents a non-regular condition with the true parameter lying in a nonidentifiable subset of the larger parameter space. Thus the condition for the standard (LR) test statistic, Wald test statistic or the score test statistic fails to hold. Based on Hamilton's 1989 model, Hansen (1992) approximated the LR statistic under non-regular condition using empirical process theory; Garcia (1998) derived analytically the asymptotic null distribution of the LR test by treating transition probabilities as nuisance parameters; Cho and

White (2007) derived limiting distribution for quasi LR statistics for various special cases of mixture models. Besides, AIC, BIC and marginal likelihood have been used to deal with model selection problems (not limited to estimation of number of regimes) for Markov switching models; See Wang and Puterman (1999) and Frühwirth-Schnatter (2004). There is more room to be explored in these areas in future research.

1.4 Contributions of this thesis

In this dissertation, I continue to explore the Markov switching regression model where the regressors can be endogenous and exogenous. The model parameters (including the variance) are still determined by the outcomes of a discrete-state Markov chain. The first contribution of this model is that the regression parameters (including error variance) are no longer piecewise constants for each regime, which has not been discussed in the current literature to my best knowledge. This is accomplished by bringing in an additional assumption that the joint distribution of the model parameters relies on an indicator variable of whether the regime changes from a previous time point. This new assumption makes the model parameters to be related to regimes but not entirely restricted by the regime, in other words, the space of model parameter is continuous and infinite even under the same regime. For a particular regime, the number of sets of parameters to be estimated is related to the number of transitions the system has made from other regimes to the specified one. This contribution is non-trivial, because this model gives more flexibility to handle the difference within a regime. It has the great advantage in economic application simply because, for example, the levels and variations of economic down (up) turn may not necessarily be the same in different time periods. Thus a small number of regimes in this model can explain a complex time series data that would have to be otherwise analyzed by a large number of regimes under the traditional Markov switching models.

The second innovation of this model is the specification of a regime relevant stochastic time variable. This new variable describes the most recent change-of-regime time point before or after a particular time point at which the hidden regime assumes a particular state. Lai et al. (2005) and Lai and Xing (2011) delved the similar concept in non-Markovian regime switching models and derived the recursive formula for its probability distribution. Invoked by their ideas, we derived the recursive formula for the distribution of this new variable in Markovian environment. The introduction of this time variable and its probability distribution bring further simplification in parameter estimation. With the conjugate prior assumption, Bayesian estimates of model parameters and the error variance have closed form solutions; prior estimates via EM algorithm all have explicit solutions; the estimation of the regime at each time point is the byproduct of the recursive formula, so is the marginal likelihood function.

1.5 Outline

This thesis is planned as follows. Chapter 2 proposes a stochastic regime switching model where both the regression coefficients and the error variance depend on a Markov chain. Inference on the regime status and model parameters is derived based on the Bayesian framework and EM algorithm is also discussed to estimate hyperparameters in this chapter. Chapter 3 explores the simulation studies to compare and contrast the accuracy and the efficiency of Bayes and BCMIX estimation method for various series length and various transition probabilities. BCMIX estimation is compared with the estimation from the classical Markov switching model. This chapter also analyzes the effectiveness of EM algorithm on hyperparameters and the impact of hyperparameter estimation on the model parameters. Real data analysis of the proposed model is given in chapter 4 for 3 economic time series data: unemployment rate series, industrial production series and real manufacturing and trade inventory series. A comparison analysis with classical Markov switching regression model is also given in the end of this chapter. Concluding remarks are given in chapter 5. Lengthy proofs and formula derivations are shown in Appendix A.

Chapter 2

A New Regime Switching Regression Model and Its Inference

2.1 A New Regime Switching Regression Model

Assume that $\{y_t\}$ is a univariate stochastic process that follows

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta}_t + \sigma_t\epsilon_t, \quad \epsilon_t \sim N(0,\ 1), \tag{2.1}$$

where $\boldsymbol{x}_t$ is a $(d \times 1)$ vector that may include both endogenous (history of y_t) and exogenous variables and accordingly $\boldsymbol{\beta}_t$ is a $(d \times 1)$ regression parameter vector. Parameters $\boldsymbol{\beta}_t$ and σ_t depend on the hidden state S_t that satisfies the following assumptions:

(A1) $\{S_t = 1, \ldots, K | t \geq 0\}$ follows a first order, irreducible and reversible Markov chain with the transition matrix

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1K} \\ p_{21} & p_{22} & \cdots & p_{2K} \\ \vdots & \vdots & \vdots & \vdots \\ p_{K1} & p_{K2} & \cdots & p_{KK} \end{bmatrix} \tag{2.2}$$

(A2) Define $S_1 \neq S_0$. Let $\tau_t = (2\sigma_t^2)^{-1}$ and define $\boldsymbol{\theta}_t := (\boldsymbol{\beta}_t,\ \tau_t)$, then

$$\boldsymbol{\theta}_t = \mathbf{1}_{\{S_t = S_{t-1}\}}\boldsymbol{\theta}_{t-1} + \mathbf{1}_{\{S_t \neq S_{t-1}\}}(\boldsymbol{Z}_t, \gamma_t),$$

where $\boldsymbol{Z}_t | \gamma_t, S_t = k \sim N\Big(\boldsymbol{z}^{(k)}, \frac{\boldsymbol{V}^{(k)}}{2\gamma_t}\Big)$ whose distribution is denoted as $f_{0,0}^{(k)}$ and $\gamma_t | S_t = k \sim \text{Gamma}(g^{(k)},\ \lambda^{(k)})$ denoted as $h_{0,0}^{(k)}$ and $\boldsymbol{z}^{(k)}, \boldsymbol{V}^{(k)}, g^{(k)}$ and $\lambda^{(k)}$ are hyperparameters $\forall\ k = 1, \ldots, K$.

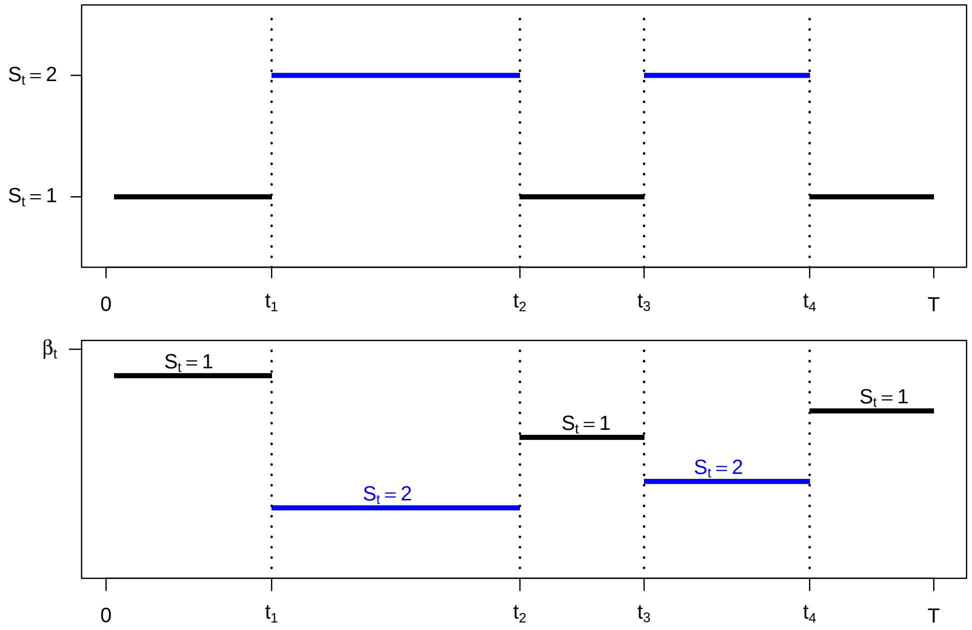
Assumption (A1) indicates that this Markov chain has the stationary probability distribution $\boldsymbol{\pi'} = (\pi_1, \ldots, \pi_K)$ and $\boldsymbol{\pi}$ has the following relation with the transition matrix P.

$$\boldsymbol{\pi}' P = \boldsymbol{\pi}' \tag{2.3}$$

As is well known in a classical Markov switching model, the distribution of model parameters depends on which hidden state the system is on at a particular time point and so model parameters are piecewise constants and each of them is limited to the number of regimes specified in the system. Assumption (A2) brings new features into model (2.1) and sheds light on some Markov property for model parameters. In (A2), the distribution of $\boldsymbol{\theta}_t$ relies not only on the regime state at a particular time point, but also on the state of the previous time point. For example, the system may be on state k at time t_1 and may have been through regime changes for some time, but at a future time t_2 the system may switch back to state k. At time t_2, the distribution of model parameters is from the same family distribution as in time t_1, however, the values of these parameters may not necessarily be the same, since (A2) allows the regeneration of the parameters at every regime switching point. Although the number of regimes is limited to K, the true values of each parameter are not limited. They are related to the number of switching to a particular state, which is uncertain in the whole stochastic process. In short, parameters are not necessarily piecewise constants at each regime, but piecewise constant between the adjacent regime changes.

To make the above concept more concrete, Figure 2.1 presents the state change and the corresponding parameter values in a simple two-state system. The process has been in state 1 until t_1 and been through the change to state 2 for some time and come back to state 1 again at t_2. However, the value of β_{S_t} until t_1 is not the same as that between t_2 and t_3, though both β_t's are in state 1, due to the regeneration mechanism in assumption (A2).

Figure 2.1: Illustration of the regime changes (upper panel) and the values of single parameter β_{S_t} at each regime (lower panel) in a two-state system.



Besides, Bayesian estimation is used to estimate $\boldsymbol{\beta}_t$ and σ_t based on the prior assumptions in (A2). Since the number of pieces of the parameters to be estimated is uncertain in this model, ML estimation fails to carry the task with an unknown number of parameters. Bayesian method makes it viable along with the definition of a new random time variable in equation (2.5), motivated by the idea of Lai et al. (2005, 2008); Lai and Xing (2011), with which the filter and smoother of the model parameters can be derived in the following sections.

2.2 Forward Filtering Estimation of Parameters

The goal of this section is to estimate the model parameters $\boldsymbol{\beta}_t$ and σ_t and probabilities of the regimes at each time t given all the historical information at time t. Such a time related inference is usually called the forward filtering estimation. First we need to define notations and useful variables. Let

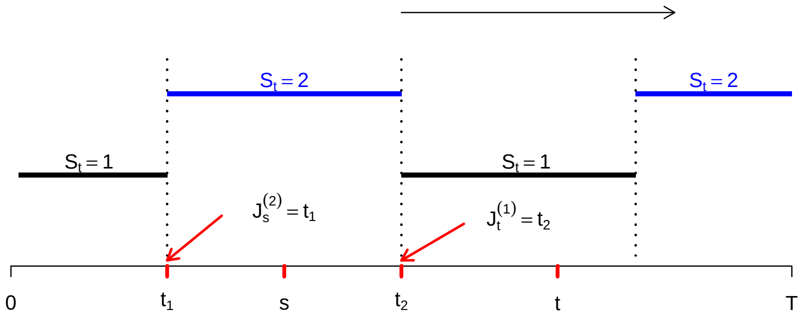
$$\begin{aligned} &\boldsymbol{y}_{ij} = (y_i, \ldots, y_j),\ \boldsymbol{x}_{ij} = (\boldsymbol{x}_i, \ldots, \boldsymbol{x}_j), \\ &\mathcal{F}_t = (\boldsymbol{x}_{1t}, \boldsymbol{y}_{1t}),\ \mathcal{F}_{ij} = (\boldsymbol{x}_{ij}, \boldsymbol{y}_{ij}) \end{aligned} \tag{2.4}$$

A new random variable describing the time index of the most recent regime change is defined as $J_t^{(k)}$, notationally,

$$J_t^{(k)} := \max\{i \leq t : S_{i-1} \neq S_i = \cdots = S_t = k\} \tag{2.5}$$

To be more specific, the system is in state k at time t, $J_t^{(k)}$ is the time index that the system moves onto state k from other states most recently at or before time t. Figure 2.2 shows that at time s, the system is on state 2, the most recent index to move to state 2 is $J_s^{(2)} = t_1$ and similarly, $J_t^{(1)} = t_2$ for system on state 1 at time t. To estimate a system where the change points are unknown, there is no way to be certain of $J_t^{(k)}$. So the range of $J_t^{(k)}$ is $1, \ldots, t$. The probability of $J_t^{(k)} = i$ given all information available at time t

Figure 2.2: Illustration of $J_t^{(k)}$ for a concrete two-state system



is defined as

$$\xi_{i,t}^{(k)} := P(J_t^{(k)} = i | \mathcal{F}_t), \tag{2.6}$$

Equation (2.6) is a short notation of $P(J_t^{(S_t)} = i, S_t = k|\mathcal{F}_t)$, a joint distribution of the state and the most recent time to change. In addition, the probability of the regime at each time given all information available at time t, i.e. $P(S_t = k|\mathcal{F}_t)$, is actually $\sum_{i=1}^{t} P(J_t^{(S_t)} = i, S_t = k|\mathcal{F}_t)$. Thus there is a simple relation between the filtering estimation of the regime and $\xi_{i,t}^{(k)}$ defined below.

$$\xi_t^{(k)} := P(S_t = k|\mathcal{F}_t) = \sum_{i=1}^{t} \xi_{i,t}^{(k)} \tag{2.7}$$

where $1 \leq i \leq t$ and $1 \leq k \leq K$.

Next, we estimate $(\boldsymbol{\beta}_t, \sigma_t)$ via the posterior distribution of $(\boldsymbol{\beta}_t, \tau_t)$, given all the information available at t since $\tau_t = (2\sigma_t^2)^{-1}$ defined in (A2) is easier to work with. By definitions in (2.5), (2.6), the posterior distribution of $(\boldsymbol{\beta}_t, \tau_t)$ is

$$\begin{aligned} f(\boldsymbol{\beta}_t, \tau_t|\mathcal{F}_t) &= \sum_{k=1}^{K}\sum_{i=1}^{t} f(\boldsymbol{\beta}_t, \tau_t, J_t^{(k)} = i|\mathcal{F}_t) \\ &= \sum_{k=1}^{K}\sum_{i=1}^{t} f(\boldsymbol{\beta}_t|\tau_t, J_t^{(k)} = i, \mathcal{F}_t) f(\tau_t|J_t^{(k)} = i, \mathcal{F}_t)\xi_{i,t}^{(k)} \qquad (2.8) \end{aligned}$$

which is a weighted average of conditional posterior distributions of $\boldsymbol{\beta}_t$ and τ_t given $J_t^{(k)} = i$.

By Bayes' theorem, the first term in the right hand side of equation (2.8) is

$$\begin{aligned} f(\boldsymbol{\beta}_t|\tau_t, J_t^{(k)} = i, \mathcal{F}_t) &= \frac{f(\boldsymbol{y}_{1t}|\boldsymbol{\beta}_t, \tau_t, J_t^{(k)} = i, \boldsymbol{x}_{1t}) f(\boldsymbol{\beta}_t|\tau_t, J_t^{(k)} = i, \boldsymbol{x}_{1t})}{f(\boldsymbol{y}_{1t}|\tau_t, J_t^{(k)} = i, \boldsymbol{x}_{1t})} \\ &\propto f(\boldsymbol{y}_{it}|\boldsymbol{\beta}_t, \tau_t, J_t^{(k)} = i, \boldsymbol{x}_{1t}) f(\boldsymbol{\beta}_t|\tau_t, J_t^{(k)} = i, \boldsymbol{x}_{1t}) \end{aligned}$$

Since the information before time i is irrelevant to the most recent changes to state k, the estimation of $\boldsymbol{\beta}_t$ uses only the most recent $t - i + 1$ pieces of information. Also because conditional y_r's are independent under the last regime change, the above equation can be further simplified to

$$\begin{aligned} &f(\boldsymbol{\beta}_t|\tau_t, J_t^{(k)} = i, \mathcal{F}_t) \\ &\propto \prod_{r=i}^{t} \exp\Big\{ - \tau_t(y_r - \boldsymbol{x}_r'\boldsymbol{\beta}_t)^2\Big\} \exp\Big\{ - \tau_t\big(\boldsymbol{\beta}_t - \boldsymbol{z}^{(k)}\big)'\big(\boldsymbol{V}^{(k)}\big)^{-1}\big(\boldsymbol{\beta}_t - \boldsymbol{z}^{(k)}\big)\Big\} \end{aligned}$$

$$\propto \exp\Big\{ -\tau_t\boldsymbol{\beta}_t'\Big(\sum_{r=i}^{t}\boldsymbol{x}_r\boldsymbol{x}_r' + (\boldsymbol{V}^{(k)})^{-1}\Big)\boldsymbol{\beta}_t + 2\tau_t\boldsymbol{\beta}_t'\Big(\sum_{r=i}^{t}\boldsymbol{x}_r y_r + \big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}^{(k)}\Big)\Big\}$$

$$\propto \exp\Big\{\tau_t\big(\boldsymbol{\beta}_t - \boldsymbol{z}_{i,t}^{(k)}\big)'\big(\boldsymbol{V}_{i,t}^{(k)}\big)^{-1}\big(\boldsymbol{\beta}_t - \boldsymbol{z}_{i,t}^{(k)}\big)\Big\}$$

where

$$\big(\boldsymbol{V}_{i,j}^{(k)}\big)^{-1} = \sum_{r=i}^{j}\boldsymbol{x}_r\boldsymbol{x}_r' + (\boldsymbol{V}^{(k)})^{-1}, \tag{2.9}$$

and

$$\boldsymbol{z}_{i,j}^{(k)} = \boldsymbol{V}_{i,j}^{(k)}\Bigg(\sum_{r=i}^{j}\boldsymbol{x}_r y_r + \big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}^{(k)}\Bigg) \tag{2.10}$$

In short, the conditional distribution of $\boldsymbol{\beta}_t$ given τ_t, $J_t^{(k)} = i$ and $\mathcal{F}_t$ follows a normal distribution with mean $\boldsymbol{z}_{i,t}^{(k)}$ and variance $\frac{\boldsymbol{V}_{i,t}^{(k)}}{2\tau_t}$, notationally,

$$\boldsymbol{\beta}_t|\tau_t, J_t^{(S_t)} = i, S_t = k, \mathcal{F}_t \sim \mathrm{N}\Big(\boldsymbol{z}_{i,t}^{(k)},\ \frac{\boldsymbol{V}_{i,t}^{(k)}}{2\tau_t}\Big) \text{ with p.d.f. } f_{i,t}^{(k)} \tag{2.11}$$

By property of marginal distribution and again with Bayes' theorem, the second item in the right hand side of equation (2.8) is shown to be

$$f(\tau_t|J_t^{(k)} = i, \mathcal{F}_t)$$

$$= \int \frac{f(\tau_t, \boldsymbol{\beta}_t, \boldsymbol{y}_{1t}|J_t^{(S_t)} = i, S_t = k, \boldsymbol{x}_{1t})}{f(\boldsymbol{y}_{1t}|J_t^{(S_t)} = i, S_t = k, \boldsymbol{x}_{1t})}\,\mathrm{d}\boldsymbol{\beta}_t$$

$$\propto \int f(\boldsymbol{y}_{it}|\tau_t, \boldsymbol{\beta}_t, J_t^{(k)} = i, \boldsymbol{x}_{1t}) f(\boldsymbol{\beta}_t|\tau_t, J_t^{(k)} = i, \boldsymbol{x}_{1t})\mathrm{d}\boldsymbol{\beta}_t f(\tau_t|J_t^{(k)} = i, \boldsymbol{x}_{1t})$$

$$\propto \tau_t^{g^{(k)}+\frac{t-i+1}{2}-1}\exp\Big\{-\Big(\frac{1}{\lambda^{(k)}} - \big(\boldsymbol{z}_{i,t}^{(k)}\big)'\big(\boldsymbol{V}_{i,t}^{(k)}\big)^{-1}\boldsymbol{z}_{i,t}^{(k)} + \sum_{r=i}^{t} y_r^2 + (\boldsymbol{z}^{(k)})'\big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}^{(k)}\Big)\tau_t\Big\}.$$

Let

$$g_{i,j}^{(k)} := g^{(k)} + \frac{j-i+1}{2} \tag{2.12}$$

and

$$\frac{1}{\lambda_{i,j}^{(k)}} := \frac{1}{\lambda^{(k)}} - \big(\boldsymbol{z}_{i,j}^{(k)}\big)'\big(\boldsymbol{V}_{i,j}^{(k)}\big)^{-1}\boldsymbol{z}_{i,j}^{(k)} + \sum_{r=i}^{j} y_r^2 + (\boldsymbol{z}^{(k)})'\big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}^{(k)}. \tag{2.13}$$

Thus, the conditional distribution of τ_t given $J_t^{(k)}$ and $\mathcal{F}_t$ follows a Gamma distribution with shape parameter $g_{i,t}^{(k)}$ and rate parameter $\lambda_{i,t}^{(k)}$. Notationally,

$$\tau_t|J_t^{(S_t)} = i, S_t = k, \mathcal{F}_t \sim \text{ Gamma } \big(g_{i,t}^{(k)}, \lambda_{i,t}^{(k)}\big) \text{ with p.d.f. } h_{i,t}^{(k)} \tag{2.14}$$

The posterior distribution of $(\boldsymbol{\beta}_t,\ \tau_t)$ in equation (2.8) is then a mixture of well defined distributions with weights $\xi_{it}^{(k)}$ defined in (2.6). If $\xi_{it}^{(k)}$ is known, the posterior means of the estimator would have closed form solutions. $\xi_{it}^{(k)}$ is so crucial in the parameter estimation that we call this quantity the forward weight. It turns out there is a recursive formula for the forward weight. Let us begin with the expansion of the definition of $\xi_{it}^{(k)}$ as follows:

$$\begin{aligned}
\xi_{i,t}^{(k)} &= P(J_t^{(S_t)} = i, S_t = k|\mathcal{F}_t) \\
&= \sum_{l=1}^{K} P(J_t^{(S_t)} = i, S_t = k, S_{t-1} = l|\mathcal{F}_t) \\
&= \sum_{l=1}^{K} f(y_t, J_t^{(S_t)} = i, S_t = k, S_{t-1} = l|\mathcal{F}_{t-1})/f(y_t|\mathcal{F}_{t-1})
\end{aligned}$$

$$= \sum_{l=1}^{K} f(y_t|J_t^{(k)} = i, S_{t-1} = l, \mathcal{F}_{t-1}) P(J_t^{(S_t)} = i|S_t = k, S_{t-1} = 1, \mathcal{F}_{t-1}) \times$$

$$P(S_t = k|S_{t-1} = l, \mathcal{F}_{t-1}) P(S_{t-1} = l|\mathcal{F}_{t-1}) / f(y_t|\mathcal{F}_{t-1}) \quad (2.15)$$

To simplify the above expression, we must understand the relationship between the state at time t and $t-1$ and conditional probability of $J_t^{(S_t)}$ at or before time t. For $l \neq k$, indicating that the last jump occurs at time t, then

$$P(J_t^{(S_t)} = i|S_t = k, S_{t-1} = l, \mathcal{F}_{t-1}) = \begin{cases} 1 & i = t \\ 0 & i < t \end{cases} \quad (2.16)$$

Due to first order Markov property $P(S_t = k|S_{t-1} = l, \mathcal{F}_{t-1}) = P(S_t = k|S_{t-1} = l) = p_{lk}$ and with the definition (2.7),

$$\xi_{i,t}^{(k)} \propto f(y_t|J_t^{(k)} = i, \mathcal{F}_{t-1}) \sum_{l \neq k} p_{lk} \xi_{t-1}^{(l)} \quad (2.17)$$

for only $i = t$, 0 otherwise. If $l = k$, the most recent jump must occur before time t and event $J_t^{(S_t)} = i|S_t = k, S_{t-1} = l, \mathcal{F}_{t-1}$ is equivalent to event $J_{t-1}^{(S_{t-1})} = i|S_{t-1} = k, \mathcal{F}_{t-1}$, i.e.

$$P(J_t^{(S_t)} = i|S_t = k = S_{t-1}, \mathcal{F}_{t-1}) = P(J_{t-1}^{(S_{t-1})} = i|S_{t-1} = k, \mathcal{F}_{t-1}) \quad i < t \quad (2.18)$$

Then for $i < t$ with $p_{kk} = P(S_t = k|S_{t-1} = k, \mathcal{F}_{t-1})$ and definition (2.6),

$$\begin{aligned}
\xi_{i,t}^{(k)} &\propto f(y_t|J_{t-1}^{(k)} = i, \mathcal{F}_{t-1}) P(J_{t-1}^{(S_{t-1})} = i|S_{t-1} = k, \mathcal{F}_{t-1}) p_{kk} P(S_{t-1} = k|\mathcal{F}_{t-1}) \\
&= f(y_t|J_{t-1}^{(k)} = i, \mathcal{F}_{t-1}) p_{kk} P(J_{t-1}^{(S_{t-1})} = i, S_{t-1} = k|\mathcal{F}_{t-1}) \\
&= f(y_t|J_{t-1}^{(k)} = i, \mathcal{F}_{t-1}) p_{kk} \xi_{i,t-1}^{(k)} \qquad (2.19)
\end{aligned}$$

To continue with the recursive formula of $\xi_{i,t}^{k}$, we prove in Theorem 1 and 2 in Appendix A that

$$f(y_t|J_t^{(S_t)} = t, S_t = k, \mathcal{F}_{t-1}) = \pi^{-\frac{1}{2}} \frac{\phi_{t,t}^{(k)}}{\phi_{0,0}^{(k)}} \qquad (2.20)$$

and

$$f(y_t|J_{t-1}^{(S_{t-1})} = i, S_{t-1} = k, \mathcal{F}_{t-1}) = \pi^{-\frac{1}{2}} \frac{\phi_{i,t}^{(k)}}{\phi_{i,t-1}^{(k)}} \qquad (2.21)$$

Where

$$\phi_{0,0}^{(k)} := \left|\boldsymbol{V}^{(k)}\right|^{\frac{1}{2}} \Gamma(g^{(k)}) (\lambda^{(k)})^{g^{(k)}} \qquad (2.22)$$

$$\phi_{i,j}^{(k)} := \left|\boldsymbol{V}_{i,j}^{(k)}\right|^{\frac{1}{2}} \Gamma(g_{i,j}^{(k)}) (\lambda_{i,j}^{(k)})^{g_{i,j}^{(k)}} \qquad \forall 1 \leq i, j \leq t \qquad (2.23)$$

and $\boldsymbol{V}_{i,j}^{(k)}, g_{i,j}^{(k)}$ and $\lambda_{i,j}^{(k)}$ are defined in (2.9), (2.12) and (2.13).

In summary the working recursive forward weight formula for $\xi_{i,t}^{(k)}$ is

$$\xi_{i,t}^{(k)*} := \begin{cases} \frac{\phi_{t,t}^{(k)}}{\phi_{0,0}^{(k)}} \sum_{l \neq k} p_{lk} \xi_{t-1}^{(l)} & i = t \\ \frac{\phi_{i,t}^{(k)}}{\phi_{i,t-1}^{(k)}} p_{kk} \xi_{i,t-1}^{(k)} & i < t \end{cases} \qquad (2.24)$$

which can be normalized to

$$\xi_{i,t}^{(k)} = \frac{\xi_{i,t}^{(k)*}}{\sum_{k=1}^{K}\sum_{i=1}^{t}\xi_{i,t}^{(k)*}} \tag{2.25}$$

At this stage, we are able to estimate the regression parameters and regime status. First, with (2.8), (2.11) and (2.25), the filtering estimation of $\boldsymbol{\beta}_t$ is defined as the posterior mean of $\boldsymbol{\beta}_t$, which can be easily shown as

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{t|t} &:= E\big[\boldsymbol{\beta}_t|\mathcal{F}_t\big] \\
&= \int\!\!\int \boldsymbol{\beta}_t f(\boldsymbol{\beta}_t|\tau_t, \mathcal{F}_t) f(\tau_t|\mathcal{F}_t)\ \mathrm{d}\boldsymbol{\beta}_t\ \mathrm{d}\tau_t \\
&= \int\!\!\int \sum_{k=1}^{K}\sum_{i=1}^{t} \boldsymbol{\beta}_t f(\boldsymbol{\beta}_t|\tau_t, J_t^{(k)} = i, \mathcal{F}_t) f(\tau_t|J_t^{(k)} = i, \mathcal{F}_t)\xi_{i,t}^{(k)}\ \mathrm{d}\boldsymbol{\beta}_t\ \mathrm{d}\tau_t \\
&= \sum_{k=1}^{K}\sum_{i=1}^{t} \boldsymbol{z}_{i,t}^{(k)}\xi_{i,t}^{(k)} \qquad (2.26)
\end{aligned}$$

and the posterior variance-covariance of $\boldsymbol{\beta}_t$ is shown below and proved in Theorem 3 in Appendix A.

$$\begin{aligned}
\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t|\mathcal{F}_t} = &\sum_{k=1}^{K}\sum_{i=1}^{t}\left(\frac{\boldsymbol{V}_{it}^{(k)}}{2\lambda_{it}^{(k)}\big(g_{it}^{(k)}-1\big)} + \boldsymbol{z}_{it}^{(k)}(\boldsymbol{z}_{it}^{k})'\right)\xi_{it}^{(k)} \\
&- \sum_{k=1}^{K}\sum_{i=1}^{t}\boldsymbol{z}_{it}^{(k)}\xi_{it}^{(k)}\left(\sum_{k=1}^{K}\sum_{i=1}^{t}\boldsymbol{z}_{it}^{(k)}\xi_{it}^{(k)}\right)' \qquad (2.27)
\end{aligned}$$

Second, the filtering estimation of σ_t, defined as the posterior mean of σ_t is

$$\widehat{\sigma}_{t|t} := E[\sigma_t|\mathcal{F}_t]$$

$$= \int \sigma_t f(\sigma_t|\mathcal{F}_t)\ \mathrm{d}\sigma_t$$

$$= \int \sigma_t \sum_{k=1}^{K}\sum_{i=1}^{t} f(\sigma_t|J_t^{(S_t)} = i, S_t = k, \mathcal{F}_t) f(J_t^{(S_t)} = i, S_t = k|\mathcal{F}_t)\ \mathrm{d}\sigma_t$$

$$= \sum_{k=1}^{K}\sum_{i=1}^{t} \xi_{i,t}^{(k)} E(\sigma_t|J_t^{(k)} = i, \mathcal{F}_t)$$

Since posterior τ_t is a gamma distribution defined in (2.14), σ_t is related to τ_t by the assumption (A2) and also by using properties of gamma distribution in Theorem 7 in Appendix A, the following holds.

$$\widehat{\sigma}_{t|t} = \sum_{k=1}^{K}\sum_{i=1}^{t} \xi_{i,t}^{(k)} (2\lambda_{i,t}^{(k)})^{-\frac{1}{2}} \frac{\Gamma(g_{i,t}^{(k)} - \frac{1}{2})}{\Gamma(g_{i,t}^{(k)})} \tag{2.28}$$

The variance of posterior σ_t is shown below and proved also in Theorem 3 in Appendix A.

$$\begin{aligned}\Sigma_{\sigma_t|\mathcal{F}_t} = \sum_{k=1}^{K}\sum_{i=1}^{t} & \frac{\xi_{it}^{(k)}}{2\lambda_{it}^{(k)}\big(g_{it}^{(k)} - 1\big)} \\ & - \left(\sum_{k=1}^{K}\sum_{i=1}^{t} \xi_{it}^{(k)} \big(2\lambda_{it}^{(k)}\big)^{-\frac{1}{2}} \frac{\Gamma\big(g_{it}^{(k)} - \frac{1}{2}\big)}{\Gamma\big(g_{it}^{(k)}\big)}\right)^2 \end{aligned} \tag{2.29}$$

To complete this section, it is worth mentioning how to make the inference on the regime. Being at state k at time t can be viewed as a Bernoulli random variable. Thus the posterior mean of $S_t = k$ is the estimated $P(S_t = k|\mathcal{F}_t)$

defined in equation (2.7) which is the sum of forward weight (2.25), i.e.,

$$E\big[I_{\{S_t=k\}}|\mathcal{F}_t\big] = \xi_t^{(k)}, \quad \text{for each } k = 1, \ldots, K \tag{2.30}$$

Naturally the variance of S_t at state k is

$$Var\big(I_{\{S_t=k\}}|\mathcal{F}_t\big) = \xi_t^{(k)}(1-\xi_t^{(k)}), \quad \text{for each } k = 1, \ldots, K \tag{2.31}$$

2.3 Backward Filtering Estimation of Parameters

Since Markov chain $\{S_t\}$ is reversible from assumption (A1), there exits a backward transition probability matrix Q defined similar to (2.2) with transition probabilities q_{ij}'s. The transition probabilities q_{ij} are related to forward transition probabilities by

$$\begin{aligned} q_{ij} = P(S_{t-1}=j|S_t=i) &= \frac{P(S_{t-1}=j, S_t=i)}{P(S_t=i)} \\ &= \frac{P(S_{t-1}=j)P(S_t=i|S_{t-1}=j)}{P(S_t=i)} = \frac{\pi_j}{\pi_i}p_{ji} \end{aligned} \tag{2.32}$$

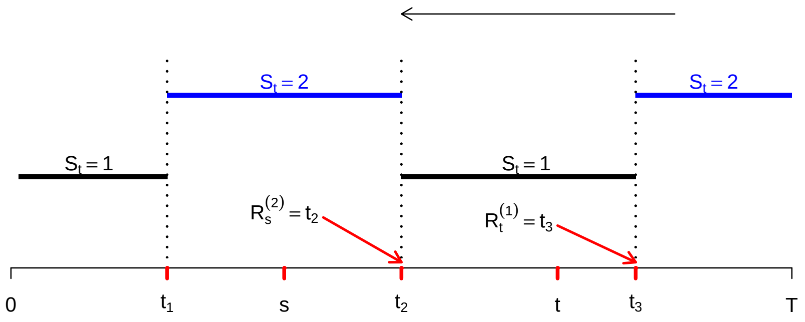
The stationary probability distribution of the reverse chain is denoted by $\widetilde{\boldsymbol{\pi}}$ related to Q in a similar manner as in (2.3). Estimation of the parameters at time t based on information from t to a future time index T is called the backward filtering estimation. If we focus on the system from t to T and read the time index backward from T to t, it would not be difficult to understand the mechanism of backward filtering is the mirror image of that of forward filtering.

Now we use notations $\boldsymbol{y}_{t,T}$, $\boldsymbol{x}_{t,T}$, and $\mathcal{F}_{t,T}$ defined the same as in (2.4) and define the most recent jump to the state k at time t from other state in reverse time order as

$$R_t^{(k)} := \min\{j \geq t : k = S_t = \cdots = S_{j-1} \neq S_j\} \tag{2.33}$$

Figure 2.3 gives an example of $R_t^{(k)}$ in a two-state system. For example, at time t the system is in state 1 and $R_t^{(1)}$ is the most recent switching index from state 2 viewing time index backwards from T to 1. This index is obviously t_3 indicated by arrow on the graph.

Figure 2.3: Illustration of $R_t^{(k)}$ for a concrete two-state system



Very similar to the definition in (2.6), the probability of $R_t^{(k)} = j$ given information from T to t is called backward recursive weight and defined as

$$\eta_{t,j}^{(k)} := P(R_t^{(S_t)} = j, S_t = k|\mathcal{F}_{t,T}) = P(R_t^{(k)} = j|\mathcal{F}_{t,T}) \tag{2.34}$$

and this quantity is also very important in the statistical estimation. In the same spirit of (2.7), the backward filtering estimation of the regime $P(S_t = k|\mathcal{F}_{t,T})$ is the sum of the backward weights for all $j \geq t$, i.e.

$$\eta_t^{(k)} = P(S_t = k|\mathcal{F}_{t,T}) = \sum_{j=t}^{T} \eta_{t,j}^{(k)} \tag{2.35}$$

The goal of this section is again to estimate the posterior distributions of $\boldsymbol{\beta}_t$ and σ_t and the probability of the system is on a certain regime at a given time point based on the information from t through T. A backward weight recursive formula is derived for the purpose of achieving this goal. For every step or formula presented below, readers may find counterparts in section 2.2. Lengthy proofs are skipped to keep this thesis readable. Key results are presented with relevant explanations.

Like equation (2.8), posterior distribution of model parameters is a mixture of the product of the conditional distributions of $\boldsymbol{\beta}_t$ and σ_t weighted by backward weights.

$$f(\boldsymbol{\beta}_t, \tau_t|\mathcal{F}_{t,T}) = \sum_{k=1}^{K} \sum_{j=t}^{T} f(\boldsymbol{\beta}_t|\tau_t, R_t^{(k)} = j, \mathcal{F}_{t,T}) f(\tau_t|R_t^{(k)} = j, \mathcal{F}_{t,T}) \eta_{t,j}^{(k)} \tag{2.36}$$

It can be shown that for $j \geq t$,

$$\begin{aligned}
& f(\boldsymbol{\beta}_t | \tau_t, R_t^{(k)} = j, \mathcal{F}_{t,T}) \\
& \propto f(\boldsymbol{y}_{t,j} | \boldsymbol{\beta}_t, \tau_t, R_t^{(k)} = j, \boldsymbol{x}_{t,T}) f(\boldsymbol{\beta}_t | \tau_t, R_t^{(k)} = j, \boldsymbol{x}_{t,T}) \\
& \propto \exp\Big\{ - \tau_t \boldsymbol{\beta}_t' \Big(\sum_{r=t}^{j} \boldsymbol{x}_r \boldsymbol{x}_r' + (\boldsymbol{V}^{(k)})^{-1} \Big) \boldsymbol{\beta}_t + 2\tau_t \boldsymbol{\beta}_t' \Big(\sum_{r=t}^{j} \boldsymbol{x}_r y_r + \big(\boldsymbol{V}^{(k)}\big)^{-1} \boldsymbol{z}^{(k)} \Big) \Big\} \\
& \propto \exp\Big\{ \tau_t \big(\boldsymbol{\beta}_t - \boldsymbol{z}_{t,j}^{(k)}\big)' \big(\boldsymbol{V}_{t,j}^{(k)}\big)^{-1} \big(\boldsymbol{\beta}_t - \boldsymbol{z}_{t,j}^{(k)}\big) \Big\}
\end{aligned}$$

where $\boldsymbol{V}_{i,j}^{(k)}$ and $\boldsymbol{z}_{i,j}^{(k)}$ are defined in (2.9) and (2.10) respectively. The conditional posterior distribution of $\boldsymbol{\beta}_t$ is a normal distribution, i.e,

$$\boldsymbol{\beta}_t | \tau_t, R_t^{(k)} = j, \mathcal{F}_{t,T} \sim \ \mathrm{N} \left(\boldsymbol{z}_{t,j}^{(k)}, \ \frac{\boldsymbol{V}_{t,j}^{(k)}}{2\tau_t} \right) \tag{2.37}$$

Similarly,

$$\begin{aligned}
& f(\tau_t | R_t^{(k)} = j, \mathcal{F}_{t,T}) \\
& = \int \frac{f(\tau_t, \boldsymbol{\beta}_t, \boldsymbol{y}_{t,T} | R_t^{k}, \boldsymbol{x}_{t,T})}{f(\boldsymbol{y}_{t,T} | R_t^{(k)} = j, \boldsymbol{x}_{t,T})} \mathrm{d}\boldsymbol{\beta}_t \\
& \propto \int f(\boldsymbol{y}_{t,j} | \tau_t, \boldsymbol{\beta}_t, R_t^{(k)} = j \boldsymbol{x}_{t,T}) f(\boldsymbol{\beta}_t | \tau_t, R_t^{(k)} = j, \boldsymbol{x}_{t,T}) f(\tau_t | R_t^{(k)} = j, \boldsymbol{x}_{t,T}) \mathrm{d}\boldsymbol{\beta}_t \\
& \propto \tau_t^{\frac{j-t+1}{2}} \exp\Big\{ - \tau_t \Big(- \big(\boldsymbol{z}_{t,j}^{(k)}\big)' \big(\boldsymbol{V}_{t,j}^{(k)}\big)^{-1} \boldsymbol{z}_{t,j}^{(k)} + \sum_{r=t}^{j} y_r^2 + (\boldsymbol{z}^{(k)})' \big(\boldsymbol{V}^{(k)}\big)^{-1} \boldsymbol{z}^{(k)} \Big) \Big\} \\
& \quad \cdot \tau_t^{g^{(k)}-1} \exp\Big\{ - \frac{\tau_t}{\lambda^{(k)}} \Big\} \\
& \propto \tau_t^{g_{t,j}^{(k)}-1} \exp\Big\{ - \frac{\tau_t}{\lambda_{t,j}^{(k)}} \Big\}
\end{aligned}$$

where $g_{i,j}^{(k)}$ and $\lambda_{ij}^{(k)}$ are defined in (2.12) and (2.13) respectively. The conditional posterior of τ_t is a gamma distribution, i.e.,

$$\tau_t | R_t^{(S_t)} = j, S_t = k, \mathcal{F}_{t,T} \sim \text{ Gamma } \left(g_{t,j}^{(k)},\ \lambda_{t,j}^{(k)} \right) \tag{2.38}$$

Next, to find the recursive formula for backward weight, $\eta_{t,j}^{(k)}$ can be expanded as, similar to equation (2.15),

$$\begin{aligned} \eta_{t,j}^{(k)} &= \sum_{l=1}^{K} P(R_t^{(S_t)} = j, S_t = k, S_{t+1} = l | \mathcal{F}_{t,T}) \\ &= \sum_{l=1}^{K} f(y_t | R_t^{(k)} = j, \mathcal{F}_{t+1,T}) P(R_t^{(S_t)} = j | S_t = k, \mathcal{F}_{t+1,T}) q_{lk} \eta_{t+1}^{(l)} / f(y_t | \mathcal{F}_{t+1,T}) \end{aligned}$$

If $l \neq k$, by definition of $R_t^{(S_t)}$, the nearest switch after time t must be at time t and be impossible at other times, so

$$P(R_t^{(S_t)} = j | S_t = k, S_{t+1} = l, \mathcal{F}_{t+1,T}) = \begin{cases} 1 & j = t \\ 0 & j > t \end{cases} \tag{2.39}$$

If $l = k$, then switch time j must be greater than t and so event $R_t^{(S_t)} = j | S_t = k = S_{t+1}, \mathcal{F}_{t+1,T}$ is equivalent to event $R_{t+1}^{(S_{t+1})} = j | S_{t+1} = k, \mathcal{F}_{t+1,T}$, i.e.

$$P(R_t^{(S_t)} = j | S_t = k, S_{t+1} = l, \mathcal{F}_{t+1,T}) = P(R_{t+1}^{(S_{t+1})} = j | S_{t+1} = k, \mathcal{F}_{t+1,T}) \tag{2.40}$$

Also with the fact that $P(R_{t+1}^{(S_{t+1})} = j|S_{t+1} = k, \mathcal{F}_{t+1,T})P(S_{t+1} = k|\mathcal{F}_{t+1,T}) = P(R_{t+1}^{(k)} = j|\mathcal{F}_{t+1,T})$,

$$\eta_{t,j}^{(k)} \propto \begin{cases} f(y_t|R_t^{(k)} = j, \mathcal{F}_{t+1,T}) \sum_{l \neq k} q_{lk} \eta_{t+1}^{(l)} & j = t, \\ f(y_t|R_{t+1}^{(k)} = j, \mathcal{F}_{t+1,T}) q_{kk} \eta_{t+1,j}^{(k)} & j > t \end{cases} \tag{2.41}$$

It can be proved, technically similar to the proofs in Theorem 1 and 2 in Appendix A, that

$$f(y_t|R_t^{(k)} = t, \mathcal{F}_{t+1,T}) = \pi^{-\frac{1}{2}} \frac{\phi_{t,t}^{(k)}}{\phi_{0,0}^{(k)}} \tag{2.42}$$

and

$$f(y_t|R_{t+1}^{(k)} = j, \mathcal{F}_{t+1,T}) = \pi^{-\frac{1}{2}} \frac{\phi_{t,j}^{(k)}}{\phi_{t+1,j}^{(k)}} \tag{2.43}$$

where $\phi_{0,0}^{(k)}$ and $\phi_{i,j}^{(k)}$ are defined in equation (2.22) and (2.23) respectively for any $i, j \in 1, 2, \ldots, T$. The recursive formula for the working backward weights is

$$\eta_{t,j}^{(k)*} := \begin{cases} \frac{\phi_{t,t}^{(k)}}{\phi_{0,0}^{(k)}} \sum_{l \neq k} q_{lk} \eta_{t+1}^{(l)} & j = t \\ \frac{\phi_{t,j}^{(k)}}{\phi_{t+1,j}^{(k)}} q_{kk} \eta_{t+1,j}^{(k)} & j > t \end{cases} \tag{2.44}$$

whose normalized version is

$$\eta_{t,j}^{(k)} = \frac{\eta_{t,j}^{(k)*}}{\sum_{k=1}^{K} \sum_{j=t}^{T} \eta_{t,j}^{(k)*}} \tag{2.45}$$

Similar to the proofs of equation (2.26) and (2.28), the estimation of $\boldsymbol{\beta}_t$ and

σ_t are their posterior means in the context of backward filtering, i.e.

$$\widehat{\boldsymbol{\beta}}_{t|t,T} = \sum_{k=1}^{K}\sum_{j=t}^{T} \boldsymbol{z}_{t,j}^{(k)}\eta_{t,j}^{(k)} \tag{2.46}$$

and

$$\widehat{\sigma}_{t|t,T} = \sum_{k=1}^{K}\sum_{j=t}^{T} \eta_{t,j}^{(k)}(2\lambda_{t,j}^{(k)})^{-\frac{1}{2}}\frac{\Gamma(g_{t,j}^{(k)}-\frac{1}{2})}{\Gamma(g_{t,j}^{(k)})} \tag{2.47}$$

respectively. The posterior variances of $\boldsymbol{\beta}_t$ and σ_t are

$$\begin{aligned}\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t|\mathcal{F}_{tT}} = &\sum_{k=1}^{K}\sum_{j=t}^{T}\left(\frac{\boldsymbol{V}_{tj}^{(k)}}{2\lambda_{tj}^{(k)}\big(g_{tj}^{(k)}-1\big)} + \boldsymbol{z}_{tj}^{(k)}(\boldsymbol{z}_{tj}^{k})'\right)\eta_{tj}^{(k)} \\ &- \sum_{k=1}^{K}\sum_{j=t}^{T}\boldsymbol{z}_{tj}^{(k)}\eta_{tj}^{(k)}\left(\sum_{k=1}^{K}\sum_{j=t}^{T}\boldsymbol{z}_{tj}^{(k)}\eta_{tj}^{(k)}\right)'\end{aligned} \tag{2.48}$$

and

$$\begin{aligned}\Sigma_{\sigma_t|\mathcal{F}_{tT}} = &\sum_{k=1}^{K}\sum_{j=t}^{T}\frac{\eta_{tj}^{(k)}}{2\lambda_{tj}^{(k)}\big(g_{tj}^{(k)}-1\big)} \\ &- \left(\sum_{k=1}^{K}\sum_{j=t}^{T}\eta_{tj}^{(k)}\big(2\lambda_{tj}^{(k)}\big)^{-\frac{1}{2}}\frac{\Gamma\big(g_{tj}^{(k)}-\frac{1}{2}\big)}{\Gamma\big(g_{tj}^{(k)}\big)}\right)^2\end{aligned} \tag{2.49}$$

Proofs are skipped but interested readers may refer to the proofs in Theorem 3 in Appendix A. Similar to (2.30) and (2.31), the inference of the regime can be presented by

$$E\big[I_{\{S_t=k\}}|\mathcal{F}_{tT}\big] = \eta_t^{(k)}, \quad \text{for each } k=1,\ldots,K \tag{2.50}$$

and

$$Var\big(I_{\{S_t=k\}}|\mathcal{F}_{tT}\big) = \eta_t^{(k)}(1-\eta_t^{(k)}), \quad \text{for each } k=1,\dots,K \qquad (2.51)$$

2.4 Predictive Estimation of Parameters

In this section, we discuss two types of predictive estimation, forward and backward prediction. As its name indicates, the information related to the interesting statistical quantities is not available at the time of interest. The estimation of the properties of the parameters at time t given all information available from 0 to time $t-1$ is called forward predictive estimation or forecast which has practical implication in time series data analysis. In particular the probability distribution of $\boldsymbol{\beta}_t$ and τ_t given information up to $t-1$ is

$$\begin{aligned}
&f(\boldsymbol{\beta}_t, \tau_t|\mathcal{F}_{t-1})\\
&=\sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{i=1}^{t} f(\boldsymbol{\beta}_t, \tau_t|J_t^{(k)}=i, S_{t-1}=l, \mathcal{F}_{t-1})P(J_t^{(S_t)}=j|S_t=k, S_{t-1}=l, \mathcal{F}_{t-1})\cdot\\
&\quad P(S_t=k|S_{t-1}=l, \mathcal{F}_{t-1})f(S_{t-1}=l|\mathcal{F}_{t-1})\\
&=\sum_{k=1}^{K}\sum_{l=1, l\neq k}^{K} f(\boldsymbol{\beta}_t, \tau_t|J_t^{(S_t)}=t, S_t=k, S_{t-1}=l, \mathcal{F}_{t-1})p_{lk}\xi_{t-1}^{(l)}+\\
&\quad\sum_{k=1}^{K}\sum_{i=1}^{t-1} f(\boldsymbol{\beta}_t, \tau_t|J_{t-1}^{(S_{t-1})}=i, S_t=S_{t-1}=k, \mathcal{F}_{t-1})p_{kk}\xi_{i,t-1}^{(k)}
\end{aligned}$$

where we use the relation in equation (2.16) and (2.18) in section 2.2 to simplify $P(J_t^{(S_t)} = j | S_t = k, S_{t-1} = l, \mathcal{F}_{t-1}) P(S_t = k | S_{t-1} = l, \mathcal{F}_{t-1}) f(S_{t-1} = l | \mathcal{F}_{t-1})$ for cases $l \neq k$ and $l = k$ respectively. Also if $l \neq k$, the regime change must occur at time t and the system regenerates at time t. The joint distribution of $(\boldsymbol{\beta}_t, \tau_t)$, given $J_t^{(S_t)} = t, S_t = k \neq S_{t-1} = l$ and $\mathcal{F}_{t-1}$ is the product of a normal and a gamma distribution with all prior parameters, i.e.

$$f(\boldsymbol{\beta}_t, \tau_t | J_t^{(S_t)} = t, S_t = k \neq S_{t-1} = l, \mathcal{F}_{t-1}) = f_{0,0}^{(k)} h_{0,0}^{(k)},$$

based on the assumption (A2). On the other hand, when $l = k$,

$$\begin{aligned} &f(\boldsymbol{\beta}_t, \tau_t | J_{t-1}^{(S_{t-1})} = i, S_t = S_{t-1} = k, \mathcal{F}_{t-1}) \\ &\qquad = f(\boldsymbol{\beta}_t | \tau_t, J_{t-1}^{(k)} = i, \mathcal{F}_{t-1}) f(\tau_t | J_{t-1}^{(k)} = i, \mathcal{F}_{t-1}) \end{aligned}$$

Since $\boldsymbol{\beta}$'s are in the same state k at time $t-1$ and t, and information at time t is not available, the estimation of $\boldsymbol{\beta}$ at time t would be the same as that at $t-1$ based on the information available at $t-1$. Thus $\boldsymbol{\beta}_t | \tau_t, J_{t-1}^{(k)} = i, \mathcal{F}_{t-1} \sim \mathrm{N}\Big(\boldsymbol{z}_{i,t-1}^{(k)},\ \frac{\boldsymbol{V}_{i,t-1}^{(k)}}{2\tau_t}\Big)$ by (2.11) and $\tau_t | J_{t-1}^{(k)} = i, \mathcal{F}_{t-1} \sim$ Gamma $\big(g_{i,t-1}^{(k)}, \lambda_{i,t-1}^{(k)}\big)$ by (2.14). Accordingly,

$$f(\boldsymbol{\beta}_t, \tau_t | J_{t-1}^{(S_{t-1})} = i, S_t = S_{t-1} = k, \mathcal{F}_{t-1}) = f_{i,t-1}^{(k)} h_{i,t-1}^{(k)}$$

The three important quantities in the forward predictive estimation setting are shown below. The forecast of $\boldsymbol{\beta}_t$ is

$$\widehat{\boldsymbol{\beta}}_{t|t-1} := \sum_{k=1}^{K} \sum_{l=1,l\neq k}^{K} \boldsymbol{z}^{(k)} p_{lk} \xi_{t-1}^{(l)} + \sum_{k=1}^{K} \sum_{i=1}^{t-1} \boldsymbol{z}_{i,t-1}^{(k)} p_{kk} \xi_{i,t-1}^{(k)} \tag{2.52}$$

and the forecast of σ_t is

$$\begin{aligned} \widehat{\sigma}_{t|t-1} := \sum_{k=1}^{K} \sum_{l=1,l\neq k}^{K} (2\lambda^{(k)})^{-\frac{1}{2}} \frac{\Gamma(g^{(k)} - \frac{1}{2})}{\Gamma(g^{(k)})} p_{lk} \xi_{t-1}^{(l)} \\ + \sum_{k=1}^{K} \sum_{i=1}^{t-1} (2\lambda_{i,t-1}^{(k)})^{-\frac{1}{2}} \frac{\Gamma(g_{i,t-1}^{(k)} - \frac{1}{2})}{\Gamma(g_{i,t-1}^{(k)})} p_{kk} \xi_{i,t-1}^{(k)} \end{aligned} \tag{2.53}$$

and finally the forecast of the probability of the regime is

$$\begin{aligned} P(S_t = k|\mathcal{F}_{t-1}) &= \sum_{l=1}^{K} P(S_t = k|S_{t-1} = l, \mathcal{F}_{t-1}) P(S_{t-1} = l|\mathcal{F}_{t-1}) \\ &= \sum_{l=1}^{K} p_{lk} \xi_{i,t-1}^{(l)} \end{aligned} \tag{2.54}$$

The second type of prediction in this section is the backward prediction which appears to be impractical in reality since time does not travel from future to the past. But it has useful theoretical properties and is an essential technical step in smoothing estimation that will be discussed in the next section. Like backward filtering in section 2.3, viewers read the time sequence from a future time T to the current time t of interest, and the difference in backward prediction is that the information at t is not available. We derive the joint distribution of $(\boldsymbol{\beta}_t,\ \tau_t)$ given $\mathcal{F}_{t+1,T}$, i.e.,

$$\begin{aligned}
&f(\boldsymbol{\beta}_t, \tau_t|\mathcal{F}_{t+1,T})\\
&=\sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{j=t}^{T} f(\boldsymbol{\beta}_t, \tau_t|R_t^{(k)}=j, S_{t+1}=l, \mathcal{F}_{t+1,T}) f(R_t^{(S_t)}=j|S_t=k, S_{t+1}=l, \mathcal{F}_{t+1,T})\cdot\\
&\quad f(S_t=k|S_{t+1}=l, \mathcal{F}_{t+1,T}) f(S_{t+1}=l|\mathcal{F}_{t+1,T})\\
&=\sum_{k=1}^{K}\sum_{l=1,l\neq k}^{K} f_{0,0}^{(k)} h_{0,0}^{(k)} q_{lk}\eta_{t+1}^{(l)} + \sum_{k=1}^{K}\sum_{j=t+1}^{T} f_{t+1,j}^{(k)} h_{t+1,j}^{(k)} q_{kk}\eta_{t+1,j}^{(k)} \qquad (2.55)
\end{aligned}$$

The above result uses the fact in equation (2.39) and (2.40), the fact that the conditional distribution of $\boldsymbol{\beta}_t$ given $R_{t+1}^{(k)}$ and $\mathcal{F}_{t+1,T}$ is $\mathrm{N}\Big(\boldsymbol{z}_{t+1,j}^{(k)},\ \frac{\boldsymbol{V}_{t+1,j}^{(k)}}{2\tau_t}\Big)$ and the fact that the conditional distribution of τ_t given $R_{t+1}^{(k)}$ and $\mathcal{F}_{t+1,T}$ is Gamma $\big(g_{t+1,j}^{(k)}, \lambda_{t+1,j}^{(k)}\big)$.

The backward prediction of the parameters are of no practical interest and therefore omitted in this section. But the usage of their probability distribution is fully realized in the next section.

2.5 Smoothing Estimation of Parameters

Suppose the observations are available from 0 to T, but the parameters to be estimated is at time t $(t \leq T)$. The estimation of parameters at time t $(t \leq T)$ using all the information available up to T is called smoothing estimation. In this section we are interested in the posterior distribution of $\boldsymbol{\beta}_t$ and σ_t given $\mathcal{F}_T$ and the posterior probabilities of the regime, i.e.,

$P(S_t = k|\mathcal{F}_T)$ for $k = 1, \ldots, K$.

By applying Bayes' theorem and Markov property, the following equation holds.

$$f(\boldsymbol{\beta}_t, \tau_t|\mathcal{F}_T) \propto \sum_{k=1}^{K} \frac{f(\boldsymbol{\beta}_t, \tau_t, S_t = k|\mathcal{F}_t) f(\boldsymbol{\beta}_t, \tau_t, S_t = k|\mathcal{F}_{t+1,T})}{f(\boldsymbol{\beta}_t, \tau_t, S_t = k)} \tag{2.56}$$

This result states that a "two-sided" conditional density of parameters is proportional to the product of the two "one-sided" conditional densities divided by its prior density (Yao, 1984, proposition 4.2). These two conditional probabilities are actually forward filter estimates of the parameters and the backward prediction of the parameters. By Markov property, y_1 depends on y_2, y_2 depends on y_3, …, y_{t-1} depends on y_t, and also given state $S_t = k$, y_t is independent of $y_{t+1}, y_{t+2}, \ldots, y_T$. Thus given the state and the parameters at time t, $\mathcal{F}_t$ and $\mathcal{F}_{t+1,T}$ are independent.

Proof. By property of conditional probability and Bayes' theorem

$$f(\boldsymbol{\beta}_t, \tau_t|\mathcal{F}_T) = \frac{\sum_{k=1}^{K} f(\mathcal{F}_t, \mathcal{F}_{t+1,T}|\boldsymbol{\beta}_t, \tau_t, S_t = k) f(\boldsymbol{\beta}_t, \tau_t, S_t = k)}{f(\mathcal{F}_T)}$$

By Markov property and independence of $y_1, \ldots, y_t$ and y_{t+1}, y_T given the state at t, the above is

$$\frac{\sum_{k=1}^{K} f(\mathcal{F}_t|\boldsymbol{\beta}_t, \tau_t, S_t = k) f(\mathcal{F}_{t+1,T}|\boldsymbol{\beta}_t, \tau_t, S_t = k) f(\boldsymbol{\beta}_t, \tau_t, S_t = k)}{f(\mathcal{F}_T)}$$

Using Bayes' theorem again, the above is further expanded into

$$\frac{1}{f(\mathcal{F}_T)}\sum_{k=1}^{K}\frac{f(\boldsymbol{\beta}_t,\tau_t,S_t=k|\mathcal{F}_t)f(\mathcal{F}_t)}{f(\boldsymbol{\beta}_t,\tau_t,S_t=k)}\cdot \\ \frac{f(\boldsymbol{\beta}_t,\tau_t,S_t=k|\mathcal{F}_{t+1,T})f(\mathcal{F}_{t+1,T})}{f(\boldsymbol{\beta}_t,\tau_t,S_t=k)}\cdot f(\boldsymbol{\beta}_t,\tau_t,S_t=k)$$

Finally,

$$f(\boldsymbol{\beta}_t,\tau_t|\mathcal{F}_T)\propto\sum_{k=1}^{K}\frac{f(\boldsymbol{\beta}_t,\tau_t,S_t=k|\mathcal{F}_t)f(\boldsymbol{\beta}_t,\tau_t,S_t=k|\mathcal{F}_{t+1,T})}{f(\boldsymbol{\beta}_t,\tau_t,S_t=k)}$$

where $f(\mathcal{F}_t), f(\mathcal{F}_{t+1,T})$ and $f(\mathcal{F}_T)$ are constant. □

Similar to section 2.2 and 2.3, the joint density of $(\boldsymbol{\beta}_t,\tau_t)$ given $\mathcal{F}_T$ is a mixture of the product of a normal and gamma distribution as follows:

$$\begin{aligned} & f(\boldsymbol{\beta}_t,\tau_t|\mathcal{F}_T) \\ &=\sum_{k=1}^{K}\sum_{i=1}^{t}\sum_{j=t}^{T} f(\boldsymbol{\beta}_t,\tau_t,S_t=k,J_t^{(S_t)}=i,R_t^{(S_t)}=j|\mathcal{F}_T) \\ &=\sum_{k=1}^{K}\sum_{i=1}^{t}\sum_{j=t}^{T} f(\boldsymbol{\beta}_t|\tau_t,S_t=k,J_t^{(S_t)}=i,R_t^{(S_t)}=j,\mathcal{F}_T) \\ &\quad\cdot f(\tau_t|S_t=k,J_t^{(S_t)}=i,R_t^{(S_t)}=j,\mathcal{F}_T)\cdot P(S_t=k,J_t^{(S_t)}=i,R_t^{(S_t)}=j|\mathcal{F}_T) \end{aligned} \tag{2.57}$$

Define smoothing weight as

$$\alpha_{itj}^{(k)}=P(S_t=k,J_t^{(S_t)}=i,R_t^{(S_t)}=j|\mathcal{F}_T) \tag{2.58}$$

It turns out that $\alpha_{itj}^{(k)}$ is a derived statistic from $\xi_{i,t}^{(k)}$ and $\eta_{t,j}^{(k)}$. The smoothed probability of the regime can be naturally defined as

$$\alpha_t^{(k)} := P(S_t = k|\mathcal{F}_T) = \sum_{i=1}^{t}\sum_{j=t}^{T} \alpha_{itj}^{(k)} \tag{2.59}$$

We will show that equation (2.56) and (2.57) are equivalent up to a normalizing constant. Let us begin with the expression on the right hand side of equation (2.56).

$$\begin{aligned}
&\frac{f(\boldsymbol{\beta}_t, \tau_t, S_t = k|\mathcal{F}_t) f(\boldsymbol{\beta}_t, \tau_t, S_t = k|\mathcal{F}_{t+1,T})}{f(\boldsymbol{\beta}_t, \tau_t, S_t = k)} \\
&= \frac{\sum_{i=1}^{t} f(\boldsymbol{\beta}_t, \tau_t, S_t = k, J_t^{(k)} = i|\mathcal{F}_t) \sum_{j=t+1}^{T} f(\boldsymbol{\beta}_t, \tau_t, S_t = k, R_{t+1}^{(k)} = j|\mathcal{F}_{t+1,T})}{f(\boldsymbol{\beta}_t|\tau_t, S_t = k) f(\tau_t|S_t = k) P(S_t = k)} \\
&= \frac{\left(\sum_{i=1}^{t} f_{i,t}^{(k)} h_{i,t}^{(k)} \xi_{i,t}^{(k)}\right)\left(\sum_{l=1,l\neq k}^{K} f_{0,0}^{(k)} h_{0,0}^{(k)} q_{lk} \eta_{t+1}^{(l)} + \sum_{j=t+1}^{T} f_{t+1,j}^{(k)} h_{t+1,j}^{(k)} q_{kk} \eta_{t+1,j}^{(k)}\right)}{f_{0,0}^{(k)} h_{0,0}^{(k)} \pi_k} \\
&= \sum_{i=1}^{t} f_{i,t}^{(k)} h_{i,t}^{(k)} \xi_{i,t}^{(k)} \sum_{l=1,l\neq k}^{K} \frac{q_{lk}\eta_{t+1}^{(l)}}{\pi_k} + \frac{q_{kk}}{\pi_k} \sum_{i=1}^{t} \sum_{j=t+1}^{T} \frac{f_{i,t}^{(k)} h_{i,t}^{(k)} f_{t+1,j}^{(k)} h_{t+1,j}^{(k)}}{f_{0,0}^{(k)} h_{0,0}^{(k)}} \xi_{i,t}^{(k)} \eta_{t+1,j}^{(k)}
\end{aligned}$$

where we use (2.8), (2.11), (2.14) in forward filtering in seciton 2.2 and equation (2.55) in backward prediction in section 2.4. It is proven in Theorem 4 of Appendix A that

$$\frac{f_{i,t}^{(k)} f_{t+1,j}^{(k)}}{f_{0,0}^{(k)}} \cdot \frac{h_{i,t}^{(k)} h_{t+1,j}^{(k)}}{h_{0,0}^{(k)}} = \frac{\phi_{0,0}^{(k)} \phi_{i,j}^{(k)}}{\phi_{i,t}^{(k)} \phi_{t+1,j}^{(k)}} f_{i,j}^{(k)} h_{i,j}^{(k)} \tag{2.60}$$

Interested readers are encouraged to go through the proofs as understanding the steps may help to comprehend many derivation in this chapter. Finally the posterior distribution of $\boldsymbol{\beta}_t$ and τ_t given all the information at T is

$$f(\boldsymbol{\beta}_t, \tau_t|\mathcal{F}_T) \propto \sum_{k=1}^{K}\sum_{i=1}^{t}\Big(\xi_{i,t}^{(k)} \sum_{l=1,l\neq k}^{K} \frac{q_{lk}\eta_{t+1}^{(l)}}{\pi_k} f_{i,t}^{(k)} h_{i,t}^{(k)} \\ + \frac{q_{kk}}{\pi_k} \sum_{j=t+1}^{T} \frac{\phi_{0,0}^{(k)}\phi_{i,j}^{(k)}}{\phi_{i,t}^{(k)}\phi_{t+1,j}^{(k)}} \xi_{i,t}^{(k)} \eta_{t+1,j}^{(k)} f_{i,j}^{(k)} h_{i,j}^{(k)}\Big) \quad (2.61)$$

From the above equation, if we define a recursive formula for working smooth weight as

$$\alpha_{itj}^{(k)*} := \begin{cases} \xi_{i,t}^{(k)} \sum_{l\neq k} \frac{q_{lk}\eta_{t+1}^{(l)}}{\pi_k} & i \leq t = j \\ \frac{\xi_{i,t}^{(k)} q_{kk}\eta_{t+1,j}^{(k)}}{\pi_k} \cdot \frac{\phi_{0,0}^{(k)}\phi_{i,j}^{(k)}}{\phi_{i,t}^{(k)}\phi_{t+1,j}^{(k)}} & i \leq t < j \leq T \end{cases} \quad (2.62)$$

and the normalized version is

$$\alpha_{itj}^{(k)} = \frac{\alpha_{itj}^{(k)*}}{\sum_{k=1}^{K}\sum_{i=1}^{t}\sum_{j=t+1}^{T}\alpha_{itj}^{(k)*}} \quad (2.63)$$

then equation (2.57) is simplified to

$$f(\boldsymbol{\beta}_t, \tau_t|\mathcal{F}_T) = \sum_{k=1}^{K}\sum_{i=1}^{t}\sum_{j=t}^{T} f_{i,j}^{(k)} h_{i,j}^{(k)} \alpha_{itj}^{(k)} \quad (2.64)$$

Similar to (2.26), (2.28), (2.46) and (2.47) in forward and backward filtering estimation in section 2.2 and 2.3, smoothing estimation of $\boldsymbol{\beta}_t$ is defined as the posterior mean, i.e.

$$\widehat{\boldsymbol{\beta}}_{t|T} := \sum_{k=1}^{K}\sum_{i=1}^{t}\sum_{j=t}^{T} \alpha_{itj}^{(k)} \boldsymbol{z}_{i,j}^{(k)} \quad (2.65)$$

and the smoothing estimation of σ_t is its posterior mean as well defined as

$$\widehat{\sigma}_{t|T} := \sum_{k=1}^{K} \sum_{i=1}^{t} \sum_{j=t}^{T} \alpha_{itj}^{(k)} (2\lambda_{i,j}^{(k)})^{-\frac{1}{2}} \frac{\Gamma\big(g_{i,j}^{(k)} - \frac{1}{2}\big)}{\Gamma\big(g_{i,j}^{(k)}\big)} \tag{2.66}$$

Their posterior variances are

$$\begin{aligned} \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t|\mathcal{F}_T} = &\sum_{k=1}^{K} \sum_{i=1}^{t} \sum_{j=t}^{T} \left(\frac{\boldsymbol{V}_{ij}^{(k)}}{2\lambda_{ij}^{(k)}\big(g_{ij}^{(k)} - 1\big)} + \boldsymbol{z}_{ij}^{(k)}(\boldsymbol{z}_{ij}^{k})' \right) \alpha_{itj}^{(k)} \\ &- \sum_{k=1}^{K} \sum_{i=1}^{t} \sum_{j=t}^{T} \boldsymbol{z}_{ij}^{(k)} \alpha_{itj}^{(k)} \left(\sum_{k=1}^{K} \sum_{i=1}^{t} \sum_{j=t}^{T} \boldsymbol{z}_{ij}^{(k)} \alpha_{itj}^{(k)} \right)' \end{aligned} \tag{2.67}$$

and

$$\begin{aligned} \Sigma_{\sigma_t|\mathcal{F}_T} = &\sum_{k=1}^{K} \sum_{i=1}^{t} \sum_{j=t}^{T} \frac{\alpha_{itj}^{(k)}}{2\lambda_{ij}^{(k)}\big(g_{ij}^{(k)} - 1\big)} \\ &- \left(\sum_{k=1}^{K} \sum_{i=1}^{t} \sum_{j=t}^{T} \alpha_{itj}^{(k)} \big(2\lambda_{ij}^{(k)}\big)^{-\frac{1}{2}} \frac{\Gamma\big(g_{ij}^{(k)} - \frac{1}{2}\big)}{\Gamma\big(g_{ij}^{(k)}\big)} \right)^2 \end{aligned} \tag{2.68}$$

Finally, the regime at each time t can be estimated by

$$E\big[I_{\{S_t=k\}}|\mathcal{F}_T\big] = \alpha_t^{(k)} = \sum_{i=1}^{t} \sum_{j=t}^{T} \alpha_{itj}^{(k)}, \quad \text{for each } k = 1, \ldots, K \tag{2.69}$$

and

$$Var\big(I_{\{S_t=k\}}|\mathcal{F}_T\big) = \alpha_t^{(k)}(1 - \alpha_t^{(k)}), \quad \text{for each } k = 1, \ldots, K \tag{2.70}$$

2.6 Estimation of Hyperparameters

In previous sections, the estimation of $\boldsymbol{\beta}_t$, τ_t and probability of the regime are all based on the assumption that the priors $\boldsymbol{z}^{(k)}$, $\boldsymbol{V}^{(k)}$, $g^{(k)}$, $\lambda^{(k)}$ and the transition probabilities p_{ij} are initially given for every $i, j, k \in 1, \ldots, K$. In this section, we discuss the issues of how to estimate these prior parameters. The parameter that is of no direct interest but essential in the model estimation is called a nuisance parameter or hyperparameter.

It turns out that the likelihood function is the byproduct in the derivation of recursive forward weight. By equation (2.15) and (2.17), for $l \neq k$ and $i = t$,

$$\xi_{i,t}^{(k)} = \frac{1}{f(y_t|\mathcal{F}_{t-1})} f(y_t|J_t^{(S_t)} = t, S_t = k, \mathcal{F}_{t-1}) \sum_{l=1, l\neq k}^{K} p_{lk}\xi_{t-1}^{(l)}.$$

Likewise by equation (2.15) and (2.19), for $l = k$ and $i < t$,

$$\xi_{i,t}^{(k)} = \frac{1}{f(y_t|\mathcal{F}_{t-1})} f(y_t|J_{t-1}^{(S_{t-1})}, S_{t-1} = k, \mathcal{F}_{t-1}) p_{kk}\xi_{i,t-1}^{(k)}.$$

Since $\sum_{k=1}^{K}\sum_{i=1}^{t}\xi_{i,t}^{(k)} = 1$ and also with equation (2.20) and (2.21), the conditional likelihood of y_t given $\mathcal{F}_{t-1}$ is

$$f(y_t|\mathcal{F}_{t-1}) = \sum_{k=1}^{K}\left(\sum_{i=1}^{t-1}\pi^{-\frac{1}{2}}\frac{\phi_{i,t}^{(k)}}{\phi_{i,t-1}^{(k)}}p_{kk}\xi_{i,t-1}^{(k)} + \pi^{-\frac{1}{2}}\frac{\phi_{t,t}^{(k)}}{\phi_{0,0}^{(k)}}\sum_{l\neq k}p_{lk}\xi_{t-1}^{(l)}\right) \tag{2.71}$$

Equation (2.71) is a function of hyperparameters $\boldsymbol{\theta} = (P, \boldsymbol{z}^{(k)}, \boldsymbol{V}^{(k)}, g^{(k)}, \lambda^{(k)}\ k =$

$1, 2, \ldots, K$), where P is a $k \times k$ Markov chain transition probability matrix, $\boldsymbol{z}^{(k)}$ is a $d \times 1$ vector, $g^{(k)}$ and $\lambda^{(k)}$ are scalars and $\boldsymbol{V}^{(k)}$ is a $d \times d$ matrix. In transition matrix P, we only estimate p_{kl} for $l \neq k$ then $p_{kk} = 1 - \sum_{l \neq k} p_{kl}$ $\forall k = 1, \ \ldots, \ K$. Stationary probability of this Markov Chain $\boldsymbol{\pi'} = (\pi_1, \pi_2, \ldots, \pi_K)$ can be calculated by (2.3) under assumption (A2), therefore excluded from the estimation. Covariance matrix $\boldsymbol{V}^{(k)}$ is symmetric and only the upper triangle of matrix needs to be estimated. The total number of hyperparameters to be estimated is $\frac{(K+d^2+d+3)K}{2}$ accordingly.

Notice that y_t's are dependent, the likelihood function of y_t is $f(y_t, t = 1, 2, \ldots, T|\boldsymbol{\theta}, \boldsymbol{x}_{1t}) \neq \prod_{t=1}^{T} f(y_t|\boldsymbol{\theta}, \boldsymbol{x}_{1t})$. But y_t's are independent given information at and before $t-1$, i.e. likelihood function

$$f(y_t, t = 1, 2, \ldots, T|\boldsymbol{\theta}, \boldsymbol{x}_{1t}) = f(y_1) \prod_{t=2}^{T} f(y_t|\mathcal{F}_{t-1}, \boldsymbol{\theta}) \tag{2.72}$$

Maximum likelihood (ML) estimation is difficult to implement, simply because its log likelihood form is very complex. The solution is to resort to Expectation-Maximization (EM) algorithm for a simpler likelihood function. In the following discussion, we show that by EM algorithm every hyperparameter has a closed form solution.

Since $\boldsymbol{\beta}_t, \tau_t$ and S_t are unobserved, we treat them as latent variables, so the complete data likelihood function can be written as

$$L_c(\boldsymbol{\theta}|y_t, \boldsymbol{\beta}_t, \tau_t, S_t, t = 1, \cdots, T) = f(y_t, \boldsymbol{\beta}_t, \tau_t, S_t, t = 1, \ldots, \ T|\boldsymbol{\theta})$$

$$
\begin{aligned}
&= f(y_t, \boldsymbol{\beta}_t, \tau_t | S_t, t = 1, \dots,\ T) P(S_t, t = 1, \dots,\ T) \\
&= \Big(\prod_{t=1}^{T} f(y_t | \boldsymbol{\beta}_t, \tau_t, S_t) f(\boldsymbol{\beta}_t | \tau_t, S_t) f(\tau_t | S_t) \Big) \prod_{t=2}^{T} P(S_t | S_{t-1}) P(S_1) \\
&= \Big(\prod_{t=1}^{T} \sum_{k=1}^{K} \mathbf{1}_{\{S_t = k\}} f(y_t | \boldsymbol{\beta}_t, \tau_t, S_t = k) f(\boldsymbol{\beta}_t | \tau_t, S_t = k) f(\tau_t | S_t = k) \Big) \cdot \\
&\quad \prod_{t=2}^{T} \sum_{k=1}^{K} \sum_{l=1}^{K} \mathbf{1}_{\{S_t = k, S_{t-1} = l\}} P(S_t = k | S_{t-1} = l) P(S_1 = l) \\
&= \prod_{t=1}^{T} \Big(\sum_{k=1}^{K} \mathbf{1}_{\{S_t = k\}} f(y_t | \boldsymbol{\beta}_t, \tau_t, S_t = k) f_{0,0}^{(k)}(\boldsymbol{\beta}_t) h_{0,0}^{(k)}(\tau_t) \Big) \cdot \\
&\prod_{t=2}^{T} \Big(\sum_{k=1}^{K} \sum_{l=1}^{K} \mathbf{1}_{\{S_t = k, S_{t-1} = l\}} p_{lk} \pi_l^{(1)} \Big)
\end{aligned}
$$

$\boldsymbol{\theta}$ is omitted for easy representation. We know that y_t's are dependent, so are $\boldsymbol{\beta}_t$'s and S_t for $t = 1, \dots, T$. $y_t | \boldsymbol{\beta}_t$'s and $\boldsymbol{\beta}_t, \tau_t | S_t$'s are independent for $t = 1, \dots, T$ respectively which yields the first product of the second line. Since $\{S_t\}$ is a Markov Chain, $P(S_t, t = 1, \dots, T) = \prod_{t=2}^{T} P(S_t | S_{t-1}) P(S_1)$. Based on the model assumption that $S_1 \neq S_0$, $P(S_1 = l) = \sum_{r \neq l}^{K} P(S_1 = l |\ S_0 = r) P(S_0 = r)$. This quantity depends on an initial state probability and initial transition probabilities which possibly differ from Markov chain transition matrix P since the probability of self transition is zero. The estimation of the initial transition probability model is not the concern of this model, so we can assume an arbitrary initial value for $P(S_1 = l)$ and let it be $\pi_l^{(1)}$. This explains the second product of the second line. Third line holds, because at each time t there is only one state of S_t and S_{t-1} for state k from $1, \dots, K$. Indicator function is used to guarantee there is only one $f(\boldsymbol{\beta}_t | S_t)$ and only one $P(S_t|S_{t-1})$ at each time t. Finally, it can be shown that $\boldsymbol{\beta}_t, \tau_t|S_t = k$ are i.i.d with density $f_{0,0}^{(k)}(\boldsymbol{\beta}_t)h_{0,0}^{(k)}(\tau_t)$ regardless of the state S_{t-1}. Then complete data log likelihood function can be written as

$$
\begin{aligned}
l_c(\boldsymbol{\theta}) &= \log\ L(\boldsymbol{\theta}|y_t, \boldsymbol{\beta}_t, \tau_t, S_t, t = 1, \cdots, T) \\
&= \sum_{t=1}^{T} \log\Big(\sum_{k=1}^{K} \mathbf{1}_{\{S_t=k\}} f(y_t|\boldsymbol{\beta}_t, \tau_t, S_t = k) f_{0,0}^{(k)}(\boldsymbol{\beta}_t) h_{0,0}^{(k)}(\tau_t) \Big) \\
&\quad + \sum_{t=2}^{T}\ \log\Big(\sum_{k=1}^{K}\sum_{l=1}^{K} \mathbf{1}_{\{S_t=k, S_{t-1}=l\}} p_{lk}\pi_l^{(1)} \Big) \\
&= \sum_{t=1}^{T} \Big\{ \sum_{k=1}^{K} \mathbf{1}_{\{S_t=k\}} \log f(y_t|\boldsymbol{\beta}_t, \tau_t, S_t = k) + \sum_{k=1}^{K} \mathbf{1}_{\{S_t=k\}} \Big(\log\ f_{0,0}^{(k)}(\boldsymbol{\beta}_t) + \log\ h_{0,0}^{(k)}(\tau_t) \Big) \Big\} \\
&\quad + \sum_{t=2}^{T}\sum_{k=1}^{K}\sum_{l=1}^{K} \mathbf{1}_{\{S_t=k,\ S_{t-1}=l\}} \big(\ \log\ p_{lk} +\ \log\ \pi_l^{(1)} \big) \\
&= \sum_{t=1}^{T}\sum_{k=1}^{K} \Big\{ \mathbf{1}_{\{S_t=k\}} \Big(-\frac{1}{2}\ \log\ \pi + \frac{1}{2}\ \log\ \tau_t - \tau_t(y_t - \boldsymbol{\beta}_t'\boldsymbol{x}_t)^2 \Big) \\
&\quad + \mathbf{1}_{\{S_t=k\}} \Big(-\frac{d}{2} \log\ \pi + \frac{d}{2}\ \log\ \tau_t - \frac{1}{2}\ \log\ |\boldsymbol{V}^{(k)}| - \tau_t(\boldsymbol{\beta}_t - \boldsymbol{z}^{(k)})'(\boldsymbol{V}^{(k)})^{-1}(\boldsymbol{\beta}_t - \boldsymbol{z}^{(k)}) \Big) \\
&\quad + \mathbf{1}_{\{S_t=k\}} \Big(-\ \log\ \Gamma\big(g^{(k)}\big) - g^{(k)}\ \log\ \big(\lambda^{(k)}\big) + \big(g^{(k)} - 1\big)\ \log\ \tau_t - \frac{\tau_t}{\lambda^{(k)}} \Big) \Big\} \\
&\quad \sum_{t=2}^{T}\sum_{k=1}^{K}\sum_{l=1}^{K} \Big(\mathbf{1}_{\{S_t=k,\ S_{t-1}=l\}}\ \log\ (p_{lk}) + \mathbf{1}_{\{S_t=k, S_{t-1}=l\}} \log\ \pi_l^{(1)} \Big)
\end{aligned}
$$

2.6.1 Expectation Step

Take expectation of the above complete data log likelihood function with respect to latent variables $\boldsymbol{\beta}_t$, S_t and τ_t given hyperparameters at previous iteration and all observations of y_t and x_t for $t = 1, \dots, T$.

$$
\begin{aligned}
&E\big(l_c(\boldsymbol{\theta})|\mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}\big) \\
&= \sum_{t=1}^{T}\sum_{k=1}^{K}\bigg\{ -\frac{1+d}{2}(\ \log\,\pi)E\big[\mathbf{1}_{\{S_t=k\}}|\mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}\big] + \frac{1+d}{2}E\big[\mathbf{1}_{\{S_t=k\}}(\ \log\,\tau_t)|\mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}\big] \\
&\quad - E\big[\mathbf{1}_{\{S_t=k\}}\tau_t(y_t - \boldsymbol{\beta}_t'\boldsymbol{x}_t)^2|\mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}\big] - \frac{1}{2}\text{log}|\boldsymbol{V}^{(k)}|E\big[\mathbf{1}_{\{S_t=k\}}|\mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}\big] \\
&\quad - E\big[\mathbf{1}_{\{S_t=k\}}\tau_t(\boldsymbol{\beta}_t - \boldsymbol{z}^{(k)})'(\boldsymbol{V}^{(k)})^{-1}(\boldsymbol{\beta}_t - \boldsymbol{z}^{(k)})|\mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}\big] \\
&\quad - \Big(\ \log\,\Gamma\big(g^{(k)}\big) + g^{(k)}\ \log\,\big(\lambda^{(k)}\big)\Big)E\big[\mathbf{1}_{\{S_t=k\}}|\mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}\big] \\
&\quad + \big(g^{(k)} - 1\big)E\big[\mathbf{1}_{\{S_t=k\}}(\ \log\,\tau_t)|\mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}\big] - \frac{1}{\lambda^{(k)}}E\big[\mathbf{1}_{\{S_t=k\}}(\tau_t)|\mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}\big]\bigg\} \\
&\quad + \sum_{t=2}^{T}\sum_{k=1}^{K}\sum_{l=1}^{K}\big(\ \text{log}p_{lk} + \text{log}\pi_l^{(1)}\big)E\big[\mathbf{1}_{\{S_t=k,\ S_{t-1}=l\}}|\mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}\big] \qquad (2.73)
\end{aligned}
$$

Continue to simplify (2.73), we need to compute the following six quantities: $E\big[\mathbf{1}_{\{S_t=k\}}|\mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}\big]$, $E\big[\mathbf{1}_{\{S_t=k,\ S_{t-1}=l\}}|\mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}\big]$, $E\big[\mathbf{1}_{\{S_t=k\}}\text{log}\tau_t|\mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}\big]$, $E\big[\mathbf{1}_{\{S_t=k\}}\tau_t|\mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}\big]$, $E\big[\mathbf{1}_{\{S_t=k\}}\tau_t(y_t - \boldsymbol{\beta}_t'\boldsymbol{x}_t)^2|\mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}\big]$, and $E\big[\mathbf{1}_{\{S_t=k\}}\tau_t(\boldsymbol{\beta}_t - \boldsymbol{z}^{(k)})'(\boldsymbol{V}^{(k)})^{-1}(\boldsymbol{\beta}_t - \boldsymbol{z}^{(k)})|\mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}\big]$. By equation (2.59),

$$
E\big[\mathbf{1}_{\{S_t=k\}}|\mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}\big] = P(S_t = k|\mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}) = \sum_{i=1}^{t}\sum_{j=t}^{T}\alpha_{itj}^{(k)} = \alpha_t^{(k)}. \qquad (2.74)
$$

$E\big[\mathbf{1}_{\{S_t=k,\ S_{t-1}=l\}}|\mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}\big] = P(S_t = k, S_{t-1} = l|\mathcal{F}_T, \boldsymbol{\theta}_{\text{old}})$. It is proved in Theorem 5 in Appendix A, that

$$P(S_t = k, S_{t-1} = l|\mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}) = \begin{cases} \frac{\phi_{t,t}^{(k)}\xi_{t-1}^{(l)}}{\phi_{0,0}^{(k)}} \frac{p_{lk}}{\pi_k} \sum_{r=1}^{K} q_{rk}\eta_{t+1}^{(r)} \Big/ A_t & k \neq l \\ \sum_{i=1}^{t-1} \frac{\phi_{i,t}^{(k)}\xi_{i,t-1}^{(k)}}{\phi_{i,t-1}^{(k)}} \frac{p_{kk}}{\pi_k} \sum_{r=1}^{K} q_{rk}\eta_{t+1}^{(r)} \Big/ A_t & k = l \end{cases} \tag{2.75}$$

where $A_t = \sum_{l=1}^{K} \Big(\sum_{k\neq l}^{K} \frac{\phi_{t,t}^{(k)}\xi_{t-1}^{(l)}}{\phi_{0,0}^{(k)}} \frac{p_{lk}}{\pi_k} \sum_{r=1}^{K} q_{rk}\eta_{t+1}^{(r)} + \sum_{i=1}^{t-1} \frac{\phi_{i,t}^{(k)}\xi_{i,t-1}^{(k)}}{\phi_{i,t-1}^{(k)}} \frac{p_{kk}}{\pi_k} \sum_{r=1}^{K} q_{rk}\eta_{t+1}^{(r)} \Big)$. This posterior of joint states is a special case of a general framework shown in Theorem 6 in Appendix A. Since conditional distribution of τ_t given $S_t = k, J_t^{(k)} = i, R_t^{(k)} = j, \mathcal{F}_T$ is Gamma $\big(g_{i,j}^{(k)}, \lambda_{i,j}^{(k)}\big)$ from equation (2.64), by Theorem 8 and using the notation $\psi(x)$ referring to $\frac{\Gamma'(x)}{\Gamma(x)}$, we have

$$\begin{aligned} &E\big[\mathbf{1}_{\{S_t=k\}}(\log\,\tau_t)|\mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}\big] \\ &= \sum_{i=1}^{t}\sum_{j=t}^{T} P(J_t^{(S_t)} = i, R_t^{(S_t)} = j, S_t = k|\mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}) E\big[\log\tau_t|S_t = k, J_t^{(S_t)} = i, R_t^{(S_t)} = j, \mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}\big] \\ &= \sum_{i=1}^{t}\sum_{j=t}^{T} \alpha_{itj}^{(k)} \Big(\frac{\mathrm{d}}{\mathrm{d}g_{ij}^{(k)}}\,\log\,\Gamma(g_{ij}^{(k)}) + \log\,\lambda_{ij}^{(k)} \Big) \\ &= \sum_{i=1}^{t}\sum_{j=t}^{T} \alpha_{itj}^{(k)} \Big(\psi(g_{ij}^{(k)}) + \log\,\lambda_{ij}^{(k)} \Big) \end{aligned}$$

Similarly,

$$E\big[\mathbf{1}_{\{S_t=k\}}\tau_t|\mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}\big] = \sum_{i=1}^{t}\sum_{j=t}^{T} \alpha_{itj}^{(k)} g_{ij}^{(k)} \lambda_{ij}^{(k)}.$$

Next,

$$E\big[\mathbf{1}_{\{S_t=k\}}\tau_t(y_t - \boldsymbol{\beta}_t'\boldsymbol{x}_t)^2|\mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}\big]$$

$$
\begin{aligned}
&= \sum_{i=1}^{t}\sum_{j=t}^{T} P(S_t = k, J_t^{(S_t)} = i, R_t^{(S_t)} = j|\mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}) \\
&\quad \cdot E\big[\tau_t(y_t - \boldsymbol{\beta}'_t\boldsymbol{x}_t)^2|S_t = k, J_t^{(S_t)} = i, R_t^{(S_t)} = j, \mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}\big] \\
&= \sum_{i=1}^{t}\sum_{j=t}^{T} \alpha_{itj}^{(k)} E\Big[E\big[(y_t - \boldsymbol{\beta}'_t\boldsymbol{x}_t)^2|\tau_t, S_t = k, J_t^{(S_t)} = i, R_t^{(S_t)} = j, \mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}\big] \\
&\quad \cdot \tau_t\big|S_t = k, J_t^{(S_t)} = i, R_t^{(S_t)} = j, \mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}\Big] \\
&= \sum_{i=1}^{t}\sum_{j=t}^{T} \alpha_{itj}^{(k)} E\Big[E\big[y_t^2 - 2\boldsymbol{\beta}'_t\boldsymbol{x}_t y_t + \boldsymbol{x}'_t\boldsymbol{\beta}_t\boldsymbol{\beta}'_t\boldsymbol{x}_t|\tau_t, S_t = k, J_t^{(S_t)} = i, R_t^{(S_t)} = j, \mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}\big] \\
&\quad \cdot \tau_t\Big|S_t = k, J_t^{(S_t)} = i, R_t^{(S_t)} = j, \mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}\Big]
\end{aligned}
$$

In additional to conditional distribution of τ_t, $\boldsymbol{\beta}_t\big|\tau_t, S_t = k, J_t^{(S_t)} = i, R_t^{(S_t)} = j, \mathcal{F}_T, \boldsymbol{\theta}_{\text{old}} \sim \mathrm{N}\Big(\boldsymbol{z}_{i,j}^{(k)}, \frac{\boldsymbol{V}_{i,j}^{(k)}}{2\tau_t}\Big)$ by (2.64),

$$
E\big[\boldsymbol{\beta}_t|\tau_t, S_t = k, J_t^{(S_t)} = i, R_t^{(S_t)} = j, \mathcal{F}_t, \boldsymbol{\theta}_{\text{old}}\big] = \boldsymbol{z}_{i,j}^{(k)},
$$

$$
E\big[\boldsymbol{\beta}_t\boldsymbol{\beta}'_t|S_t = k, J_t^{(S_t)} = i, R_t^{(S_t)} = j, \mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}\big] = \frac{\boldsymbol{V}_{i,j}^{(k)}}{2\tau_t} + \boldsymbol{z}_{i,j}^{(k)}\big(\boldsymbol{z}_{i,j}^{(k)}\big)'
$$

and

$$
E\big[\tau_t\big|S_t = k, J_t^{(S_t)} = i, R_t^{(S_t)} = j, \mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}\big] = g_{ij}^{(k)}\lambda_{i,j}^{(k)}
$$

Then,

$$
\begin{aligned}
&E\big[\mathbf{1}_{\{S_t=k\}}\tau_t(y_t - \boldsymbol{\beta}'_t\boldsymbol{x}_t)^2|\mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}\big] \\
&\quad = \sum_{i=1}^{t}\sum_{j=t}^{T} \alpha_{itj}^{(k)}\Big(\big(y_t^2 - 2\boldsymbol{z}_{i,j}^{(k)}{}'\boldsymbol{x}_t y_t + \boldsymbol{x}'_t\boldsymbol{z}_{i,j}^{(k)}\boldsymbol{z}_{i,j}^{(k)}{}'\boldsymbol{x}_t\big)g_{ij}^{(k)}\lambda_{ij}^{(k)} + \frac{1}{2}\boldsymbol{x}'_t\boldsymbol{V}_{i,j}^{(k)}\boldsymbol{x}_t\Big)
\end{aligned}
$$

Finally, by properties of matrix expectation in Theorem 9 in Appendix A,

$$
\begin{aligned}
&E\big[\mathbf{1}_{\{S_t=k\}}\tau_t(\boldsymbol{\beta}_t-\boldsymbol{z}^{(k)})'\big(\boldsymbol{V}^{(k)}\big)^{-1}(\boldsymbol{\beta}_t-\boldsymbol{z}^{(k)})|\mathcal{F}_T,\boldsymbol{\theta}_{\text{old}}\big]\\
&=\sum_{i=1}^{t}\sum_{j=t}^{T}P(S_t=k,J_t^{(S_t)}=i,R_t^{(S_t)}=j|\mathcal{F}_T,\boldsymbol{\theta}_{\text{old}})\\
&\quad\cdot E\big[\tau_t(\boldsymbol{\beta}_t-\boldsymbol{z}^{(k)})'\big(\boldsymbol{V}^{(k)}\big)^{-1}(\boldsymbol{\beta}_t-\boldsymbol{z}^{(k)})|S_t=k,J_t^{(S_t)}=i,R_t^{(S_t)}=j,\mathcal{F}_T,\boldsymbol{\theta}_{\text{old}}\big]\\
&=\sum_{i=1}^{t}\sum_{j=t}^{T}\alpha_{itj}^{(k)}E\Big[E\Big[\Big(\boldsymbol{\beta}_t'\big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{\beta}_t-\boldsymbol{\beta}_t'\big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}^{(k)}-\boldsymbol{z}^{(k)\prime}\big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{\beta}_t+\boldsymbol{z}^{(k)\prime}\big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}^{(k)}\Big)\Big|\\
&\quad\tau_t,S_t=k,J_t^{(S_t)}=i,R_t^{(S_t)}=j,\mathcal{F}_T,\boldsymbol{\theta}_{\text{old}}\Big]\cdot\tau_t\big|S_t=k,J_t^{(S_t)}=i,R_t^{(S_t)}=j,\mathcal{F}_T,\boldsymbol{\theta}_{\text{old}}\Big]\\
&=\sum_{i=1}^{t}\sum_{j=t}^{T}\alpha_{itj}^{(k)}E\Big[\tau_t\text{tr}\big((\boldsymbol{V}^{(k)})^{-1}\boldsymbol{V}_{i,j}^{(k)}/2\tau_t\big)+\tau_t\boldsymbol{z}_{i,j}^{(k)\prime}(\boldsymbol{V}^{(k)})^{-1}\boldsymbol{z}_{i,j}^{(k)}-\tau_t\boldsymbol{z}_{i,j}^{(k)\prime}\big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}^{(k)}\\
&\quad-\tau_t\boldsymbol{z}^{(k)\prime}\big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}_{ij}^{(k)}+\tau_t\boldsymbol{z}^{(k)\prime}\big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}^{(k)}\big|S_t=k,J_t^{(S_t)}=i,R_t^{(S_t)}=j,\mathcal{F}_T,\boldsymbol{\theta}_{\text{old}}\Big]\\
&=\sum_{i=1}^{t}\sum_{j=t}^{T}\alpha_{itj}^{(k)}\Big(\text{tr}\big((\boldsymbol{V}^{(k)})^{-1}\boldsymbol{V}_{i,j}^{(k)}/2\big)+g_{ij}^{(k)}\lambda_{ij}^{(k)}\boldsymbol{z}_{i,j}^{(k)\prime}(\boldsymbol{V}^{(k)})^{-1}\boldsymbol{z}_{i,j}^{(k)}-g_{ij}^{(k)}\lambda_{ij}^{(k)}\boldsymbol{z}_{i,j}^{(k)\prime}\big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}^{(k)}\\
&\quad-g_{ij}^{(k)}\lambda_{ij}^{(k)}\boldsymbol{z}^{(k)\prime}\big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}_{ij}^{(k)}+g_{ij}^{(k)}\lambda_{ij}^{(k)}\boldsymbol{z}^{(k)\prime}\big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}^{(k)}\Big)
\end{aligned}
\tag{2.76}
$$

where

$$
\begin{aligned}
E\big[\boldsymbol{\beta}_t'\big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{\beta}_t|S_t=k,J_t^{(S_t)}=i,R_t^{(S_t)}=j,\mathcal{F}_T,\boldsymbol{\theta}_{old}\big]&\\
=\text{tr}\big((\boldsymbol{V}^{(k)})^{-1}\boldsymbol{V}_{i,j}^{(k)}/2\tau_t\big)+\boldsymbol{z}_{i,j}^{(k)\prime}(\boldsymbol{V}^{(k)})^{-1}\boldsymbol{z}_{i,j}^{(k)}&.
\end{aligned}
$$

2.6.2 Maximization Step

Find (p_{lk}), for $l \neq k$ and $\boldsymbol{z}^{(k)}$, $\boldsymbol{V}^{(k)}, g^{(k)}$ and $\lambda^{(k)}$ that maximize the expected complete data log likelihood function (2.73) in E-step. The only term related to p_{lk} is $E\big[\mathbf{1}_{\{S_t=k,S_{t-1}=l\}}|\mathcal{F}_T, \boldsymbol{\theta}_{old}\big]$, then

$$\frac{\partial E\big[l_c(\boldsymbol{\theta})|\mathcal{F}_T, \boldsymbol{\theta}_{old}\big]}{\partial p_{lk}} = \frac{\partial\Bigg(\sum_{t=2}^{T}\sum_{k=1}^{K}\sum_{l=1}^{K} P(S_t = k, S_{t-1} = l|\mathcal{F}_T, \boldsymbol{\theta}_{old}) \log p_{lk}\Bigg)}{\partial p_{lk}}$$

To find p_{lk} that maximize $f(\cdot)$ along with constraint $\sum_{k=1}^{K} p_{lk} = 1$, we consider using Lagrange multiplier. The constraint is denoted by $g(\cdot) = \sum_{k=1}^{K} p_{lk} = 1$. In this technique, if there is a function $f(\cdot)$ subject to a constraint $g(\cdot) =$ constant, introduce parameter λ so that $\bigtriangledown f(\cdot) = \lambda \bigtriangledown g(\cdot)$. Gradient is taken with respect to variables such as p_{lk} in this case and λ. Solve p_{lk}'s and λ

$$\sum_{t=2}^{T} P(S_t = k, S_{t-1} = l\big|\mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}) = \lambda p_{lk} \quad \forall l, k = 1, 2, \ldots, K$$

Therefore

$$\lambda = \sum_{r=1}^{K}\sum_{t=2}^{T} P(S_t = r, S_{t-1} = l\big|\mathcal{F}_T, \boldsymbol{\theta}_{\text{old}})$$

and

$$p_{lk} = \frac{\sum_{t=2}^{T} P(S_t = k, S_{t-1} = l\big|\mathcal{F}_T, \boldsymbol{\theta}_{\text{old}})}{\sum_{r=1}^{K}\sum_{t=2}^{T} P(S_t = r, S_{t-1} = l\big|\mathcal{F}_T, \boldsymbol{\theta}_{\text{old}})} \quad \forall l, k = 1, 2, \ldots, K$$

The term related to $\boldsymbol{z}^{(k)}$ in (2.73) is $\sum_{t=1}^{T}\sum_{k=1}^{K} E\big[\mathbf{1}_{\{S_t=k\}}\tau_t(\boldsymbol{\beta}_t-\boldsymbol{z}^{(k)})'(\boldsymbol{V}^{(k)})^{-1}(\boldsymbol{\beta}_t-\boldsymbol{z}^{(k)})|\mathcal{F}_T,\boldsymbol{\theta}_{\text{old}}\big]$ and with expression (2.76),

$$
\begin{aligned}
&\frac{\partial E\big[l_c(\boldsymbol{\theta})|\mathcal{F}_T,\boldsymbol{\theta}_{\text{old}}\big]}{\partial \boldsymbol{z}^{(k)}}\\
&\quad=\sum_{t=1}^{T}\sum_{i=1}^{t}\sum_{j=t}^{T}\alpha_{itj}^{(k)}\Big(-2g_{ij}^{(k)}\lambda_{ij}^{(k)}\big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}_{i,j}^{(k)}+2g_{ij}^{(k)}\lambda_{ij}^{(k)}\big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}^{(k)}\Big)
\end{aligned}
$$

Setting the above equation to 0, then

$$
\begin{aligned}
&\sum_{t=1}^{T}\sum_{i=1}^{t}\sum_{j=t}^{T}\alpha_{itj}^{(k)}g_{ij}^{(k)}\lambda_{ij}^{(k)}\Big(\big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}_{i,j}^{(k)}-\big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}^{(k)}\Big)=0\\
&\Longrightarrow\sum_{t=1}^{T}\sum_{i=1}^{t}\sum_{j=t}^{T}\alpha_{itj}^{(k)}g_{ij}^{(k)}\lambda_{ij}^{(k)}\Big(\boldsymbol{z}_{i,j}^{(k)}-\boldsymbol{z}^{(k)}\Big)=0\\
&\Longrightarrow\boldsymbol{z}^{(k)}=\frac{\sum_{t=1}^{T}\sum_{i=1}^{t}\sum_{j=t}^{T}\alpha_{itj}^{(k)}g_{ij}^{(k)}\lambda_{ij}^{(k)}\boldsymbol{z}_{i,j}^{(k)}}{\sum_{t=1}^{T}\sum_{i=1}^{t}\sum_{j=t}^{T}\alpha_{itj}^{(k)}g_{ij}^{(k)}\lambda_{ij}^{(k)}}
\end{aligned}
$$

The term related to $\boldsymbol{V}^{(k)}$ in (2.73) is

$$
\begin{aligned}
&\sum_{t=1}^{T}\sum_{k=1}^{K}\frac{1}{2}\log\,|\boldsymbol{V}^{(k)-1}|E[\mathbf{1}_{\{S_t=k\}}|\mathcal{F}_T,\boldsymbol{\theta}_{\text{old}}]\\
&-\sum_{t=1}^{T}\sum_{k=1}^{K}E\big[\mathbf{1}_{\{S_t=k\}}\tau_t(\boldsymbol{\beta}_t-\boldsymbol{z}^{(k)})'(\boldsymbol{V}^{(k)})^{-1}(\boldsymbol{\beta}_t-\boldsymbol{z}^{(k)})|\mathcal{F}_T,\boldsymbol{\theta}_{\text{old}}\big]\\
&=\sum_{t=1}^{T}\sum_{k=1}^{K}\frac{1}{2}\log\,|\boldsymbol{V}^{(k)-1}|\alpha_t^{(k)}\\
&\quad-\sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{i=1}^{t}\sum_{j=t}^{T}\alpha_{itj}^{(k)}\bigg(\text{tr}\big((\boldsymbol{V}^{(k)})^{-1}\boldsymbol{V}_{i,j}^{(k)}/2\big)+g_{ij}^{(k)}\lambda_{ij}^{(k)}\boldsymbol{z}_{i,j}^{(k)\prime}(\boldsymbol{V}^{(k)})^{-1}\boldsymbol{z}_{i,j}^{(k)}
\end{aligned}
$$

$$- g_{ij}^{(k)}\lambda_{ij}^{(k)}\boldsymbol{z}_{i,j}^{(k)\prime}\big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}^{(k)} - g_{ij}^{(k)}\lambda_{ij}^{(k)}\boldsymbol{z}^{(k)\prime}\big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}_{ij}^{(k)} + g_{ij}^{(k)}\lambda_{ij}^{(k)}\boldsymbol{z}^{(k)\prime}\big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}^{(k)}\Big)$$

By Theorem 10, the partial derivative with respect to a vector $\boldsymbol{V}^{(k)-1}$ is then

$$\frac{\partial E\big[l_c(\boldsymbol{\theta})|\mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}\big]}{\partial\big(\boldsymbol{V}^{(k)}\big)^{-1}}$$

$$= \frac{1}{2}\big(2\boldsymbol{V}^{(k)} - \text{diag }(\boldsymbol{V}^{(k)})\big)\sum_{t=1}^{T}\alpha_t^{(k)} - \sum_{t=1}^{T}\sum_{i=1}^{t}\sum_{j=t}^{T}\alpha_{itj}^{(k)}\Big(\boldsymbol{V}_{ij}^{(k)} - \text{diag }(\boldsymbol{V}_{i,j}^{(k)}/2)$$

$$+ g_{ij}^{(k)}\lambda_{ij}^{(k)}\boldsymbol{z}_{i,j}^{(k)}\boldsymbol{z}_{i,j}^{(k)\prime} - g_{ij}^{(k)}\lambda_{ij}^{(k)}\boldsymbol{z}_{i,j}^{(k)}\boldsymbol{z}^{(k)\prime} - g_{ij}^{(k)}\lambda_{ij}^{(k)}\boldsymbol{z}^{(k)}\boldsymbol{z}_{ij}^{(k)\prime} + g_{ij}^{(k)}\lambda_{ij}^{(k)}\boldsymbol{z}^{(k)}\boldsymbol{z}^{(k)\prime}\Big)$$

Set the above expression to zero, thus

$$\boldsymbol{V}^{(k)} - \text{diag }(\boldsymbol{V}^{(k)}/2) = \frac{1}{\sum_{t=1}^{T}\alpha_t^{(k)}}\sum_{t=1}^{T}\sum_{i=1}^{t}\sum_{j=t}^{T}\alpha_{itj}^{(k)}\Big(\boldsymbol{V}_{ij}^{(k)} - \text{diag }(\boldsymbol{V}_{i,j}^{(k)}/2)$$

$$+ g_{ij}^{(k)}\lambda_{ij}^{(k)}\boldsymbol{z}_{i,j}^{(k)}\boldsymbol{z}_{i,j}^{(k)\prime} - g_{ij}^{(k)}\lambda_{ij}^{(k)}\boldsymbol{z}_{i,j}^{(k)}\boldsymbol{z}^{(k)\prime} - g_{ij}^{(k)}\lambda_{ij}^{(k)}\boldsymbol{z}^{(k)}\boldsymbol{z}_{ij}^{(k)\prime} + g_{ij}^{(k)}\lambda_{ij}^{(k)}\boldsymbol{z}^{(k)}\boldsymbol{z}^{(k)\prime}\Big)$$

The term relating to $g^{(k)}$ in (2.73) is

$$-\sum_{t=1}^{T}\sum_{k=1}^{K}\Big(\log\Gamma\big(g^{(k)}\big) + \big(g^{(k)}\big)\log\big(\lambda^{(k)}\big)\Big)\alpha_t^{(k)}$$

$$+\sum_{t=1}^{T}\sum_{k=1}^{K}\big(g^{(k)} - 1\big)E\big[\mathbf{1}_{\{S_t=k\}}(\log\tau_t)|\mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}\big]$$

Taking first derivative with respective to $g^{(k)}$ yields the following equation.

$$\frac{\partial E\big[l_c(\boldsymbol{\theta})|\mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}\big]}{\partial g^{(k)}}$$

$$= \sum_{t=1}^{T} \Big(-\psi(g^{(k)}) - \log\, (\lambda^{(k)}) \Big) \alpha_t^{(k)} + \sum_{t=1}^{T} E\big[\mathbf{1}_{\{S_t=k\}}\ \log\, \tau_t | \mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}\big]$$

$$= \sum_{t=1}^{T} \Big(-\psi(g^{(k)}) - \log\, (\lambda^{(k)}) \Big) \alpha_t^{(k)} + \sum_{t=1}^{T} \sum_{i=1}^{t} \sum_{j=t}^{T} \alpha_{itj}^{(k)} \Big(\psi\big(g_{ij}^{(k)}\big) +\ \log\, \big(\lambda_{ij}^{(k)}\big) \Big)$$

Set the above equation to zero, then

$$\psi(g^{(k)}) +\ \log\, \big(\lambda^{(k)}\big) = \frac{1}{\sum_{t=1}^{T} \alpha_t^{(k)}} \sum_{t=1}^{T} \sum_{i=1}^{t} \sum_{j=t}^{T} \alpha_{itj}^{(k)} \Big(\psi\big(g_{ij}^{(k)}\big) +\ \log\, \big(\lambda_{ij}^{(k)}\big) \Big)$$

Finally, find $\lambda^{(k)}$ that maximize (2.73).

$$\begin{aligned} &\frac{\partial E\big[l_c(\boldsymbol{\theta}) | \mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}\big]}{\partial \lambda^{(k)}} \\ &= -\frac{g^{(k)}}{\lambda^{(k)}} \sum_{t=1}^{T} \alpha_t^{(k)} + \frac{1}{\lambda^{(k)2}} \sum_{t=1}^{T} E\big[\mathbf{1}_{\{S_t=k\}}(\tau_t) | \mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}\big] \\ &= -\frac{g^{(k)}}{\lambda^{(k)}} \sum_{t=1}^{T} \alpha_t^{(k)} + \frac{1}{\lambda^{(k)2}} \sum_{t=1}^{T} \sum_{i=1}^{t} \sum_{j=t}^{T} \alpha_{itj}^{(k)} g_{ij}^{(k)} \lambda_{ij}^{k} \end{aligned}$$

Set the above equation to zero and solve for $\lambda^{(k)}$

$$\begin{aligned} &\lambda^{(k)} = \frac{\sum_{t=1}^{T} \sum_{i=1}^{t} \sum_{j=t}^{T} \alpha_{itj}^{(k)} g_{ij}^{(k)} \lambda_{ij}^{(k)}}{g^{(k)} \sum_{t=1}^{T} \alpha_t^{(k)}} \\ &\Longrightarrow \log\, \lambda^{(k)} =\ \log\, \Big(\sum_{t=1}^{T} \sum_{i=1}^{t} \sum_{j=t}^{T} \alpha_{itj}^{(k)} g_{ij}^{(k)} \lambda_{ij}^{(k)} \Big) -\ \log\, \big(g^{(k)}\big) -\ \log\, \Big(\sum_{t=1}^{T} \alpha_t^{(k)} \Big) \end{aligned}$$

Plug $\log\, \lambda^{(k)}$ into the equation that estimates $g^{(k)}$ and solve for $g^{(k)}$.

$$\psi(g^{(k)}) - \ \log\ \big(g^{(k)}\big) = \frac{1}{\sum_{t=1}^{T} \alpha_t^{(k)}} \sum_{t=1}^{T} \sum_{i=1}^{t} \sum_{j=t}^{T} \alpha_{itj}^{(k)} \Bigg(\psi\big(g_{ij}^{(k)}\big) + \ \log\ \big(\lambda_{ij}^{(k)}\big) \Bigg)$$

$$- \ \log\ \big(\sum_{t=1}^{T} \sum_{i=1}^{t} \sum_{j=t}^{T} \alpha_{itj}^{(k)} g_{ij}^{(k)} \lambda_{ij}^{(k)} \big) + \ \log\ \big(\sum_{t=1}^{T} \alpha_t^{(k)} \big)$$

Estimation of hyperparameters all has closed form solutions. Given the initial prior values, hyperparameters can be updated by the following formula which are a summary of M-step.

$$p_{lk,\text{new}} = \frac{\sum_{t=2}^{T} P(S_t = k, S_{t-1} = l | \mathcal{F}_T, \boldsymbol{\theta}_{\text{old}})}{\sum_{r=1}^{K} \sum_{t=2}^{T} P(S_t = r, S_{t-1} = l | \mathcal{F}_T, \boldsymbol{\theta}_{\text{old}})} \quad \forall l, k \in 1, 2, \ldots, K \tag{2.77}$$

$$\boldsymbol{z}_{\text{new}}^{(k)} = \frac{\sum_{t=1}^{T} \sum_{i=1}^{t} \sum_{j=t}^{T} \alpha_{itj}^{(k)} g_{ij}^{(k)} \lambda_{ij}^{(k)} \boldsymbol{z}_{i,j}^{(k)}}{\sum_{t=1}^{T} \sum_{i=1}^{t} \sum_{j=t}^{T} \alpha_{itj}^{(k)} g_{ij}^{(k)} \lambda_{ij}^{(k)}} \tag{2.78}$$

$$\boldsymbol{V}_{\text{new}}^{(k)} - \ \text{diag}\ (\boldsymbol{V}_{\text{new}}^{(k)}/2) = \frac{1}{\sum_{t=1}^{T} \alpha_t^{(k)}} \sum_{t=1}^{T} \sum_{i=1}^{t} \sum_{j=t}^{T} \alpha_{itj}^{(k)} \Big(\boldsymbol{V}_{i,j}^{(k)} - \ \text{diag}\ (\boldsymbol{V}_{i,j}^{(k)}/2) +$$

$$g_{ij}^{(k)} \lambda_{ij}^{(k)} \boldsymbol{z}_{i,j}^{(k)} \boldsymbol{z}_{i,j}^{(k)\prime} - g_{ij}^{(k)} \lambda_{ij}^{(k)} \boldsymbol{z}_{i,j}^{(k)} \boldsymbol{z}_{\text{new}}^{(k)\ \prime} - g_{ij}^{(k)} \lambda_{ij}^{(k)} \boldsymbol{z}_{\text{new}}^{(k)} \boldsymbol{z}_{ij}^{(k)\prime} + g_{ij}^{(k)} \lambda_{ij}^{(k)} \boldsymbol{z}_{\text{new}}^{(k)} \boldsymbol{z}_{\text{new}}^{(k)\ \prime} \Big) \tag{2.79}$$

$$\psi(g_{\text{new}}^{(k)}) - \ \log\ \big(g_{\text{new}}^{(k)}\big) = \frac{1}{\sum_{t=1}^{T} \alpha_t^{(k)}} \sum_{t=1}^{T} \sum_{i=1}^{t} \sum_{j=t}^{T} \alpha_{itj}^{(k)} \Bigg(\psi\big(g_{ij}^{(k)}\big) + \ \log\ \big(\lambda_{ij}^{(k)}\big) \Bigg)$$

$$- \ \log\ \big(\sum_{t=1}^{T} \sum_{i=1}^{t} \sum_{j=t}^{T} \alpha_{itj}^{(k)} g_{ij}^{(k)} \lambda_{ij}^{(k)} \big) + \ \log\ \big(\sum_{t=1}^{T} \alpha_t^{(k)} \big) \tag{2.80}$$

$$\lambda_{\text{new}}^{(k)} = \frac{1}{g_{\text{new}}^{(k)} \sum_{t=1}^{T} \alpha_t^{(k)}} \sum_{t=1}^{T} \sum_{i=1}^{t} \sum_{j=t}^{T} \alpha_{itj}^{(k)} g_{ij}^{(k)} \lambda_{ij}^{(k)} \tag{2.81}$$

2.7 Bounded Complexity Mixture (BCMIX) Approximation

The estimation of regression parameters and hyperparameters in Section 2.2 - 2.6 heavily relies on the filtering recursive weights $\xi_{i,t}^{(k)}$, $\eta_{t,j}^{(l)}$ and a derived smoothing weight $\alpha_{itj}^{(r)}$. Since every of the indice i, j, t must go through time 1 to T, these weights need to be computed in polynomial time and require roughly T^2 or T^3 memory space. Regression parameters estimated based on these recursive weights are called Bayes estimators. To increase the computational efficiency, we consider an approximation procedure with lower order computational complexity, yet comparable to the Bayes estimates in statistical efficiency. These procedures are discussed in earlier works by Lai et al. (2005), Lai et al. (2008) and Lai and Xing (2011) are modified to adapt to the models in this thesis.

For forward filtering weights, only a fixed number M of weights are kept at every time t and among M weights, m of most recent weights with respect to the evaluation time t are preserved. Usually m is between 1 (inclusive) and M (exclusive). The remaining $M - m$ weights are the largest weights before $m + 1$. Concretely, let $\mathcal{K}_{t-1}^{(k)}$ be the set of indices for which $\xi_{i,t-1}^{(k)}$ in (2.25) is kept at stage $t - 1$ for regime k. There are M elements in set

$\mathcal{K}_{t-1}^{(k)}$ including most recent index set $\{t-m, t-m+1, \ldots, t-1\}$. When a new observation at time t arrives, the new index set denoted by $\mathcal{K}_t^{(k)}$ is updated to include index t and thus the most recent m indice in $\mathcal{K}_t^{(k)}$ are $\{t-m+1, t-m+2, \ldots, t-1, t\}$. The remaining $M-m$ indices are selected by exclusion. Let r be an index in $\mathcal{K}_{t-1}^{(k)}$ with $r \leq t-m$ where $\xi_{r,t}^{(k)}$ is the minimum, i.e.

$$\arg\min_{r} \{\xi_{r,t}^{(k)} | r \in \mathcal{K}_{t-1}^{(k)} \text{ and } r \leq t-m\} \tag{2.82}$$

If there are more than one element in (2.82), choose the index farthest away from t. Now the updated index set is

$$\mathcal{K}_t^{(k)} = \{t\} \cup \mathcal{K}_{t-1}^{(k)} \setminus \{r\}. \tag{2.83}$$

Quantity in equation (2.25) is modified to

$$\xi_{i,t}^{(k)} \text{ where } i \in \mathcal{K}_t^{(k)} \tag{2.84}$$

In practice, we modify the quantity $\xi_{i,t}^{(k)*}$ first and normalize it to $\xi_{i,t}^{(k)}$. The estimation of regression parameters using (2.84) is called the BCMIX approximation to the forward filtering estimation.

Likewise, BCMIX approximation of backward filtering estimation of regression parameters can be easily constructed based on the modified backward recursive weights in equation (2.45). The total number of indices kept at each time t is M where m $(1 \leq m < M)$ of them are the most recent indices reading time sequence backward. Let $\widetilde{\mathcal{K}}_{t+1}^{(k)}$ be a set of indices j for which $\eta_{t+1,j}^{(k)}$ is kept at time $t+1$ for regime k. Again, $\widetilde{\mathcal{K}}_{t+1}^{(k)}$ includes $\{t+1, t+2, \ldots, t+m\}$ and other indices updated from its previous step. At time t, the updated index set is

$$\widetilde{\mathcal{K}}_{t}^{(k)} = \{t\} \cup \widetilde{\mathcal{K}}_{t+1}^{(k)} \setminus \{r\} \tag{2.85}$$

where r is the index farthest away from time t if the result is not unique from the following equation:

$$\arg\min_{r} \{\eta_{t,r}^{(k)} | r \in \widetilde{\mathcal{K}}_{t+1}^{(k)} \text{ and } r \geq t+m\}. \tag{2.86}$$

Since smoothing recursive weight $\alpha_{itj}^{(k)}$ is a derived statistics based on $\xi_{i,t}^{(k)}$ and $\eta_{t,j}^{(k)}$, BCMIX approximation to $\alpha_{itj}^{(k)}$ can use the index set defined in equations (2.83) and (2.85). Readers can also easily compute the posterior mean and variance of $\boldsymbol{\beta}_t$ and σ_t in three estimation scenarios (forward and backward filtering and smoothing) by keeping track of indices in $\mathcal{K}_t^{(k)}$ and $\widetilde{\mathcal{K}}_{t}^{(k)}$. I will not belabor the process.

The BCMIX procedure reduces the computation time and memory space to $O(TM)$ in filtering estimation and reduces to $O(TM^2)$ in smoothing estimation. Efficiency can be achieved for $M \ll T$. The accuracy and efficiency of BCMIX approximation clearly depend on the magnitude of M and m; If they are too small, important weights may be discarded at too early a stage and thus accuracy is sacrificed; if too large, time and memory space is compromised. In Section 3.2 and 3.3, we will discuss the effect of M and m on statistical inference and compare the BCMIX results with Bayes' estimates.

2.8 Computational Issues

As we can see from previous sections in this chapter, recursive updating the statistics has repeatedly occurred in the statistical inference in our model. To increase the computational efficiency in programming, one needs to implement dynamic programming which recursively calls the results from previous steps. Failure to apply dynamic programming scheme may increase the computation time substantially.

The estimation algorithm of this model is coded in C++ with the aid of TNT[1]and BOOST[2] library. Summary statistics, figures and tables are generated from R and MatLab. Simulation studies in Chapter 3 have been implemented on Stampede cluster provided by Texas Advanced Computing Center[3] (TACC). TACC Stampede system is a 10 PFLOPS(PF) Dell Linux Cluster based on 6400+ Dell PowerEdge server nodes, in which I use normal compute nodes and large memory nodes. Normal compute nodes (one server in the cluster) contain two 8-core 2.7 GHz Intel Xeon E5-2680 processors and 32 GB of DDR3 memory. Large memory nodes contain 4 E5-4650 8-core processors and 1024GB of DDR3 memory. Programs are operated in multiple

[1] http://math.nist.gov/tnt/

[2] http://www.boost.org/

[3] https://www.tacc.utexas.edu/

cores and multiple nodes in serial parallel. The following table presents an rough estimation of memory space required and running time for a single series estimation of a certain length. Operating multiple codes and multiple nodes in parallel can save computational time to within 24 hours for one panel of simulation, which would otherwise take over a hundred days in a dual-core PC with similar hardware configuration.

Table 2.1: Memory usage and running time for a single series with EM updating only once on Stampede cluster

Series Length	Memory required (GB)	Node type	Run time/node(min)
2000	8	Normal	1
3000	32	Large	2.11
4000	32	Large	3.12
5000	64	Large	5.24

Chapter 3

Simulation Studies

3.1 Diagnostics

To evaluate the goodness of fit of the estimated parameters, we propose the following 5 diagnostic statistics. Sum of squared errors (SSE) is a way to assess the performance of model parameter $\boldsymbol{\beta}_t$. For model (2.1), SSE is defined as

$$SSE := \frac{1}{T}\sum_{t=1}^{T}(y_t - \boldsymbol{x}_t'\widehat{\boldsymbol{\beta}}_t)^2 \tag{3.1}$$

Another useful statistic to evaluate the performance of $\boldsymbol{\beta}_t$ estimates is L_2 norm. L_2 norm measures the difference between the true parameters and the estimated parameters and is defined as

$$L_2 = \frac{1}{T}\sum_{t=1}^{T}\|\boldsymbol{\beta}_t - \widehat{\boldsymbol{\beta}}_t\| \tag{3.2}$$

Traditional SSE as a metric for model fitness is also an estimator of variance of random errors which cannot be explained by the regression model. It is inefficient to use SSE to evaluate model (2.1) because the error variances are estimated in the model. We propose a new metric called sum of squares of standardized errors (SSSE) to reflect that fact that the only error unexplained follows a standard normal distribution and all the rest of the information is explained by the model. SSSE is defined as

$$SSSE := \frac{1}{T}\sum_{t=1}^{T}\left(\frac{y_t - \boldsymbol{x}_t'\hat{\boldsymbol{\beta}}_t}{\hat{\sigma}_t}\right)^2 \tag{3.3}$$

An interesting statistic that assesses the accuracy of both $\hat{\boldsymbol{\beta}}_t$ and $\hat{\sigma}_t$ is named Kullback-Leibler (KL) divergence. KL divergence measures the difference between the true model in (2.1) and the estimated one by comparing their probability distributions. Let us define $\boldsymbol{\theta}_t = (\boldsymbol{\beta}_t, \sigma_t)$. KL divergence proved in Theorem 11 in Appendix A is

$$KL(\boldsymbol{\theta}_t, \widehat{\boldsymbol{\theta}}_t) = \frac{1}{2}\left(\frac{\big(\boldsymbol{x}_t'(\boldsymbol{\beta}_t - \widehat{\boldsymbol{\beta}}_t)\big)^2}{\hat{\sigma}_t^2} + \frac{\sigma_t^2}{\hat{\sigma}_t^2} - 1 - \log\left(\frac{\sigma_t^2}{\hat{\sigma}_t^2}\right)\right)$$

for every time point $t = 1, \ldots, T$. Then KL divergence for a time series is

$$KL = \frac{1}{T}\sum_{t=1}^{T} KL(\boldsymbol{\theta}_t, \widehat{\boldsymbol{\theta}}_t) \tag{3.4}$$

Ideally KL equals zero, if the estimated model is the same as the true model. In practice, the closer the KL values to zero the better the model fit.

The third important quantity in the estimation is the inference on the regime. Depending on the methods chosen, we have forward estimation of the regime in equation (2.7), or backward estimation in (2.35) or the smoothing estimation as in (2.59). These quantities are the estimated probabilities of system on a certain regime, whose values are between 0 and 1 theoretically. We may use a naive Bayesian classifier to classify the regimes. S_t equals to the state whose probability estimated is the largest. For example, in smoothing estimation, the estimated probability is from (2.59), i.e. $\alpha_t^{(k)} = \sum_{i=1}^{t} \sum_{j=t}^{T} \alpha_{itj}^{(k)}$, then the estimated regime at time t is

$$\widehat{S}_t = \arg\max_k \{\alpha_t^{(k)}\}$$

for $k = 1, \ldots, K$. Notice that there is only one regime at each time point. To evaluate the goodness of classification, we compute a rate called identification ratio (IR) to represent the percentage of correctly classified regime in a time series:

$$IR := \frac{\sum_{t=1}^{T} \sum_{k=1}^{K} \mathbf{1}_{\{\widehat{S}_t = k \,\cap\, S_t = k\}}}{T} \tag{3.5}$$

where $\mathbf{1}$ is an indicator function and S_t is the true state. IR is also easy to define in the context of forward, backward estimation and so on, therefore omitted.

In Monte Carlo simulation, we compare and contrast the means and standard errors of the above statistics. To keep this dissertation manageable, we only present the results by smoothing estimation.

3.2 Comparison of Bayes and BCMIX Estimation

The goal of this section is to compare the two estimation methods (Bayes and BCMIX) and show that BCMIX is an efficient estimation method. In addition, the simulation setting is constructed so that the model estimates can be compared with the results from piecewise linear regression models which will be discussed later in this section. We begin with a simple model,

$$y_t = \beta_t y_{t-1} + \sigma_t \epsilon_t \tag{3.6}$$

The data are generated from a two-regime system by the following prior parameters. $\gamma_t^{(k)} = \frac{1}{2(\sigma_t^{(k)})^2} \sim$ Gamma $(g^{(k)}, \lambda^{(k)})$ where $g^{(1)} = 2.5$, $g^{(2)} = 1.2$, $\lambda^{(1)} = 0.8$, $\lambda^{(2)} = 1$. $\beta_t^{(k)}$ is generated from a truncated normal distribution $N\left(z^{(k)}, \frac{V^{(k)}}{2\gamma^{(k)}}\right)$ where $z^{(1)} = 0.3$, $z^{(2)} = -0.3$, $V^{(1)} = 0.16$, $V^{(2)} = 0.16$ and $|\beta_t^{(k)}| < 1$ so that the series is stationary. 500 series are generated and each with length T=1000.

In a comprehensive Bayesian statistical analysis, hyperparameters need to be estimated first and then used to estimate model parameters which will be discussed in later section. Since the purpose of this section is to compare the effects of piece-wise linear regression, BAYES method and BCMIX method on the parameter estimation, the estimation begins with true prior parameters.

Also in this simulation, the regime switching points are manually fixed.

Thus Markov transition probability matrix P is not used to generate the data. However, there is still a need to choose a proper P to make the estimation more reasonable. It is expected that the more transition points in a series, the larger transition probability would be between the two states. Because the expected number of switches from state 1 to state 2 is $\sum_{t=1}^{T} P(S_t = 2|S_t = 1)P(S_t = 1) = \sum_{t=1}^{1000} p_{12} \times \pi_1$ and in the same vain the expected number of switches from state 2 to state 1 is $\sum_{t=1}^{1000} p_{21} \times \pi_2$, different prior transition probabilities are chosen for the following scenarios. Assume the series begin with state 1.

Scenario 1 Single change point at t = 501. $S_t = 1$ for $1 \leq t \leq 500$ and $S_t = 2$ for $501 \leq t \leq 1000$. Assume $p_{11} = 0.998, p_{12} = 0.002, p_{21} = 0.002$ and $p_{22} = 0.998$.

Scenario 2 Two transition points at $t = 351$ and $t = 701$. $S_t = 1$ for $1 \leq t \leq 350$ and $701 \leq t \leq 1000$; $S_t = 2$ for $351 \leq t \leq 700$. Assume $p_{11} = 0.998, p_{12} = 0.002, p_{21} = 0.002$ and $p_{22} = 0.998$.

Scenario 3 Three transition points at $t = 251$, $t = 501$ and $t = 751$. $S_t = 1$ for $1 \leq t \leq 250$ and $501 \leq t \leq 750$; $S_t = 2$ for $251 \leq t \leq 500$ and $751 \leq t \leq 1000$. Assume $p_{11} = 0.996, p_{12} = 0.004, p_{21} = 0.004$ and $p_{22} = 0.996$.

Scenario 4 Four transition points at $t = 201, 401, 601,$ and 801. $S_t = 1$ for $1 \leq t \leq 200$, $401 \leq t \leq 600$ and $801 \leq t \leq 1000$; $S_t = 2$ for $201 \leq t \leq$

400, and $601 \leq t \leq 800$. Assume $p_{11} = 0.996, p_{12} = 0.004, p_{21} = 0.004$ and $p_{22} = 0.996$.

Scenario 5 Five transition points at $t = 201, 351, 501, 651$ and 801. $S_t = 1$ for $1 \leq t \leq 200$, $351 \leq t \leq 500$ and $651 \leq t \leq 800$; $S_t = 2$ for $201 \leq t \leq 350$, $501 \leq t \leq 650$ and $801 \leq t \leq 1000$. Assume $p_{11} = 0.994, p_{12} = 0.006, p_{21} = 0.006$ and $p_{22} = 0.994$.

Table 3.1 shows the means and standard errors of 500 simulations in each scenario for the five diagnostic statistics discussed in section 3.1. The results of least square regression are shown in the column titled by "Oracle"; results of Bayes estimation titled by "Bayes" and those of BCMIX estimation titled by "BCMIX" in the last four columns of every subtable. Bayes method uses comprehensive recursive weights whereas BCMIX method uses the selected weights depending on M and m. In this simulation, we choose $M = 15, 20, 30, 40$ and $m = 10, 10, 15, 20$ corresponding to previous M.

Table 3.1a shows that means of KL statistics in Bayes method are no larger than those in BCMIX method for every scenario, the difference is very small. In fact in scenario 1 and 2, BCMIX is just as good as Bayes, where all yields the same KL statistics; in scenario 3 to 5, when $M = 15$, $m = 10$, KL statistics of BCMIX are only 0.0002 or 0.0003 higher (roughly 2.3% to 2.8% more) than those of Bayes. As M and m increase, KL of BCMIX converges to that of Bayes. When $M = 40$, $m = 20$, BCMIX estimation is almost the same as Bayes estimation, comparing the third and the last column of Table 3.1a. Standard errors of KL statistics are as small as the order of 10^{-4} and become smaller when M and m increase. This confirms the fact that BCMIX is an efficient alternative of Bayes method. In addition, as the number of regime switching increases from one to five (scenario 1 to 5), mean KL statistics increase correspondingly. This pattern is reasonable, because the more switches, the more parameters to estimate, and thus the bigger estimation errors. Since model (3.6) is in linear regression form and data between the two change points are generated from the same probability distribution, least square estimation is an ideal choice for this simple linear regression. Least square regression, though infeasible when the change points are unknown, is a good benchmark to compare how well the proposed method behaves. It is not surprising that among all scenarios mean KL statistics and standard errors under "Oracle" are the smallest, an indication that $\widehat{\beta}_t$ and $\widehat{\sigma}_t$ are closer to the true parameter values. KL statistics by Bayes or BCMIX are twice of those of "Oracle". The overall mean KL statistics are small and close to zero, a sign of good model fit.

To evaluate the performance of $\widehat{\beta}_t$ alone, Table 3.1b and 3.1c shows the means and standard errors of SSE and L_2 norm respectively. In both tables, Bayes method shows the smallest statistics compared with BCMIX method except one case (scenario 1 in Table 3.1c) due to possible errors in random sampling. In BCMIX method, as M and m increase, the means and standard errors of both statistics become smaller for every scenario. It is logic to induce that BCMIX estimates converge to Bayes estimates. The difference caused by Bayes and BCMIX is also small. For example, in 3.1b, the results of BCMIX(20, 10) are only 0.0001 to 0.0002 higher (roughly 0.009% to 0.02% more) than those of Bayes; and in Table 3.1c, except scenario 1, the results of BCMIX(40, 20) are more than its Bayes counterparts by less than 0.0001 (roughly no more than 0.2%). The overall L_2 statistics become larger when the number of switches goes up from one to five (scenario 1 to 5). This pattern does not apply to SSE statistics, simply because (3.1) measures the mean errors which are not constant and stochastically generated by a gamma distribution in model specification. The overall magnitude of the errors relies on the random generation mechanism associated with the computer software. In general SSE and L_2 statistics do not give a panorama of the model performance.

Like KL statistic, the invented $SSSE$ statistic, whose means and standard errors are shown in Table 3.1d, evaluates the estimation of both model parameters β_t and σ_t. This statistic is specially designed for the proposed model in this thesis, and uses the estimated σ_t at each time $t = 1, \ldots, T$, which is not applicable for least squares linear regression. $SSSE$ is also an estimate of the error in a standard normal distribution, thus the closer this statistic to 1, the better the model performance. Table 3.1d shows that all the means are greater than 0.96 with standard errors at the order of 10^{-4}. When the number of changes increase from one to five (scenario 1 to 5), the mean statistics are further away from 1. The largest of this statistic is almost 0.99, achieved in one change point setting. For scenario 2 to 5, means of BCMIX(15, 10) slightly drop and the standard errors increase a couple of units than those under Bayes but still in the same order of 10^{-4}; as M and m increase, $SSSE$ statistics regain and standard errors drop towards the levels in Bayes. Interesting though, when $(M, m) = (40, 20)$, the mean statistics appear to be even closer to 1 and the standard errors are still slightly larger than those in Bayes method. The improvement in BCMIX estimation may be taken as a confirmation that BCMIX is an efficient method for Bayes. Or this improvement could be possibly spurious due to the fact that the more errors in a statistical estimation tend to boost the diagnostic statistics.

The last diagnostic statistic to be discussed in this section is the identification ratio (IR) defined in (3.5). Unlike any of the previous statistics, IR evaluates the performance of the inference in the regime, i.e. the percentage of regimes that are correctly classified. From Table 3.1e, in the two-change-point setting (scenario 2), IR is as high as 99%; whereas in the five-change-point setting (scenario 5), IR is roughly 3% lower. For a fixed series length, when the number of change points increases, the IR statistics naturally drop, because the estimation is rather fuzzy around every change point. Thus the more change points, the more errors in the estimation, and the lower IR values. From column 2 to column 6 (Table 3.1e), the mean IR statistics for each scenario first drop in BCMIX(15, 10) by about 0.1% and then regain to Bayes levels in BCMIX(20, 10). In BCMIX setting, standard errors show a pattern of decreasing to the levels in Bayes method with the increasing M and m. Within BCMIX framework alone, for every scenario the means of IR statistics increases a little and standard errors decrease when (M, m) changes from (15, 10) to (40, 20) and eventually stabilize for large M and m. This results again confirm that BCMIX is an efficient method to Bayes.

As discussed earlier, BCMIX method is computationally efficient and from the analysis of Table 3.1, BCMIX is also statistically efficient compared to Bayes method. The last question to answer before the closure of this section is the choice of M and m in BCMIX. Though the larger M and m the better estimation, computational time is proportional to M in the order of $O(TM)$ in forward or backward filtering and in the order of $O(TM^2)$ in smoothing estimation where T is the series length. Let us re-

Table 3.1: Monte Carlo means of diagnostic statistics by Oracle, Bayes and BCMIX methods. Standard errors are shown in parenthesis.

Scenarios	Oracle	Bayes	BCMIX			
			(15, 10)	(20, 10)	(30, 15)	(40, 20)
Scenario 1	0.0020 (6.4e-05)	0.0033 (1.0e-04)	0.0033 (1.0e-04)	0.0033 (1.0e-04)	0.0033 (1.0e-04)	0.0033 (1.0e-04)
Scenario 2	0.0031 (7.6e-05)	0.0070 (1.3e-03)	0.0070 (1.3e-03)	0.0070 (1.3e-03)	0.0070 (1.3e-03)	0.0070 (1.3e-03)
Scenario 3	0.0039 (8.7e-05)	0.0081 (2.8e-04)	0.0083 (2.9e-04)	0.0082 (2.9e-04)	0.0082 (2.8e-04)	0.0082 (2.8e-04)
Scenario 4	0.0050 (1.0e-04)	0.0107 (3.1e-04)	0.0110 (3.2e-04)	0.0108 (3.1e-04)	0.0107 (3.1e-04)	0.0107 (3.1e-04)
Scenario 5	0.0061 (1.2e-04)	0.0132 (3.3e-04)	0.0135 (3.5e-04)	0.0133 (3.4e-04)	0.0132 (3.4e-04)	0.0132 (3.4e-04)

(a) Kullback-Leibler (KL) divergence

Table 3.1: Continued:

Scenarios	Oracle	Bayes	BCMIX			
			(15, 10)	(20, 10)	(30, 15)	(40, 20)
Scenario 1	0.8625 (9.8e-02)	0.8616 (9.8e-02)	0.8617 (9.8e-02)	0.8617 (9.8e-02)	0.8617 (9.8e-02)	0.8617 (9.8e-02)
Scenario 2	1.0437 (2.4e-01)	1.0426 (2.4e-01)	1.0427 (2.4e-01)	1.0427 (2.4e-01)	1.0427 (2.4e-01)	1.0426 (2.4e-01)
Scenario 3	2.1695 (1.1e+00)	2.1677 (1.1e+00)	2.1679 (1.1e+00)	2.1679 (1.1e+00)	2.1678 (1.1e+00)	2.1678 (1.1e+00)
Scenario 4	1.6482 (7.2e-01)	1.6460 (7.2e-01)	1.6461 (7.2e-01)	1.6461 (7.2e-01)	1.6461 (7.2e-01)	1.6461 (7.2e-01)
Scenario 5	0.9745 (7.8e-02)	0.9709 (7.8e-02)	0.9712 (7.8e-02)	0.9711 (7.8e-02)	0.9711 (7.8e-02)	0.9710 (7.8e-02)

(b) Sum of Squared Error (SSE)

Scenarios	Oracle	Bayes	BCMIX			
			(15, 10)	(20, 10)	(30, 15)	(40, 20)
Scenario 1	0.0327 (7.8e-04)	0.0361 (7.7e-04)	0.0358 (7.8e-04)	0.0357 (7.8e-04)	0.0358 (7.7e-04)	0.0358 (7.7e-04)
Scenario 2	0.0402 (7.7e-04)	0.0443 (7.9e-04)	0.0447 (8.3e-04)	0.0446 (8.1e-04)	0.0445 (8.0e-04)	0.0444 (8.0e-04)
Scenario 3	0.0438 (7.6e-04)	0.0519 (8.1e-04)	0.0528 (8.9e-04)	0.0523 (8.5e-04)	0.0520 (8.3e-04)	0.0520 (8.3e-04)
Scenario 4	0.0503 (7.8e-04)	0.0598 (8.6e-04)	0.0610 (9.4e-04)	0.0603 (9.1e-04)	0.0599 (8.8e-04)	0.0598 (8.8e-04)
Scenario 5	0.0547 (7.7e-04)	0.0667 (8.8e-04)	0.0680 (9.9e-04)	0.0671 (9.4e-04)	0.0669 (9.2e-04)	0.0668 (9.1e-04)

(c) L_2 norm

Table 3.1: Continued

Scenarios	Bayes	BCMIX			
		(15, 10)	(20, 10)	(30, 15)	(40, 20)
Scenario 1	0.9891 (2.7e-04)	0.9893 (3.3e-04)	0.9893 (3.0e-04)	0.9894 (2.9e-04)	0.9894 (2.8e-04)
Scenario 2	0.9822 (3.4e-04)	0.9817 (5.3e-04)	0.9821 (4.7e-04)	0.9824 (3.7e-04)	0.9824 (3.6e-04)
Scenario 3	0.9760 (3.8e-04)	0.9747 (6.0e-04)	0.9758 (5.0e-04)	0.9762 (4.3e-04)	0.9762 (4.2e-04)
Scenario 4	0.9697 (4.0e-04)	0.9674 (6.5e-04)	0.9692 (5.1e-04)	0.9699 (4.4e-04)	0.9700 (4.2e-04)
Scenario 5	0.9640 (4.5e-04)	0.9622 (7.0e-04)	0.9636 (6.0e-04)	0.9642 (5.1e-04)	0.9642 (4.9e-04)

(d) Sum of Squares of Standardized Error (SSSE)

Scenarios	Bayes	BCMIX			
		(15, 10)	(20, 10)	(30, 15)	(40, 20)
Scenario 1	0.9835 (4.0e-03)	0.9802 (4.4e-03)	0.9807 (4.6e-03)	0.9844 (4.1e-03)	0.9845 (4.0e-03)
Scenario 2	0.9851 (1.9e-03)	0.9867 (2.0e-03)	0.9871 (2.0e-03)	0.9872 (2.0e-03)	0.9867 (1.9e-03)
Scenario 3	0.9751 (2.8e-03)	0.9700 (3.6e-03)	0.9728 (3.1e-03)	0.9765 (2.7e-03)	0.9763 (2.7e-03)
Scenario 4	0.9730 (2.1e-03)	0.9687 (2.7e-03)	0.9713 (2.3e-03)	0.9725 (2.2e-03)	0.9731 (2.1e-03)
Scenario 5	0.9606 (2.8e-03)	0.9540 (3.3e-03)	0.9572 (3.1e-03)	0.9576 (3.0e-03)	0.9588 (2.9e-03)

(e) Identification Ratio (IR)

view Table 3.1a again. In each scenario of 1 or 2, means and standard errors are all the same for both Bayes and BCMIX methods. In scenario 3 to 5, the mean statistics improve by 0.0001 or 0.0002 points and standard errors either remain the same or become slightly smaller from BCMIX(15, 10) to BCMIX(20, 10) setting. As M and m increase, the improvement is barely conspicuous. Besides, the estimates under BCMIC(20, 10) are just as good as those under Bayes. Similar phenomena apply to Table 3.1b. In Table 3.1c, mean statistic drops by 0.0005, 0.0007 and 0.0009 and the standard error drops by 0.01, 0.3 and 0.5 of 10^{-4} in scenario 3-5 when comparing BCMIX(15, 10) with BCMIX(20, 10) setting. When comparing BCMIX(20, 10) with BCMIX(30, 15) the mean statistics decrease by 0.0003, 0.0004 and 0.0002 and standard errors decrease by 0.2, 0.3 and 0.2 of order 10^{-4} in the same scenarios. This result shows that the improvement of the estimation is bigger from BCMIX(15, 10) to BCMIX(20, 10) than from BCMIX(20, 10) to BCMIX(30, 15). The analysis of Table 3.1d and 3.1e is similar and therefore omitted. In general $(M, m) = (20, 10)$ is a good choice for BCMIX method. We will use BCMIX(20, 10) in larger simulation environment in the following sections.

3.3 Estimation with Estimated Hyperparameters

The simulation in this section is very similar to that in section 3.2 in ways of fixed position switching points and the series length. The difference is that we choose a little complex model and estimate the hyperparameters with EM algorithm defined in section 2.6. First an AR(1) model is defined as

$$y_t = \alpha_t + \beta_t y_{t-1} + \sigma_t \epsilon_t \tag{3.7}$$

The true series are generated by the prior values such as $\boldsymbol{z}^{(1)\prime} = (0.3,\ 0.5)$, $\boldsymbol{z}^{(2)\prime} = (-0.2,\ -0.5)$ and $\boldsymbol{V}^{(1)} = \boldsymbol{V}^{(2)} = \begin{pmatrix} 0.16 & 0 \\ 0 & 0.16 \end{pmatrix}$ and $g^{(1)} = 2.5$, $g^{(2)} = 1.2$, $\lambda^{(1)} = 0.8$ and $\lambda^{(2)} = 1$. We use EM algorithm to estimate prior values $\boldsymbol{z}$, $\boldsymbol{V}$, $\boldsymbol{g}$, λ and transition matrix $\boldsymbol{P}$. The initial prior values are chosen as follows: $\boldsymbol{z} = \begin{pmatrix} 0.2 & 0.4 \\ -0.2 & -0.3 \end{pmatrix}$, $\boldsymbol{V} = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}$, $g = \begin{pmatrix} 3 & 1.5 \end{pmatrix}$, $\lambda = \begin{pmatrix} 1 & 2 \end{pmatrix}$, and $P = \begin{pmatrix} 0.99 & 0.01 \\ 0.01 & 0.99 \end{pmatrix}$ We consider the following 7 fixed change-point scenarios.

Scenario 1 There is only one transition point at $t = 501$. $S_t = 1$ for $1 \leq t \leq 500$ and $S_t = 2$ for $501 \leq t \leq 1000$.

Scenario 2 There are two transition points at $t = 351$ and $t = 701$. $S_t = 1$

for $1 \leq t \leq 350$, $S_t = 2$ for $351 \leq t \leq 700$ and $S_t = 1$ for $701 \leq t \leq 1000$.

Scenario 3 There are three transition points at $t = 251$, $t = 501$ and $t = 751$. $S_t = 1$ for $1 \leq t \leq 250$ and $501 \leq t \leq 750$; $S_t = 2$ for $251 \leq t \leq 500$ and $751 \leq t \leq 1000$.

Scenario 4 There are four transition points at $t = 201, 401, 601,$ and 801. $S_t = 1$ for $1 \leq t \leq 200$, $401 \leq t \leq 600$ and $801 \leq t \leq 1000$; $S_t = 2$ for $201 \leq t \leq 400$ and $601 \leq t \leq 800$.

Scenario 5 There are five transition points at $t = 201, 351, 501, 651$ and 801. $S_t = 1$ for $1 \leq t \leq 200$, $351 \leq t \leq 500$ and $651 \leq t \leq 800$; $S_t = 2$ for $201 \leq t \leq 350$, $501 \leq t \leq 650$ and $801 \leq t \leq 1000$.

Scenario 6 There are six transition points at $t = 141, 281, 421, 561, 701$ and 851. $S_t = 1$ for $1 \leq t \leq 140$, $281 \leq t \leq 420$, $561 \leq t \leq 700$, $851 \leq t \leq 1000$; $S_t = 2$ for $141 \leq t \leq 280$, $421 \leq t \leq 560$ and $701 \leq t \leq 850$.

Scenario 7 There are eight transition points at $t = 111, 221, 331, 441, 551, 661, 771$ and 881. $S_t = 1$ for $1 \leq t \leq 110$, $221 \leq t \leq 330$, $441 \leq t \leq 550$, $661 \leq t \leq 770$ and $881 \leq t \leq 1000$; $S_t = 2$ for $111 \leq t \leq 220$, $331 \leq t \leq 440$, $551 \leq t \leq 660$ and $771 \leq t \leq 880$.

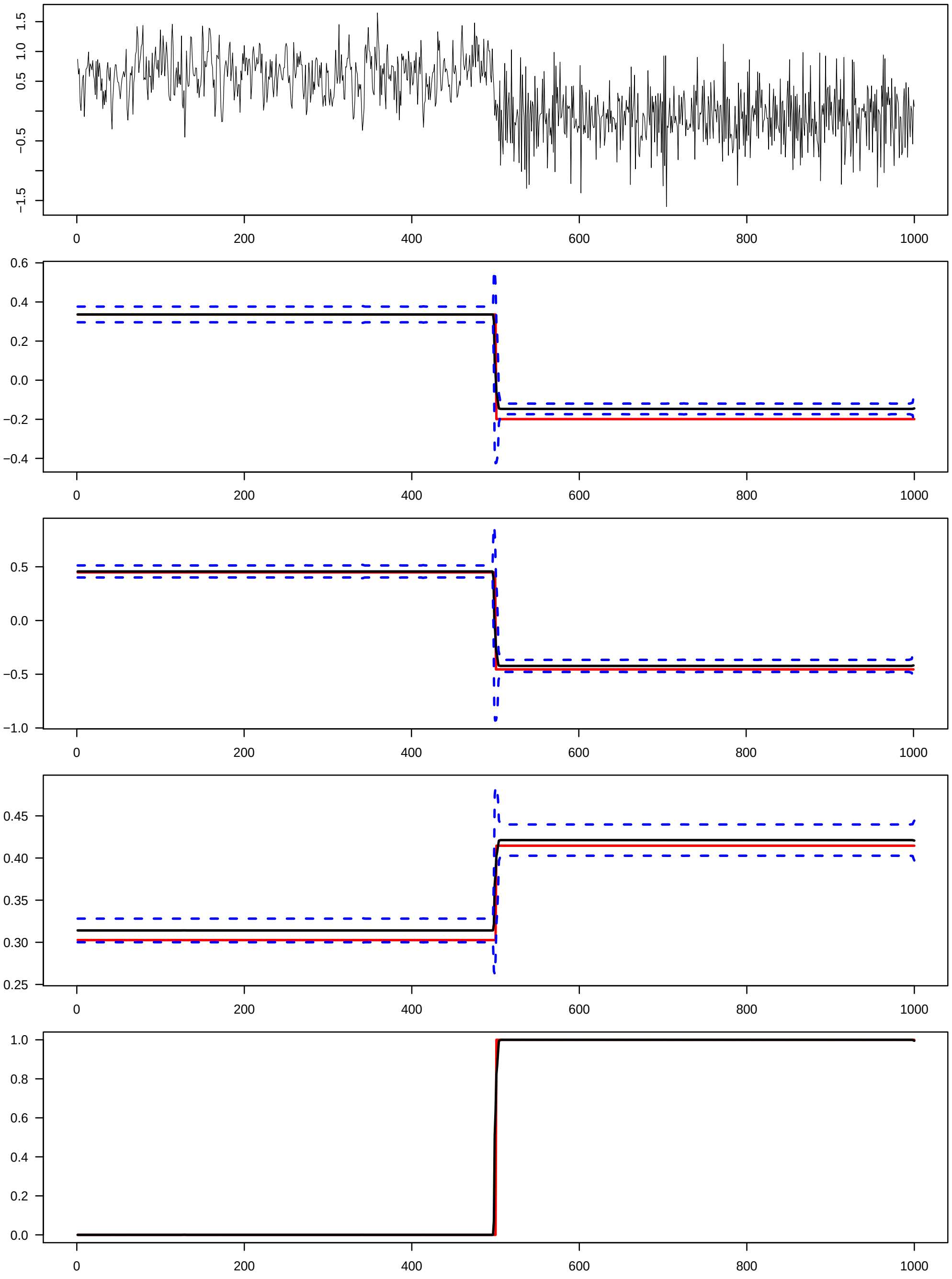
Section 3.2 has proven that BCMIX method is as efficient as the time consuming Bayes method. Thus table 3.2 shows only the results of BCMIX method in comparison with "Oracle" (least squares linear regression when the change points are known). As a confirmation of the results in section 3.2, the diagnostic statistics in Table 3.2 become stable when BCMIX parameters M and m become larger. The statistics in BCMIX method are also very close to the statistics under "Oracle" if applicable under all scenarios and settings. In general KL statistics are very close to zero and SSSE and IR statistics are close to one, which are indicators of good model fit. Readers interested in the in-depth analysis of Table 3.2 may refer to section 3.2 for details. We avoid the repetition of the similar analysis.

Instead in this section, we focus on graphical presentations of the estimation and model fit. Figure 3.1 shows the estimation of the model parameters and the regime status in comparison with the true parameters from selected series in each scenario. Every sub figure in Figure 3.1 has 5 plots. The first plot is a simulated time series; the rest regards to the estimation of $\alpha_t, \beta_t, \sigma_t$ and $P(S_t = 2)$. The true parameters are indicated by solid red line; the estimates by solid black line and the confidence intervals are shown by dashed blue line. Here we choose a typical 95% confidence interval computed by mean estimate plus/minus 1.96 times the standard deviation of the parameter.

Take Figure 3.1c for example. The top figures shows the simulated time series y_t generated by model (3.7). For $1 \leq t \leq 250$, y_t was at regime $S_t = 1$ and was generated by $(\alpha_t, \beta_t, \sigma_t) = (0.251, 0.499, 0.563)$ plus a Gaussian noise (variance equals 1). The estimated $\widehat{\alpha}_t$'s (solid black in the second plot) at this period are time-varying and a little overestimated than the true value, and their confidence intervals (dashed blue) merely cover the true parameter. As far as β_t is concerned, the estimated $\widehat{\beta}_t$'s (solid black in the third plot) are all below the true value 0.499 and the confidence intervals (dashed blue) are a little off the true parameter value. σ_t estimation seems to be good during this period, as the estimates are all close to 0.563 and the confidence intervals include this true value. For regime status, the bottom plot shows the probability that the regime is at state 2 and this value would be either 0 or 1 if the true regime is known. For t from 1 through 250, the true state is 1, and thus the probability of the regime at 2 is 0 (solid red). In a two-regime system, the regime at time t belongs to the state whose probability is larger than 0.5 according to naive Bayesian classifier. The estimation of the regime $\widehat{P}(S_t = 2)$ (solid black) is close to zero during and thus the estimated state is 1 during this period.

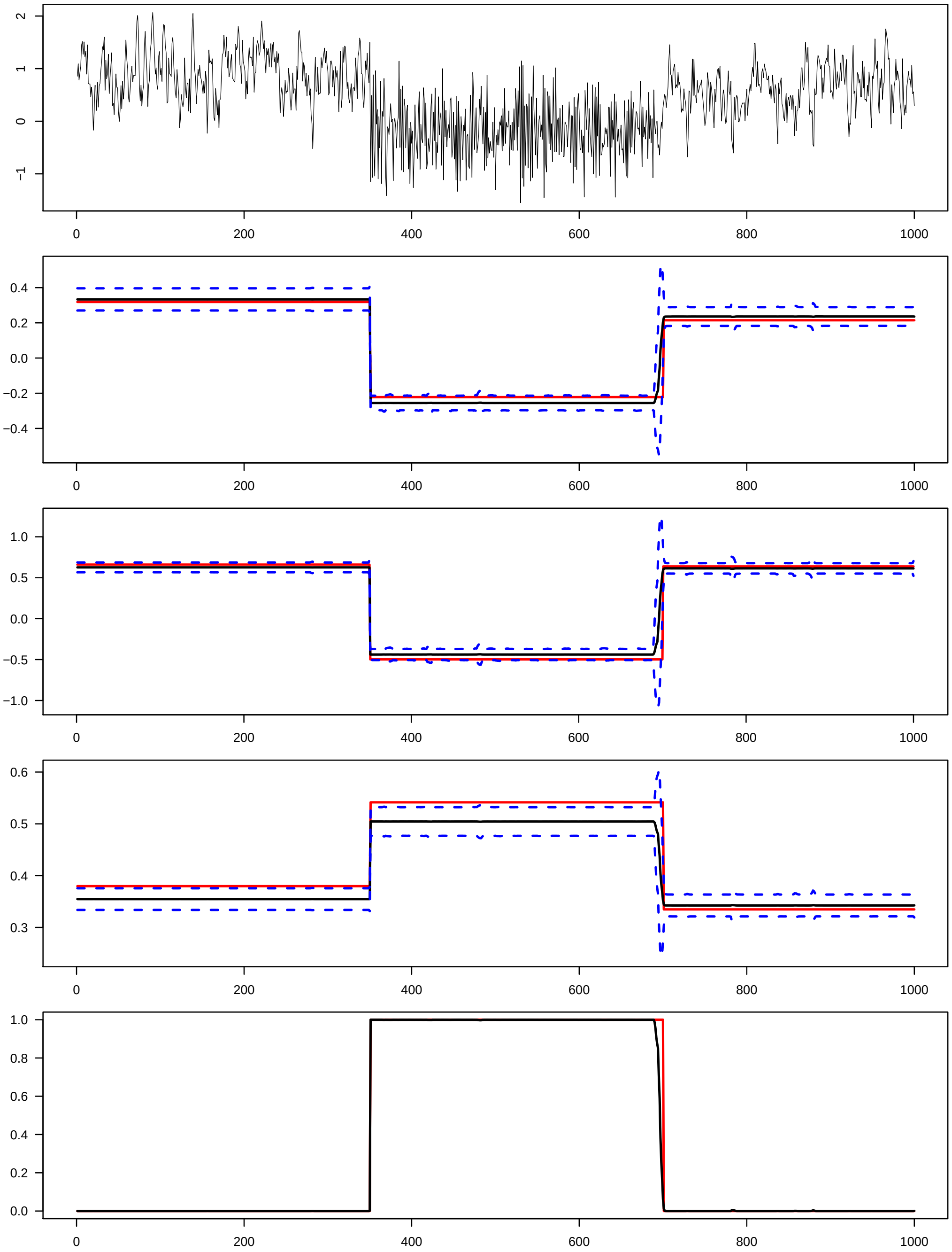
Furthermore, for time period between 251 and 500, the true regime is 2 and true $(\alpha_t, \beta_t, \sigma_t) = (-0.435,\ -0.396,\ 0.584)$; for t between 501 and 750, true regime is 1 and true $(\alpha_t, \beta_t, \sigma_t) = (0.415,\ 0.355,\ 0.409)$ and finally for t from 751 to the end of the series, the true regime is again 2 and true $(\alpha_t, \beta_t, \sigma_t) = (-0.025,\ -0.415,\ 0.467)$. In each of the three sub periods, the estimated $\widehat{\alpha}_t, \widehat{\beta}_t$, $\widehat{\sigma}_t$ and $\widehat{P}(S_t = 2)$ are close to their true counterparts and their confidence intervals include the corresponding true parameters with no exception. When the series is making a transition from

Figure 3.1: Selected series from each scenario: the true parameters (solid red) and estimates (solid black) by BCMIX(20, 10) and 95% confidence intervals (dashed blue).



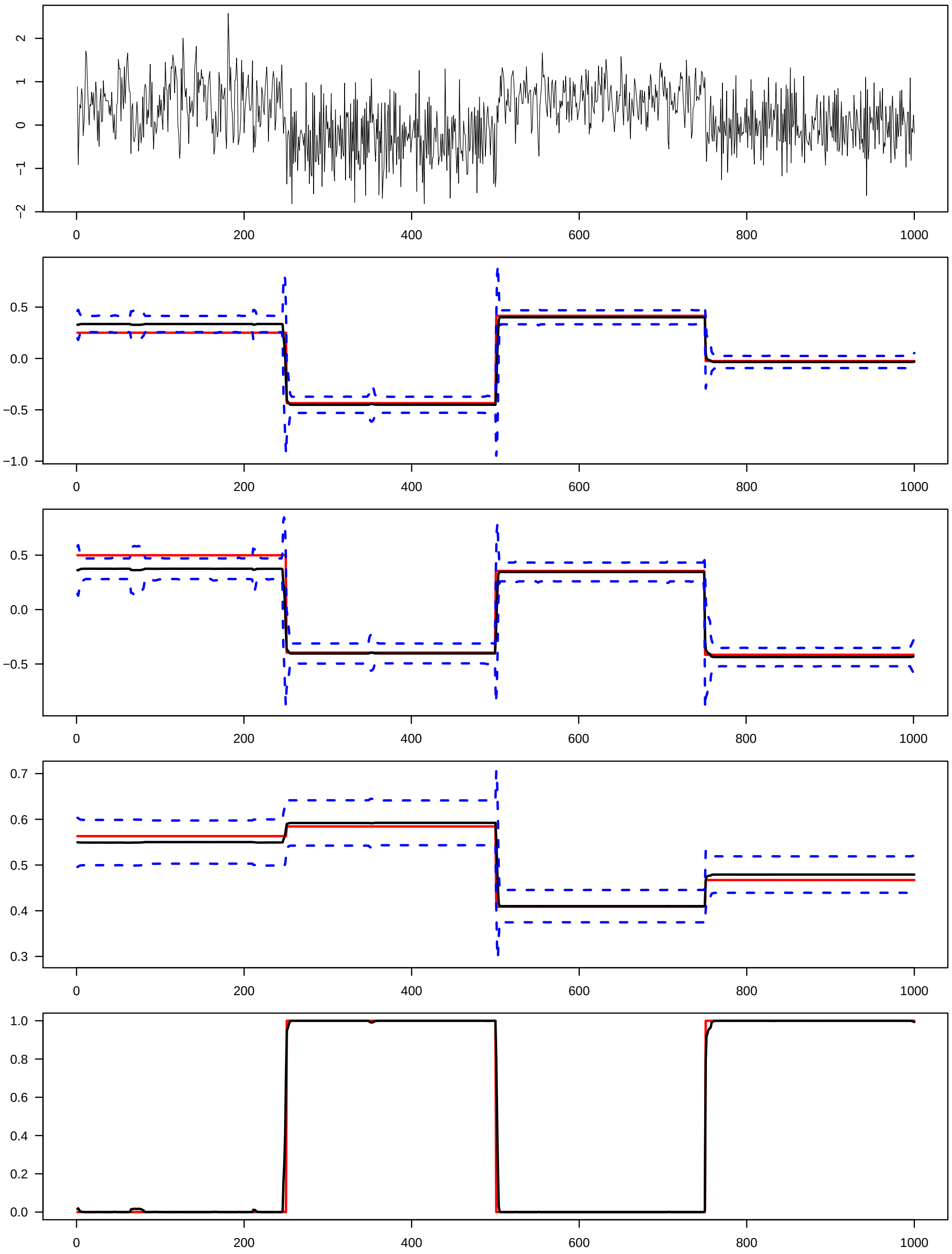
(a) Selected series y_t from scenario 1 (top), α_t (second), β_t (third), σ_t (fourth) and $P(S_t = 2)$ (bottom)

Figure 3.1: Continued



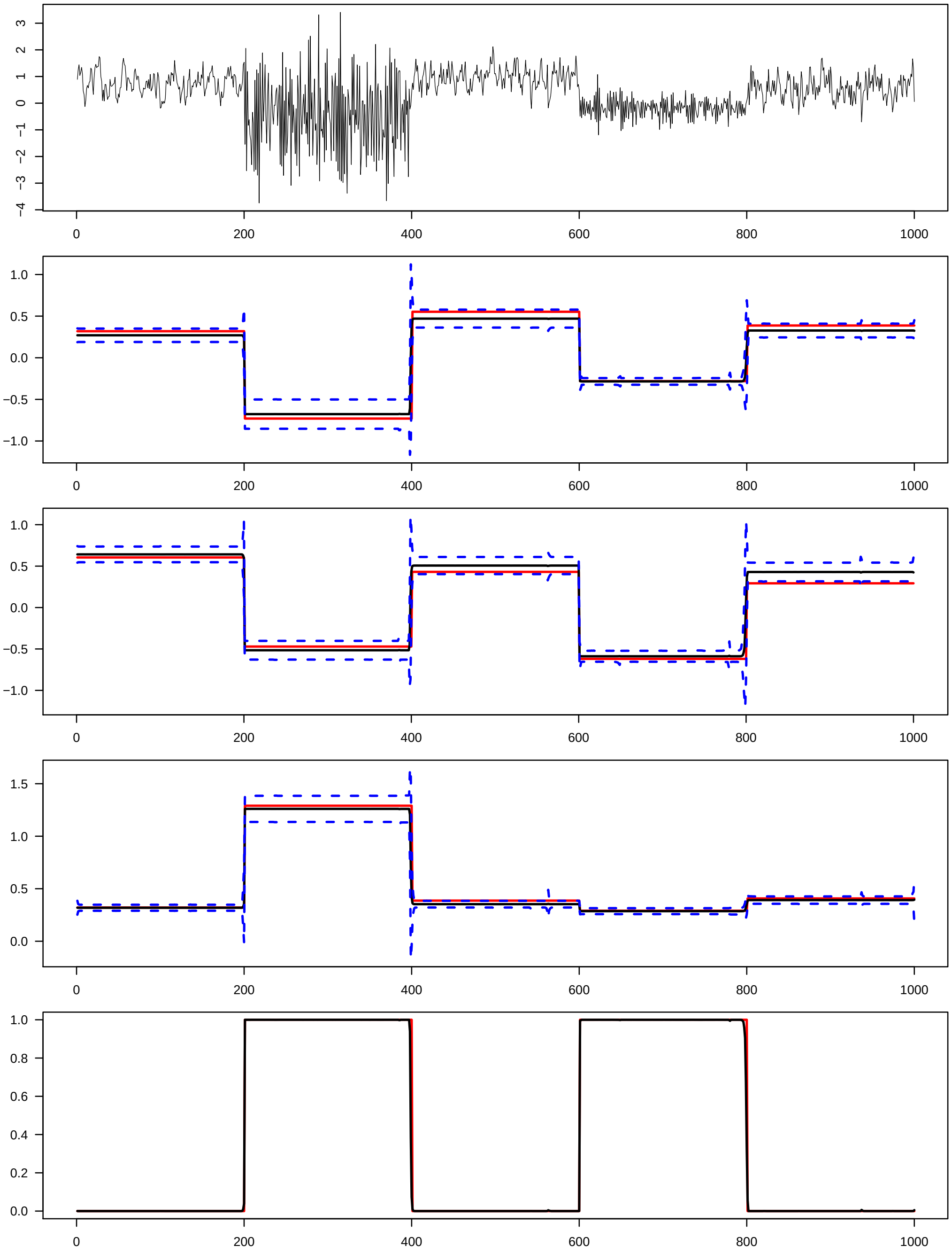
(b) Selected series y_t from scenario 2 (top), α_t (second), β_t (third), σ_t (fourth) and $P(S_t = 2)$ (bottom)

Figure 3.1: Continued



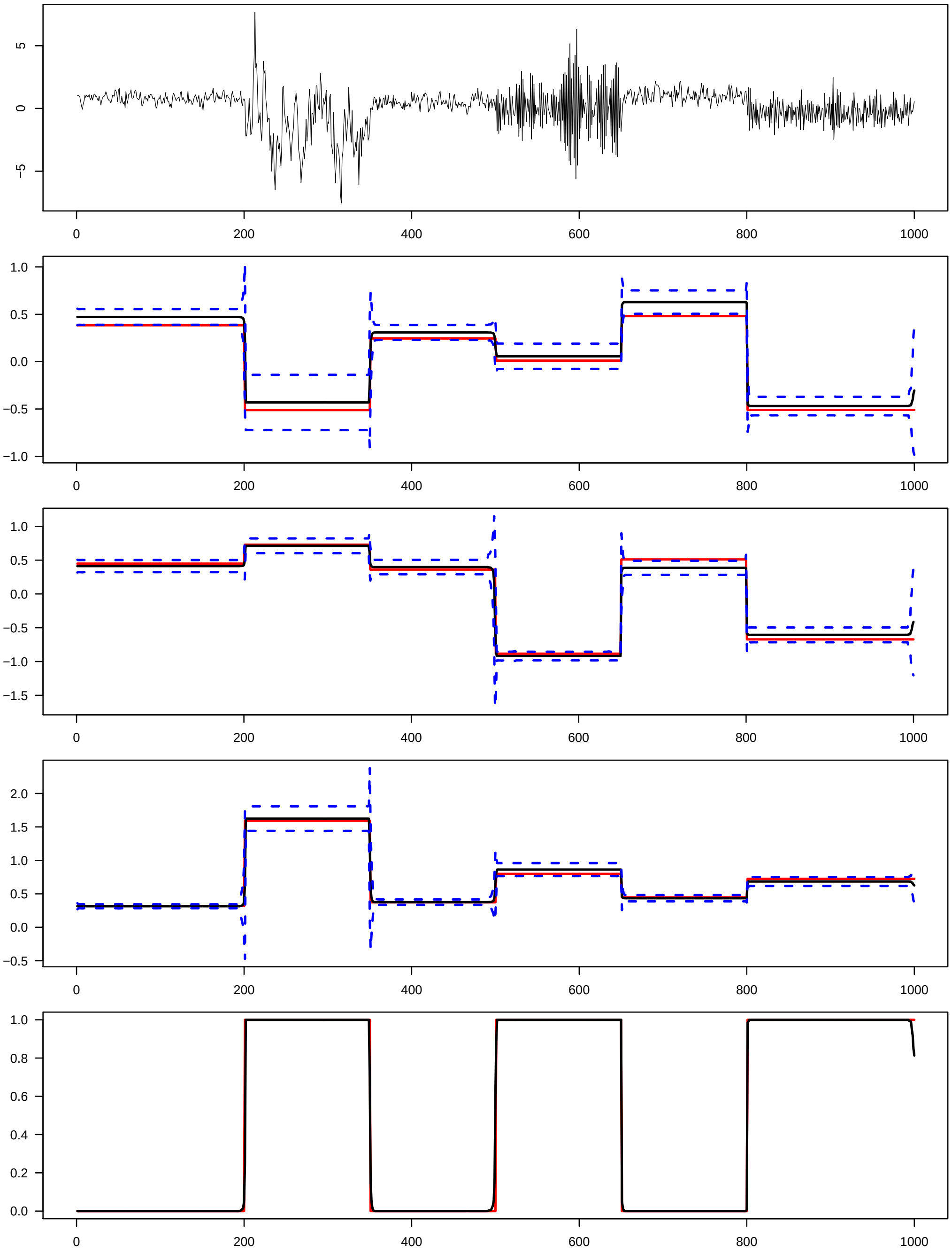
(c) Selected series y_t from scenario 3 (top), α_t (second), β_t (third), σ_t (fourth) and $P(S_t = 2)$ (bottom)

Figure 3.1: Continued



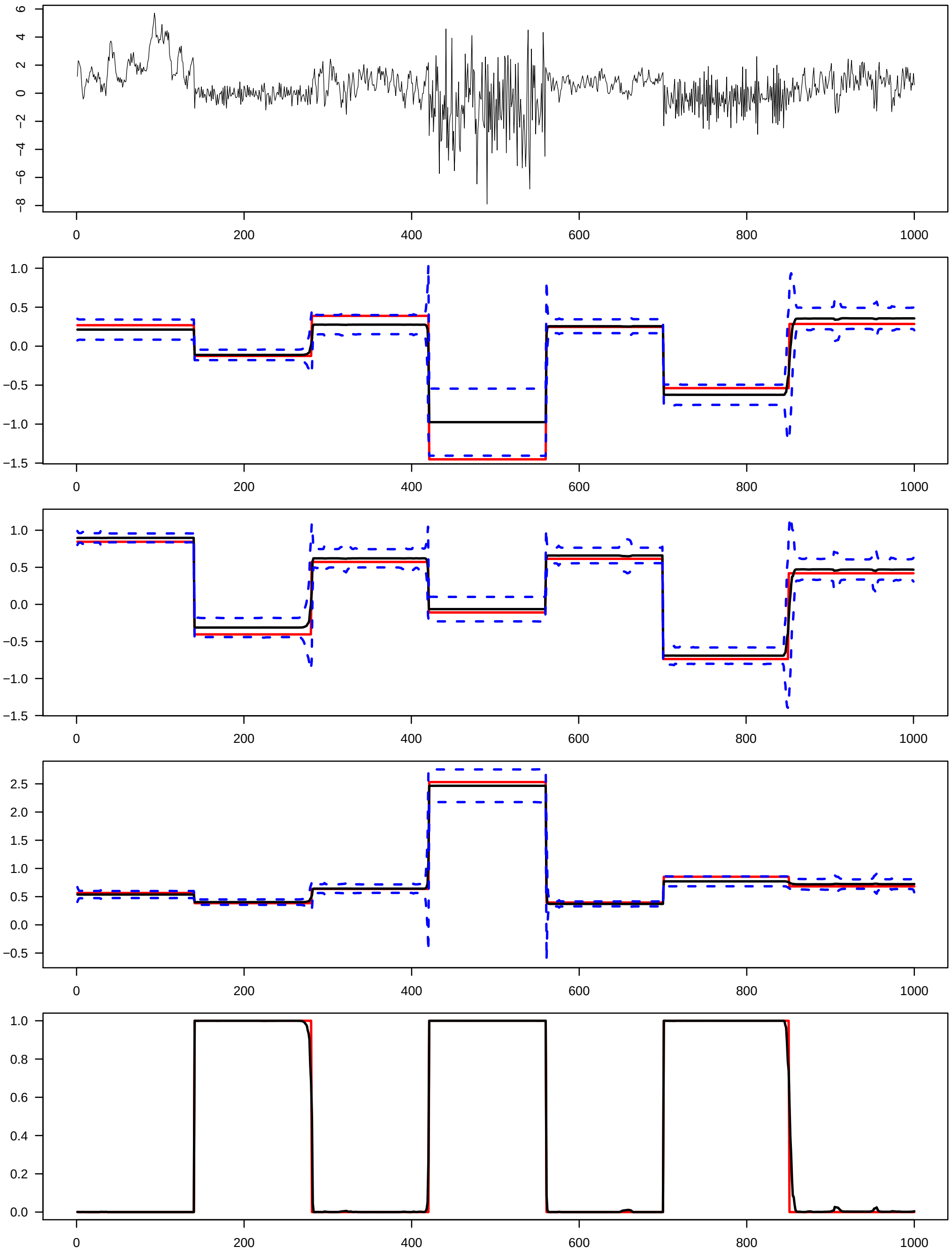
(d) Selected series y_t from scenario 4 (top), α_t (second), β_t (third), σ_t (fourth) and $P(S_t = 2)$ (bottom)

Figure 3.1: Continued



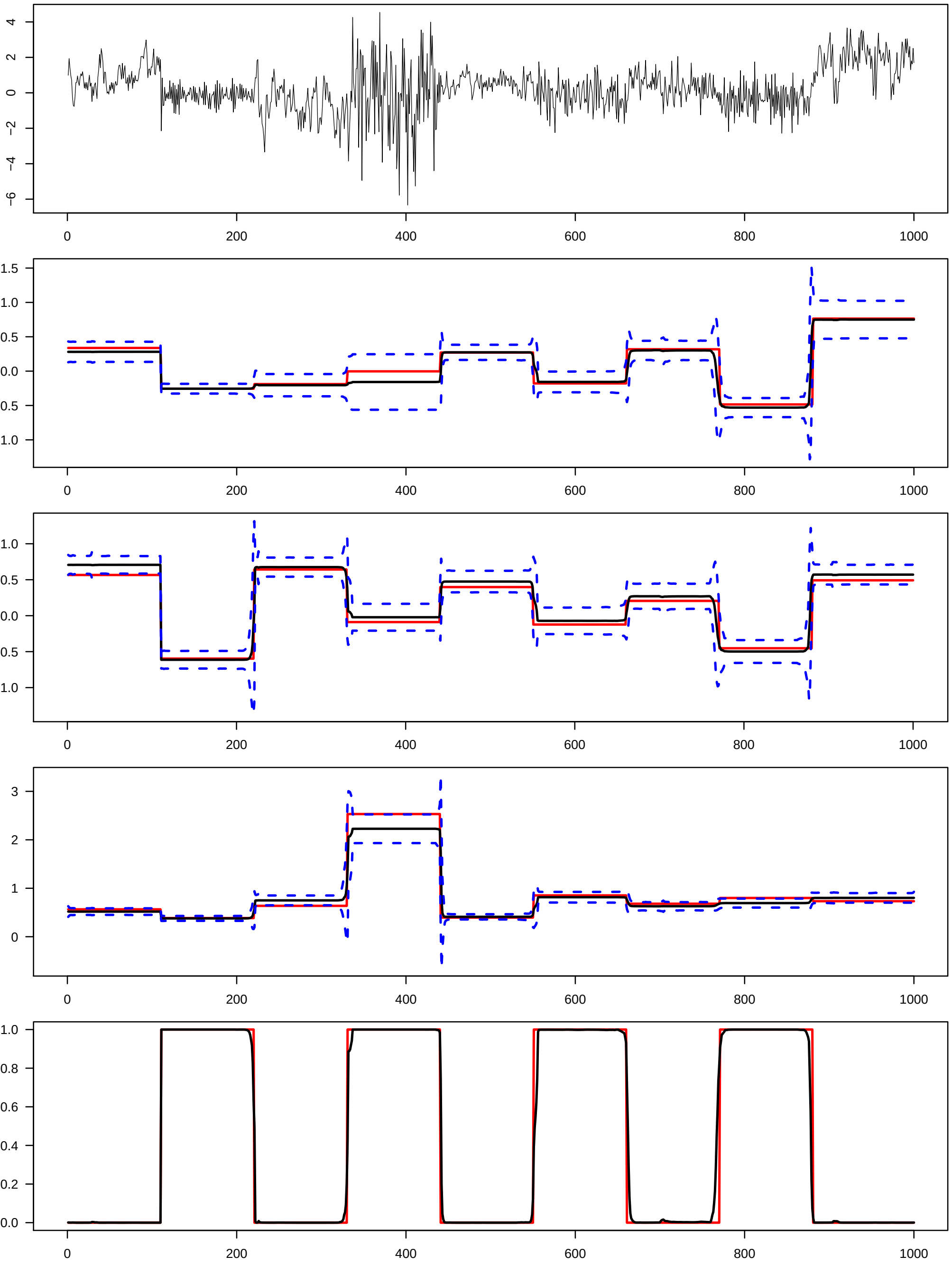
(e) Selected series y_t from scenario 5 (top), α_t (second), β_t (third), σ_t (fourth) and $P(S_t = 2)$ (bottom)

Figure 3.1: Continued



(f) Selected series y_t from scenario 6 (top), α_t (second), β_t (third), σ_t (fourth) and $P(S_t = 2)$ (bottom)

Figure 3.1: Continued



(g) Selected series y_t from scenario 7 (top), α_t (second), β_t (third), σ_t (fourth) and $P(S_t = 2)$ (bottom)

Table 3.2: Monte Carlo means of diagnostic statistics by Oracle and BCMIX methods. Standard errors are shown in parenthesis.

Scenarios	Oracle	BCMIX			
		(15, 10)	(20, 10)	(30, 15)	(40, 20)
Scenario 1	0.0031 (8.0e-05)	0.0044 (1.3e-04)	0.0044 (1.3e-04)	0.0044 (1.3e-04)	0.0044 (1.3e-04)
Scenario 2	0.0045 (9.4e-05)	0.0068 (2.0e-04)	0.0068 (2.0e-04)	0.0068 (2.0e-04)	0.0068 (2.0e-04)
Scenario 3	0.0061 (1.1e-04)	0.0102 (2.8e-04)	0.0102 (2.8e-04)	0.0102 (2.8e-04)	0.0102 (2.8e-04)
Scenario 4	0.0077 (1.3e-04)	0.0126 (3.5e-04)	0.0126 (3.5e-04)	0.0126 (3.5e-04)	0.0126 (3.5e-04)
Scenario 5	0.0093 (1.4e-04)	0.0155 (4.3e-04)	0.0155 (4.4e-04)	0.0154 (4.4e-04)	0.0154 (4.4e-04)
Scenario 6	0.0110 (1.5e-04)	0.0186 (5.7e-04)	0.0186 (5.7e-04)	0.0186 (5.7e-04)	0.0186 (5.7e-04)
Scenario 7	0.0140 (1.9e-04)	0.0237 (6.9e-04)	0.0236 (6.9e-04)	0.0236 (6.9e-04)	0.0235 (6.9e-04)

(a) Kullback-Leibler (KL) divergence

Scenarios	Oracle	BCMIX			
		(15, 10)	(20, 10)	(30, 15)	(40, 20)
Scenario 1	1.1060 (2.7e-01)	1.1047 (2.7e-01)	1.1046 (2.7e-01)	1.1045 (2.7e-01)	1.1044 (2.7e-01)
Scenario 2	0.8299 (2.1e-01)	0.8281 (2.1e-01)	0.8281 (2.1e-01)	0.8280 (2.1e-01)	0.8280 (2.1e-01)
Scenario 3	1.0344 (1.5e-01)	1.0323 (1.5e-01)	1.0322 (1.5e-01)	1.0322 (1.5e-01)	1.0321 (1.5e-01)
Scenario 4	0.9507 (1.3e-01)	0.9481 (1.3e-01)	0.9480 (1.3e-01)	0.9479 (1.3e-01)	0.9478 (1.3e-01)
Scenario 5	1.1145 (1.1e-01)	1.1113 (1.1e-01)	1.1112 (1.1e-01)	1.1112 (1.1e-01)	1.1111 (1.1e-01)
Scenario 6	0.8585 (5.9e-02)	0.8550 (5.9e-02)	0.8550 (5.9e-02)	0.8550 (5.9e-02)	0.8549 (5.9e-02)
Scenario 7	0.8858 (5.9e-02)	0.8809 (5.9e-02)	0.8809 (5.9e-02)	0.8809 (5.9e-02)	0.8808 (5.9e-02)

(b) Sum of squared Errors (SSE)

Table 3.2: Continued.

Scenarios	Oracle	BCMIX			
		(15, 10)	(20, 10)	(30, 15)	(40, 20)
Scenario 1	0.0509 (1.3e-03)	0.0522 (1.2e-03)	0.0522 (1.2e-03)	0.0523 (1.2e-03)	0.0524 (1.2e-03)
Scenario 2	0.0575 (1.1e-03)	0.0603 (1.1e-03)	0.0602 (1.1e-03)	0.0603 (1.1e-03)	0.0603 (1.1e-03)
Scenario 3	0.0707 (1.7e-03)	0.0740 (1.5e-03)	0.0740 (1.5e-03)	0.0740 (1.5e-03)	0.0741 (1.5e-03)
Scenario 4	0.0786 (1.8e-03)	0.0830 (1.6e-03)	0.0829 (1.6e-03)	0.0830 (1.6e-03)	0.0830 (1.6e-03)
Scenario 5	0.0866 (1.5e-03)	0.0912 (1.4e-03)	0.0912 (1.4e-03)	0.0912 (1.4e-03)	0.0912 (1.4e-03)
Scenario 6	0.0901 (1.1e-03)	0.0964 (1.1e-03)	0.0963 (1.1e-03)	0.0962 (1.1e-03)	0.0962 (1.1e-03)
Scenario 7	0.1041 (1.4e-03)	0.1105 (1.4e-03)	0.1102 (1.4e-03)	0.1101 (1.4e-03)	0.1101 (1.4e-03)

(c) L_2 norm

Scenarios	BCMIX			
	(15, 10)	(20, 10)	(30, 15)	(40, 20)
Scenario 1	0.9946 (2.3e-04)	0.9945 (1.9e-04)	0.9943 (1.9e-04)	0.9941 (2.0e-04)
Scenario 2	0.9911 (2.0e-04)	0.9911 (1.7e-04)	0.9910 (1.6e-04)	0.9909 (1.6e-04)
Scenario 3	0.9860 (2.1e-04)	0.9858 (2.0e-04)	0.9857 (2.0e-04)	0.9856 (2.0e-04)
Scenario 4	0.9813 (2.3e-04)	0.9810 (2.2e-04)	0.9809 (2.2e-04)	0.9809 (2.2e-04)
Scenario 5	0.9764 (2.6e-04)	0.9762 (2.3e-04)	0.9761 (2.3e-04)	0.9761 (2.3e-04)
Scenario 6	0.9716 (2.7e-04)	0.9715 (2.4e-04)	0.9714 (2.3e-04)	0.9714 (2.3e-04)
Scenario 7	0.9626 (2.9e-04)	0.9623 (2.6e-04)	0.9623 (2.5e-04)	0.9622 (2.5e-04)

(d) Sum of Squares of Standardized Errors (SSSE)

Table 3.2: Continued.

Scenarios	BCMIX			
	(15, 10)	(20, 10)	(30, 15)	(40, 20)
Scenario 1	0.9934 (2.3e-03)	0.9936 (2.3e-03)	0.9924 (2.5e-03)	0.9926 (2.5e-03)
Scenario 2	0.9933 (1.6e-03)	0.9927 (1.8e-03)	0.9941 (1.6e-03)	0.9944 (1.5e-03)
Scenario 3	0.9949 (8.5e-04)	0.9949 (8.7e-04)	0.9953 (7.1e-04)	0.9951 (7.5e-04)
Scenario 4	0.9946 (4.4e-04)	0.9942 (5.9e-04)	0.9938 (7.3e-04)	0.9934 (8.0e-04)
Scenario 5	0.9926 (7.1e-04)	0.9920 (8.6e-04)	0.9922 (8.0e-04)	0.9922 (8.1e-04)
Scenario 6	0.9906 (1.1e-03)	0.9911 (8.9e-04)	0.9918 (5.0e-04)	0.9918 (5.1e-04)
Scenario 7	0.9881 (7.2e-04)	0.9885 (6.6e-04)	0.9886 (6.3e-04)	0.9888 (6.0e-04)

(e) Identification Ratio (IR)

one regime to the other, the estimation shows a little fuzziness. In the bottom plot of Figure 3.1c, the estimated $\widehat{P}(S_t = 2)$ experiences changes from $0.005, 0.16, 0.25, 0.41, 0.71,$ to 0.95 for time $t = 246, 247, 248, 249, 250$ and 251. It takes five time steps for the estimation to correctly identify the change point. This phenomenon is prominent in the estimations of all other parameters. But in general the estimation adjusts quickly to the change. Misclassification around the transition areas may also explain why IR statistics are always less than 1 in Table 3.2e.

In the series where there are more transition points (Figure 3.1d to 3.1g), the estimates are still close to their true corresponding parameters and the confidence intervals include the true parameters most of the time. Although the more transition points are in a fixed length series, the less accurate the estimation is, usually accompanied by more fuzziness around the transition points and wider confidence intervals.

3.4 Simulation with Large Series

To continue with the AR(1) model defined in (3.7), this section explores the simulation with longer series and with stochastic change points generated by specified Markov transition probabilities. The true simulation parameters $\boldsymbol{z}$, $\boldsymbol{V}, \boldsymbol{g}$ and $\boldsymbol{\lambda}$ are the same as used for model (3.7) in Section 3.3, except for the Markov probability transition matrix P, defined as $P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$ for a two-regime system. p is the probability of making transition from state 1 to state 2; q is the reverse. For every pair of $(p,\ q)$ in the following scenarios, 500 series are simulated for series lengths equal to 2000, 3000, 4000 and 5000 respectively.

Scenario 1 $p = 0.002,\ q = 0.002$

Scenario 2 $p = 0.004,\ q = 0.004$

Scenario 3 $p = 0.008,\ q = 0.008$

Scenario 4 $p = 0.016,\ q = 0.016$

Scenario 5 $p = 0.016,\ q = 0.032$

Scenario 6 $p = 0.032,\ q = 0.032$

Scenario 7 $p = 0.05,\ q = 0.05$

Scenario 8 $p = 0.075,\ q = 0.075$

Scenario 9 $p = 0.1,\ q = 0.1$

Scenario 10 $p = 0.2,\ q = 0.2$

The number of regime changes in each simulated series is regulated by p and q. Unlike previous simulations, the number of regime changes is stochastic and the positions of the change points are also random. The larger the transition probabilities are, the more transition points are anticipated in a simulated series. The estimation starts with the same initial prior as in Section 3.3 and proceed with the estimation of forward, backward, and smoothing recursive weights by BCMIX (20, 10) methods and then estimate hyperparameters via EM algorithm. This process repeats until EM algorithm converges. At the end of the last EM iteration, we compute the estimated $\boldsymbol{\beta}_t$, σ_t and probability of the regime at every time point. In principle EM algorithm should be run until convergence, whereas in this simulation EM algorithm is run once by choosing good initial prior values. The choice of initial prior values will be fully discussed in the next section.

The major issue encountered in long series estimation is the computation cost. In fact selecting an example model in the entire simulation studies has taken the computational resources and time into consideration. On the one hand, the theoretical settings need to be complex and comprehensive enough to compare and contrast methods; on the other hand the computational resources are utilized at their (almost) optimal conditions. This is also the reason we stick with AR(1) model, although higher order AR models or models involving exogenous variables are quite feasible to implement, but a little challenging for intensive simulation on long series. Computational issues encountered in this thesis have been discussed in Section 2.8.

3.4.1 Goodness of Fit of BCMIX Method

The goal of this subsection is to understand the impact of the series length and the magnitude of transition probabilities on the estimation of model parameters by analyzing KL, SSSE, L_2 and IR statistics . In Table 3.3a, it is obvious that for a fixed series length, both the means and standard errors of KL statistics increase when the transition probabilities increase (larger p and q pairs). Likewise, for a fixed transition probability pair (a particular scenario), the means of KL statistics have a tendency to decrease when the series become longer for the first 5 scenarios and show a little up and down patterns as series length increases for the last 5 scenarios. But when the series reach 4000 and 5000 long, the differences of means are only 1 to 6 units at the order of 10^{-4}. The standard errors of KL statistics decrease as the series length increases from 2000 to 5000 across all scenarios, indicating that the estimation becomes stabilized when the series becomes larger.

The means of L_2 statistics in Table 3.3c increase as (p, q) pairs become larger in each column. Row-wise, they may have a tendency to decrease in most scenarios, but may stay flat or rise up a little for some scenarios. The standard errors become larger for series making more transitions within a fixed series length and become smaller for longer series within each scenario. The highest standard error is located at the upper left corner of the table and the lowest standard error at bottom right corner. The means of SSSE statistics in Table 3.3b decrease and standard errors increase as the number of transition points increase for a particular series length. Since this statistic is an estimation of the variance of the standard normal distribution, values closer to 1 indicate better fit. Under each scenario, the mean statistics decrease first and go up a little again, but their values remain almost the same up to the third digit after the decimal point. The standard errors of SSSE statistics do decrease as series length increases for every scenario.

The identification ratio (IR) in Table 3.3d signals no significant improvement as the series length increases. This statistic remains as high as 99% for scenario 1 to 6 for all series length. The standard errors tend to decrease within each row. Clearly, in each column of Table 3.3d, the mean statistics decrease when there are more transition points in the series. Up to scenario 6 ($p = 0.032, q = 0.032$), the model correctly identifies the regime status 99% of the time. Even in the worst case (scenario 10) the model can still classify $92\% - 93\%$ of the regimes correctly. Interesting though, the standard errors go down first till scenario 5 or 6 and then up a little again within each column, showing that model may reach its peak performance at a moderate large number of change points in a series.

Overall, the model behaves more stable when the series is longer and estimates better for longer series in some cases, but series length appears not to be a deciding advantage for using this model. The clear message

Table 3.3: Monte Carlo means of diagnostic statistics by BCMIX (20, 10) method. Standard errors are shown in parenthesis.

(p, q)	2000	3000	4000	5000
(0.002, 0.002)	0.0037(1.3e-4)	0.0034(1.0e-4)	0.0033(7.8e-5)	0.0031(7.2e-5)
(0.004, 0.004)	0.0053(1.4e-4)	0.0052(1.4e-4)	0.0050(1.1e-4)	0.0052(9.9e-5)
(0.008, 0.008)	0.0094(2.2e-4)	0.0092(1.8e-4)	0.0092(1.5e-4)	0.0089(1.4e-4)
(0.016, 0.016)	0.0134(3.0e-4)	0.0129(1.9e-4)	0.0125(1.6e-4)	0.0124(1.3e-4)
(0.016, 0.032)	0.0183(2.9e-4)	0.0187(2.5e-4)	0.0182(2.0e-4)	0.0181(1.7e-4)
(0.032, 0.032)	0.0249(3.8e-4)	0.0241(2.5e-4)	0.0237(2.2e-4)	0.0243(2.3e-4)
(0.050, 0.050)	0.0345(4.4e-4)	0.0342(3.6e-4)	0.0337(2.6e-4)	0.0338(2.5e-4)
(0.075, 0.075)	0.0514(4.6e-4)	0.0522(4.5e-4)	0.0511(3.4e-4)	0.0517(3.2e-4)
(0.100, 0.100)	0.0937(7.2e-4)	0.0922(5.9e-4)	0.0920(5.0e-4)	0.0922(4.3e-4)
(0.200, 0.200)	0.1322(7.2e-4)	0.1359(6.7e-4)	0.1366(6.1e-4)	0.1363(5.2e-4)

(a) Kullback-Leibler (KL) Divergence

(p, q)	2000	3000	4000	5000
(0.002, 0.002)	0.9947(1.9e-4)	0.9944(1.6e-4)	0.9941(1.2e-4)	0.9943(1.0e-4)
(0.004, 0.004)	0.9918(1.8e-4)	0.9912(1.4e-4)	0.9907(1.1e-4)	0.9906(9.2e-5)
(0.008, 0.008)	0.9839(1.9e-4)	0.9824(1.9e-4)	0.9824(1.8e-4)	0.9830(1.3e-4)
(0.016, 0.016)	0.9769(1.8e-4)	0.9756(1.7e-4)	0.9759(1.4e-4)	0.9761(1.2e-4)
(0.016, 0.032)	0.9664(2.3e-4)	0.9648(2.1e-4)	0.9654(1.7e-4)	0.9653(1.6e-4)
(0.032, 0.032)	0.9551(2.5e-4)	0.9537(2.2e-4)	0.9538(1.8e-4)	0.9539(1.6e-4)
(0.050, 0.050)	0.9371(2.8e-4)	0.9348(2.4e-4)	0.9353(2.1e-4)	0.9349(2.0e-4)
(0.075, 0.075)	0.9015(3.6e-4)	0.9012(2.9e-4)	0.9012(2.7e-4)	0.9014(2.4e-4)
(0.100, 0.100)	0.8284(4.6e-4)	0.8278(4.0e-4)	0.8292(3.4e-4)	0.8283(3.1e-4)
(0.200, 0.200)	0.7655(4.0e-4)	0.7568(3.9e-4)	0.7570(3.4e-4)	0.7570(3.2e-4)

(b) Sum of Squares of Standardized Errors (SSSE)

Table 3.3: Continued

(p, q)	2000	3000	4000	5000
(0.002, 0.002)	0.0410(8.7e-4)	0.0395(7.0e-4)	0.0390(5.9e-4)	0.0374(5.0e-4)
(0.004, 0.004)	0.0507(8.1e-4)	0.0504(7.1e-4)	0.0486(5.4e-4)	0.0493(5.3e-4)
(0.008, 0.008)	0.0642(7.3e-4)	0.0668(6.7e-4)	0.0657(5.8e-4)	0.0649(4.9e-4)
(0.016, 0.016)	0.0787(8.5e-4)	0.0787(6.2e-4)	0.0777(5.5e-4)	0.0769(4.7e-4)
(0.016, 0.032)	0.0940(8.0e-4)	0.0944(6.1e-4)	0.0917(5.5e-4)	0.0926(4.8e-4)
(0.032, 0.032)	0.1082(8.0e-4)	0.1073(6.2e-4)	0.1057(5.3e-4)	0.1066(5.0e-4)
(0.050, 0.050)	0.1278(7.8e-4)	0.1270(6.5e-4)	0.1251(5.3e-4)	0.1266(4.5e-4)
(0.075, 0.075)	0.1567(7.8e-4)	0.1560(6.3e-4)	0.1544(5.3e-4)	0.1548(4.9e-4)
(0.100, 0.100)	0.2067(7.2e-4)	0.2064(6.4e-4)	0.2049(5.1e-4)	0.2058(4.6e-4)
(0.200, 0.200)	0.2505(7.0e-4)	0.2519(5.6e-4)	0.2504(5.0e-4)	0.2516(4.4e-4)

(c) L_2 norm

(p, q)	2000	3000	4000	5000
(0.002, 0.002)	0.9961(1.5e-3)	0.9961(1.4e-3)	0.9976(5.4e-4)	0.9968(1.0e-3)
(0.004, 0.004)	0.9936(2.3e-3)	0.9973(6.1e-4)	0.9966(8.7e-4)	0.9973(4.8e-4)
(0.008, 0.008)	0.9947(8.4e-4)	0.9889(1.6e-3)	0.9927(1.2e-3)	0.9954(6.1e-4)
(0.016, 0.016)	0.9937(5.5e-4)	0.9931(9.3e-4)	0.9946(2.8e-4)	0.9945(2.9e-4)
(0.016, 0.032)	0.9926(3.5e-4)	0.9919(3.7e-4)	0.9918(3.6e-4)	0.9921(2.8e-4)
(0.032, 0.032)	0.9890(4.3e-4)	0.9886(4.5e-4)	0.9889(3.3e-4)	0.9893(2.5e-4)
(0.050, 0.050)	0.9840(5.0e-4)	0.9834(4.5e-4)	0.9841(3.7e-4)	0.9838(3.2e-4)
(0.075, 0.075)	0.9742(5.1e-4)	0.9736(4.5e-4)	0.9738(4.2e-4)	0.9737(3.6e-4)
(0.100, 0.100)	0.9505(6.5e-4)	0.9501(5.4e-4)	0.9503(4.5e-4)	0.9491(4.1e-4)
(0.200, 0.200)	0.9276(6.5e-4)	0.9249(5.4e-4)	0.9252(5.0e-4)	0.9243(4.1e-4)

(d) Identification Ratio (IR)

conveyed by this simulation study is that larger transition probabilities seem to jeopardize all the evaluation metrics. As far as this simulation study is concerned, the model may be best suited for a Markov chain with moderate large transition probabilities. When it comes to the comparison with existing research models, this model still perform better than the traditional Markov switching model in the cases such as scenario 9 or 10, which will be discussed in the end of this section.

3.4.2 Analysis of Hyperparameter Estimation

Table 3.4 and Table 3.5 show the means and standard errors of hyperparameters from 500 simulations for each scenario and each series length. Among all the estimations, the most interesting and widely concerned statistics are Markov chain transition probabilities, i.e. p and q in this simulation study. In Table 3.4d, regardless of the series length, the estimated p and q are almost the same under each scenario. For example in scenario 1, p's are estimated to be 0.002, exactly the same as their true prior counterpart for all series lengths; in scenario 2 q's are estimated to be 0.003 or 0.004 a little smaller than the true prior 0.004, but the estimation is consistent across series length. The estimation also appears to be reasonable in a way that estimated p and q become larger as the true transition probabilities become larger within a fixed series length. It may appear that on average the model has a tendency to underestimate the transition probabilities. For example, in scenario $(0.032, 0.032)$, the estimated (p, q)'s are only $(0.009, 0.009)$ or $(0.01, 0.01)$; in scenario $(0.2, 0.2)$, the estimated p and q's are about 0.066 or 0.067. However, the underestimation is spurious and the reason is explained by the mechanism to generate the series.

For a first order Markov chain, it is not difficult to prove that the expected number of transition points from state 1 to state 2 is simply the number of cases that the series at state 2 at time t and at state 1 at time $t-1$. Let X be the number of such transitions, then $X = \sum_{t=1}^{T} I_{\{S_t=2, S_{t-1}=1\}}$, the expected value of X is

$$EX = \sum_{t=1}^{T} P(S_t = 2, S_t = 1) = \sum_{t=1}^{T} pP(S_{t-1} = 1) \approx \frac{Tpq}{p+q} \qquad (3.8)$$

Approximation occurs due to ignoring the initial state. Similarly, the expected number of transitions from state 2 to state 1 can be proven to be the same. The total number of expected transitions in a series is the sum of the two values.

In this simulation study, the series is controlled to stay at one regime for at least 10 time points before it is allowed to move on to another regime. This intervention seems to be realistic, because economic interruption is not likely to occur at very adjacent periods. The controlled simulation scheme significantly lowers the number of change points for each scenario than the theory would predict. Table 3.6 computes the expected number of change points based on equation (3.8) and the practical average number of change points in the simulated series for every scenario and every $T = 2000, 3000, 4000$ and 5000. For scenario $(0.05, 0.05)$ of $T = 2000$, the expected number of change points is 100, whereas in practice there are roughly 25 changes on average. The estimated (p, q) is $(0.013, 0.012)$ reflecting the fact that 25% of the expected change points correspond to the estimation of 25% of the theoretical probability 0.05. Viewing Table 3.6 in another way, in series $T = 4000$ and scenario $(0.1, 0.1)$, the observed average number of change points is 137.4 (Table 3.6b), a little higher than 128 in $T = 4000$ and scenario $(0.032, 0.032)$ (Table 3.6a). The estimated (p, q) is $(0.043, 0.042)$ (Table 3.4d) for scenario $(0.05, 0.05)$ reflects the fact that the series are actually generated from lower probability scenario. In general, the estimation of (p, q) is very good, reflecting the true nature of Markov chain. It is worth mentioning earlier that the traditional Markov switching models tend to overestimate the transition probabilities. A comparison analysis will be given in the next subsection.

For completeness, the estimation of other hyperparameters is also provided in Table 3.4. A separate section will focus on the estimation of prior parameters, so the discussion here may fall short. $\boldsymbol{g}$ seems to be exaggeratively overestimated for shorter series and low transition probabilities. In the line of the first scenario, $g^{(1)}$ is estimated to be between 21.72 for $T = 5000$ and 168.14 for $T = 2000$, when the true value of $g^{(1)}$ is 2.5. But as the transition probabilities become larger, the estimation become stabilized and the estimates seem to converge to the values not far from their true priors. At the bottom line of Table 3.4a, $g^{(1)}$ converges to 2.6 or 2.7 and $g^{(2)}$ converges to about 1.5 and their true values are 2.5 and 1.2 respectively. $\boldsymbol{\lambda}$ estimates in Table 3.4b behave similarly, but the values are moving in a different direction. They begin small for shorter series with low transition probabilities and gradually increase and then become stable for larger transition probability scenarios. $\lambda^{(1)}$ converges to a value between 1.2 and 1.3 while the true prior is 0.8; and $\lambda^{(2)}$ converges to around 0.7 or 0.8 when the true value is 1. $\boldsymbol{z}$

of Table 3.4c seems to be the only estimates that are genuine reflection of the true prior values for every transition probability scenario and every series length. In Table 3.5 $\widehat{V}_{11}^{(k)}$'s are small in low probability scenario and increase and stabilize around 0.15 or 0.16, close to the true value 0.16. $\widehat{V}_{22}^{(1)}$ converges to around 0.15 and $\widehat{V}_{22}^{(2)}$ converges to 0.09 when both true values are 0.16. The estimation on the off diagonal of $\widehat{V}^{(k)}$ has no major issues since these values are small and close to 0, the true prior value.

The quick observation from Table 3.4 and 3.5 does not indicate a relationship between performance of the hyperparameter estimation and series length. As transition probabilities become larger, however, the estimation seems to be more reasonable, particular for Table 3.4a and 3.4b. Except for the estimation of (p, q) pair which is of primary interest, there is no conclusive criterion to evaluate the prior estimation. The fact that the hyperparameter estimation does not converge to the true value in the simulation does not necessarily indicate a modeling failure. The major concern is the collective effect of hyperparameters on the estimation of regression parameters. In a hyperplane, a good combination of hyperparameters may result in good estimation about model parameters. See Section 3.5.2 for more discussions about the effect of hyperparameter estimation on model parameters.

Table 3.4: Monte Carlo means of the estimated hyperparameters via EM algorithm. Standard errors are shown in parenthesis.

(p, q)	2000		3000		4000		5000	
	$\hat{g}_1$	$\hat{g}_2$	$\hat{g}_1$	$\hat{g}_2$	$\hat{g}_1$	$\hat{g}_2$	$\hat{g}_1$	$\hat{g}_2$
(0.002, 0.002)	168.14	177.28	77.97	112.91	28.28	40.69	21.72	27.51
	(1.1e+1)	(8.0)	(8.2)	(7.9)	(3.8)	(4.8)	(2.3)	(5.6)
(0.004, 0.004)	40.01	55.13	15.32	18.58	9.02	5.91	5.66	3.65
	(4.3)	(4.4)	(1.4)	(2.1)	(7.0e-1)	(5.0e-1)	(3.1e-1)	(2.6e-1)
(0.008, 0.008)	8.49	7.82	5.76	4.17	4.52	2.75	3.87	2.19
	(5.6e-1)	(9.1e-1)	(3.0e-1)	(3.1e-1)	(1.7e-1)	(1.0e-1)	(1.2e-1)	(6.0e-2)
(0.016, 0.016)	5.65	3.25	4.51	2.45	3.63	2.11	3.24	1.89
	(2.7e-1)	(1.5e-1)	(1.7e-1)	(9.6e-2)	(1.0e-1)	(5.8e-2)	(7.6e-2)	(3.9e-2)
(0.016, 0.032)	3.96	2.33	3.49	2.03	3.05	1.80	2.99	1.68
	(1.3e-1)	(8.8e-2)	(9.0e-2)	(5.2e-2)	(5.8e-2)	(3.7e-2)	(5.3e-2)	(2.7e-2)
(0.032, 0.032)	3.55	1.98	3.10	1.79	2.94	1.68	2.82	1.60
	(8.8e-2)	(4.8e-2)	(6.4e-2)	(3.0e-2)	(5.1e-2)	(2.5e-2)	(3.9e-2)	(2.1e-2)
(0.050, 0.050)	3.24	1.78	2.89	1.64	2.76	1.64	2.68	1.54
	(6.3e-2)	(3.9e-2)	(4.5e-2)	(2.1e-2)	(3.6e-2)	(2.1e-2)	(3.0e-2)	(1.5e-2)
(0.075, 0.075)	2.99	1.54	2.81	1.47	2.69	1.54	2.64	1.51
	(4.3e-2)	(2.0e-2)	(3.1e-2)	(1.5e-2)	(2.7e-2)	(1.4e-2)	(2.3e-2)	(1.2e-2)
(0.100, 0.100)	2.82	1.40	2.68	1.52	2.70	1.51	2.62	1.48
	(2.6e-2)	(1.2e-2)	(2.1e-2)	(1.1e-2)	(1.9e-2)	(1.0e-2)	(1.5e-2)	(8.6e-3)
(0.200, 0.200)	2.81	1.38	2.68	1.52	2.68	1.49	2.63	1.48
	(1.9e-2)	(8.7e-3)	(1.5e-2)	(8.8e-3)	(1.5e-2)	(7.4e-3)	(1.3e-2)	(6.5e-3)

(a) True $g^{(1)} = 2.5$ and $g^{(2)} = 1.2$

(p, q)	2000		3000		4000		5000	
	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_1$	$\hat{\lambda}_2$
(0.002, 0.002)	0.386	0.076	0.521	0.210	0.757	0.334	0.805	0.425
	(3.8e-2)	(9.6e-3)	(3.3e-2)	(2.0e-2)	(4.1e-2)	(2.3e-2)	(4.3e-2)	(2.8e-2)
(0.004, 0.004)	0.632	0.250	0.816	0.408	0.871	0.495	1.021	0.586
	(3.7e-2)	(2.1e-2)	(3.8e-2)	(2.5e-2)	(3.7e-2)	(2.9e-2)	(3.5e-2)	(2.6e-2)
(0.008, 0.008)	0.891	0.479	1.019	0.528	1.039	0.536	1.105	0.617
	(3.6e-2)	(2.6e-2)	(3.4e-2)	(2.3e-2)	(2.9e-2)	(1.8e-2)	(2.8e-2)	(2.0e-2)
(0.016, 0.016)	0.976	0.547	1.017	0.597	1.129	0.622	1.211	0.629
	(3.2e-2)	(2.2e-2)	(2.8e-2)	(2.0e-2)	(2.5e-2)	(1.7e-2)	(2.5e-2)	(1.5e-2)
(0.016, 0.032)	1.102	0.627	1.125	0.621	1.222	0.669	1.206	0.678
	(2.7e-2)	(2.1e-2)	(2.3e-2)	(1.8e-2)	(2.1e-2)	(1.4e-2)	(2.0e-2)	(1.4e-2)
(0.032, 0.032)	1.106	0.651	1.211	0.644	1.237	0.670	1.251	0.682
	(2.3e-2)	(1.7e-2)	(2.2e-2)	(1.4e-2)	(2.0e-2)	(1.2e-2)	(1.7e-2)	(1.1e-2)
(0.050, 0.050)	1.113	0.698	1.229	0.663	1.285	0.660	1.272	0.690
	(2.0e-2)	(1.6e-2)	(1.8e-2)	(1.2e-2)	(1.6e-2)	(1.1e-2)	(1.4e-2)	(9.4e-3)
(0.075, 0.075)	1.140	0.751	1.192	0.768	1.282	0.685	1.281	0.693
	(1.6e-2)	(1.2e-2)	(1.3e-2)	(1.0e-2)	(1.4e-2)	(8.3e-3)	(1.1e-2)	(7.6e-3)
(0.100, 0.100)	1.138	0.827	1.279	0.696	1.258	0.680	1.288	0.700
	(1.0e-2)	(8.9e-3)	(1.1e-2)	(7.0e-3)	(8.8e-3)	(6.3e-3)	(8.4e-3)	(5.9e-3)
(0.200, 0.200)	1.083	0.845	1.267	0.682	1.272	0.690	1.284	0.696
	(7.4e-3)	(7.1e-3)	(8.0e-3)	(5.4e-3)	(7.6e-3)	(4.9e-3)	(7.3e-3)	(4.7e-3)

(b) True $\lambda^{(1)} = 0.8$ and $\lambda^{(2)} = 1$

Table 3.4: Continued

(p, q)		2000		3000		4000		5000	
(0.002, 0.002)	$z^{(1)}$	0.3018	0.5060	0.2967	0.4987	0.3096	0.4959	0.3081	0.4971
		(6.9e-3)	(6.5e-3)	(5.8e-3)	(5.8e-3)	(4.9e-3)	(4.8e-3)	(4.6e-3)	(4.6e-3)
	$z^{(2)}$	-0.2289	-0.4608	-0.2318	-0.4365	-0.2530	-0.4608	-0.2377	-0.4672
		(2.0e-2)	(1.4e-2)	(1.5e-2)	(1.1e-2)	(1.0e-2)	(9.3e-3)	(9.3e-3)	(8.2e-3)
(0.004, 0.004)	$z^{(1)}$	0.3072	0.5055	0.3119	0.5027	0.3065	0.4970	0.3136	0.5007
		(5.5e-3)	(5.1e-3)	(4.6e-3)	(4.3e-3)	(4.0e-3)	(3.7e-3)	(3.5e-3)	(3.4e-3)
	$z^{(2)}$	-0.2652	-0.4505	-0.2421	-0.4426	-0.2526	-0.4685	-0.2469	-0.4638
		(1.3e-2)	(1.0e-2)	(9.6e-3)	(8.6e-3)	(8.1e-3)	(7.1e-3)	(6.6e-3)	(6.0e-3)
(0.008, 0.008)	$z^{(1)}$	0.3018	0.4927	0.3099	0.5038	0.3100	0.4982	0.3129	0.5053
		(4.0e-3)	(3.9e-3)	(3.3e-3)	(3.4e-3)	(2.9e-3)	(2.9e-3)	(2.6e-3)	(2.6e-3)
	$z^{(2)}$	-0.2339	-0.4520	-0.2403	-0.4837	-0.2463	-0.4762	-0.2425	-0.4732
		(8.1e-3)	(7.4e-3)	(6.3e-3)	(5.2e-3)	(5.5e-3)	(4.7e-3)	(4.9e-3)	(4.5e-3)
(0.016, 0.016)	$z^{(1)}$	0.3026	0.4919	0.3047	0.5057	0.3096	0.4982	0.3125	0.5050
		(3.2e-3)	(3.1e-3)	(2.8e-3)	(2.5e-3)	(2.3e-3)	(2.2e-3)	(2.1e-3)	(2.0e-3)
	$z^{(2)}$	-0.2428	-0.4637	-0.2470	-0.4759	-0.2285	-0.4743	-0.2355	-0.4791
		(6.3e-3)	(5.4e-3)	(5.1e-3)	(4.3e-3)	(4.2e-3)	(3.5e-3)	(3.6e-3)	(3.4e-3)
(0.016, 0.032)	$z^{(1)}$	0.3076	0.4923	0.3084	0.5032	0.3109	0.5021	0.3146	0.5030
		(2.5e-3)	(2.7e-3)	(2.2e-3)	(2.0e-3)	(1.8e-3)	(1.8e-3)	(1.7e-3)	(1.7e-3)
	$z^{(2)}$	-0.2515	-0.4561	-0.2422	-0.4767	-0.2386	-0.4825	-0.2332	-0.4773
		(4.9e-3)	(4.4e-3)	(4.0e-3)	(3.4e-3)	(3.3e-3)	(3.1e-3)	(2.9e-3)	(2.6e-3)
(0.032, 0.032)	$z^{(1)}$	0.3033	0.4909	0.3096	0.5042	0.3119	0.4990	0.3102	0.5030
		(2.2e-3)	(2.2e-3)	(1.8e-3)	(1.7e-3)	(1.5e-3)	(1.5e-3)	(1.4e-3)	(1.3e-3)
	$z^{(2)}$	-0.2421	-0.4607	-0.2439	-0.4863	-0.2373	-0.4795	-0.2343	-0.4791
		(4.1e-3)	(3.7e-3)	(3.4e-3)	(3.1e-3)	(2.8e-3)	(2.5e-3)	(2.5e-3)	(2.2e-3)
(0.050, 0.050)	$z^{(1)}$	0.3031	0.4909	0.3094	0.5023	0.3099	0.5010	0.3135	0.5016
		(1.8e-3)	(1.8e-3)	(1.5e-3)	(1.4e-3)	(1.3e-3)	(1.2e-3)	(1.2e-3)	(1.2e-3)
	$z^{(2)}$	-0.2344	-0.4598	-0.2374	-0.4798	-0.2377	-0.4793	-0.2363	-0.4798
		(3.3e-3)	(3.2e-3)	(2.7e-3)	(2.2e-3)	(2.3e-3)	(2.1e-3)	(2.1e-3)	(1.8e-3)
(0.075, 0.075)	$z^{(1)}$	0.2985	0.4808	0.3002	0.4807	0.3133	0.4989	0.3115	0.5033
		(1.4e-3)	(1.4e-3)	(1.2e-3)	(1.1e-3)	(1.1e-3)	(9.8e-4)	(9.3e-4)	(9.3e-4)
	$z^{(2)}$	-0.2317	-0.4574	-0.2326	-0.4578	-0.2402	-0.4828	-0.2382	-0.4782
		(2.6e-3)	(2.2e-3)	(2.0e-3)	(1.8e-3)	(1.8e-3)	(1.6e-3)	(1.7e-3)	(1.4e-3)
(0.100, 0.100)	$z^{(1)}$	0.2910	0.4699	0.3109	0.5022	0.3140	0.5008	0.3162	0.5006
		(8.8e-4)	(9.4e-4)	(9.0e-4)	(9.4e-4)	(8.2e-4)	(8.2e-4)	(7.5e-4)	(7.0e-4)
	$z^{(2)}$	-0.2314	-0.4487	-0.2389	-0.4801	-0.2442	-0.4788	-0.2433	-0.4795
		(1.7e-3)	(1.5e-3)	(1.6e-3)	(1.3e-3)	(1.4e-3)	(1.2e-3)	(1.2e-3)	(1.1e-3)
(0.200, 0.200)	$z^{(1)}$	0.2807	0.4633	0.3112	0.5033	0.3128	0.5031	0.3146	0.5028
		(7.4e-4)	(7.3e-4)	(8.1e-4)	(8.7e-4)	(7.1e-4)	(7.9e-4)	(7.2e-4)	(7.6e-4)
	$z^{(2)}$	-0.2265	-0.4374	-0.2432	-0.4824	-0.2420	-0.4779	-0.2419	-0.4748
		(1.2e-3)	(1.2e-3)	(1.2e-3)	(1.2e-3)	(1.0e-3)	(1.1e-3)	(9.7e-4)	(9.6e-4)

(c) True $\boldsymbol{z}^{(1)} = (0.3,\ 0.5)$ and $\boldsymbol{z}^{(2)} = (-0.2,\ -0.5)$

Table 3.4: Continued

(p, q)	2000		3000		4000		5000	
	$\hat{p}$	$\hat{q}$	$\hat{p}$	$\hat{q}$	$\hat{p}$	$\hat{q}$	$\hat{p}$	$\hat{q}$
(0.002, 0.002)	0.002	0.011	0.002	0.004	0.002	0.003	0.002	0.003
	(5.7e-5)	(2.0e-3)	(6.8e-5)	(8.1e-4)	(6.1e-5)	(7.0e-5)	(5.9e-5)	(1.0e-4)
(0.004, 0.004)	0.003	0.004	0.003	0.003	0.003	0.003	0.003	0.003
	(7.2e-5)	(4.0e-4)	(4.6e-5)	(7.0e-5)	(6.7e-5)	(5.0e-5)	(4.0e-5)	(4.1e-5)
(0.008, 0.008)	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004
	(6.4e-5)	(7.9e-5)	(6.2e-5)	(1.1e-4)	(4.0e-5)	(7.2e-5)	(5.0e-5)	(4.7e-5)
(0.016, 0.016)	0.006	0.005	0.005	0.005	0.005	0.005	0.005	0.005
	(5.3e-5)	(6.7e-5)	(4.8e-5)	(6.0e-5)	(3.9e-5)	(4.4e-5)	(3.3e-5)	(4.1e-5)
(0.016, 0.032)	0.007	0.008	0.007	0.008	0.007	0.008	0.007	0.008
	(5.0e-5)	(5.7e-5)	(4.7e-5)	(5.7e-5)	(4.0e-5)	(4.4e-5)	(3.6e-5)	(4.6e-5)
(0.032, 0.032)	0.009	0.009	0.010	0.010	0.010	0.010	0.010	0.010
	(5.2e-5)	(5.8e-5)	(4.9e-5)	(6.3e-5)	(4.5e-5)	(4.8e-5)	(4.1e-5)	(4.2e-5)
(0.050, 0.050)	0.013	0.012	0.014	0.014	0.014	0.014	0.014	0.014
	(6.3e-5)	(6.3e-5)	(6.4e-5)	(6.9e-5)	(5.5e-5)	(6.1e-5)	(4.8e-5)	(5.4e-5)
(0.075, 0.075)	0.024	0.024	0.024	0.024	0.022	0.022	0.022	0.022
	(8.7e-5)	(8.1e-5)	(6.9e-5)	(7.0e-5)	(6.5e-5)	(6.6e-5)	(6.1e-5)	(6.0e-5)
(0.100, 0.100)	0.044	0.045	0.043	0.043	0.043	0.042	0.043	0.042
	(1.3e-4)	(1.2e-4)	(1.2e-4)	(1.2e-4)	(9.7e-5)	(9.9e-5)	(8.6e-5)	(9.1e-5)
(0.200, 0.200)	0.057	0.058	0.066	0.067	0.066	0.067	0.066	0.067
	(1.1e-4)	(1.2e-4)	(1.9e-4)	(1.9e-4)	(1.7e-4)	(1.7e-4)	(1.7e-4)	(1.8e-4)

(d) True (p, q) pairs are indicated in the first column

Table 3.5: Monte Carlo mean of estimated $\hat{\boldsymbol{V}}$ by EM algorithm. Standard error are shown in parenthesis. True $v_{11}^{(k)} = v_{22}^{(k)} = 0.16$ and $v_{12}^{(k)} = v_{21}^{(k)} = 0$, for $k = 1, 2$.

(p, q)		2000		3000		4000		5000	
(0.002, 0.002)	$V^{(1)}$	0.0452	-0.0041	0.0696	-0.0025	0.0781	-0.0013	0.0943	-0.0015
		(3.7e-3)	(1.1e-3)	(4.2e-3)	(1.5e-3)	(3.5e-3)	(1.3e-3)	(4.9e-3)	(1.3e-3)
		-0.0041	0.0480	-0.0025	0.0703	-0.0013	0.0817	-0.0015	0.0846
		(1.1e-3)	(3.6e-3)	(1.5e-3)	(4.1e-3)	(1.3e-3)	(3.6e-3)	(1.3e-3)	(3.3e-3)
	$V^{(2)}$	0.0204	0.0006	0.0521	-0.0008	0.0847	-0.0007	0.0939	0.0026
		(2.3e-3)	(4.6e-4)	(4.2e-3)	(1.1e-3)	(4.9e-3)	(1.4e-3)	(5.4e-3)	(1.3e-3)
		0.0006	0.0171	-0.0008	0.0325	-0.0007	0.0544	0.0026	0.0648
		(4.6e-4)	(1.7e-3)	(1.1e-3)	(2.6e-3)	(1.4e-3)	(3.0e-3)	(1.3e-3)	(3.5e-3)

(0.004, 0.004)	$V^{(1)}$	0.0799	-0.0040	0.1023	-0.0007	0.1105	-0.0034	0.1134	-0.0039
		(4.5e-3)	(1.4e-3)	(4.3e-3)	(1.4e-3)	(3.9e-3)	(1.3e-3)	(3.7e-3)	(1.3e-3)
		-0.0040	0.0837	-0.0007	0.0978	-0.0034	0.1052	-0.0039	0.1187
		(1.4e-3)	(4.4e-3)	(1.4e-3)	(3.9e-3)	(1.3e-3)	(3.4e-3)	(1.3e-3)	(3.5e-3)
	$V^{(2)}$	0.0653	0.0004	0.0957	0.0010	0.1077	0.0028	0.1167	0.0020
		(4.4e-3)	(1.1e-3)	(4.7e-3)	(1.3e-3)	(4.7e-3)	(1.1e-3)	(4.5e-3)	(1.2e-3)
		0.0004	0.0411	0.0010	0.0636	0.0028	0.0714	0.0020	0.0857
		(1.1e-3)	(2.8e-3)	(1.3e-3)	(3.0e-3)	(1.1e-3)	(2.8e-3)	(1.2e-3)	(3.0e-3)
(0.008, 0.008)	$V^{(1)}$	0.1146	-0.0080	0.1125	-0.0005	0.1216	0.0001	0.1242	-0.0026
		(4.3e-3)	(1.4e-3)	(3.6e-3)	(1.2e-3)	(3.5e-3)	(1.4e-3)	(3.1e-3)	(9.8e-4)
		-0.0080	0.1125	-0.0005	0.1085	0.0001	0.1249	-0.0026	0.1202
		(1.4e-3)	(4.0e-3)	(1.2e-3)	(5.5e-3)	(1.4e-3)	(5.0e-3)	(9.8e-4)	(2.8e-3)
	$V^{(2)}$	0.0982	0.0005	0.1002	-0.0021	0.1173	0.0003	0.1244	-0.0016
		(4.3e-3)	(1.3e-3)	(3.4e-3)	(9.3e-4)	(3.4e-3)	(9.4e-4)	(3.3e-3)	(8.6e-4)
		0.0005	0.0756	-0.0021	0.0709	0.0003	0.0794	-0.0016	0.0937
		(1.3e-3)	(3.3e-3)	(9.3e-4)	(2.7e-3)	(9.4e-4)	(2.1e-3)	(8.6e-4)	(2.3e-3)
(0.016, 0.016)	$V^{(1)}$	0.1280	-0.0049	0.1257	-0.0035	0.1297	-0.0033	0.1371	-0.0036
		(3.7e-3)	(1.1e-3)	(3.0e-3)	(1.0e-3)	(2.7e-3)	(8.6e-4)	(2.5e-3)	(8.3e-4)
		-0.0049	0.1221	-0.0035	0.1176	-0.0033	0.1257	-0.0036	0.1332
		(1.1e-3)	(3.2e-3)	(1.0e-3)	(3.5e-3)	(8.6e-4)	(2.4e-3)	(8.3e-4)	(2.5e-3)
	$V^{(2)}$	0.1134	0.0003	0.1219	0.0007	0.1324	-0.0008	0.1309	0.0001
		(3.7e-3)	(1.1e-3)	(3.6e-3)	(8.9e-4)	(3.2e-3)	(8.6e-4)	(2.8e-3)	(7.3e-4)
		0.0003	0.0918	0.0007	0.0868	-0.0008	0.0910	0.0001	0.0968
		(1.1e-3)	(2.9e-3)	(8.9e-4)	(2.1e-3)	(8.6e-4)	(2.0e-3)	(7.3e-4)	(1.9e-3)
(0.016, 0.032)	$V^{(1)}$	0.1266	-0.0071	0.1327	-0.0057	0.1402	-0.0058	0.1380	-0.0047
		(2.7e-3)	(1.1e-3)	(2.6e-3)	(8.2e-4)	(2.5e-3)	(7.8e-4)	(2.0e-3)	(7.2e-4)
		-0.0071	0.1380	-0.0057	0.1250	-0.0058	0.1253	-0.0047	0.1326
		(1.1e-3)	(2.9e-3)	(8.2e-4)	(2.4e-3)	(7.8e-4)	(2.1e-3)	(7.2e-4)	(2.2e-3)
	$V^{(2)}$	0.1282	-0.0004	0.1397	0.0001	0.1369	-0.0008	0.1371	-0.0013
		(3.3e-3)	(9.3e-4)	(3.2e-3)	(7.7e-4)	(2.4e-3)	(7.2e-4)	(2.1e-3)	(5.8e-4)
		-0.0004	0.1032	0.0001	0.0970	-0.0008	0.0936	-0.0013	0.0980
		(9.3e-4)	(2.5e-3)	(7.7e-4)	(2.0e-3)	(7.2e-4)	(1.6e-3)	(5.8e-4)	(1.5e-3)
(0.032, 0.032)	$V^{(1)}$	0.1345	-0.0076	0.1387	-0.0080	0.1425	-0.0077	0.1449	-0.0076
		(2.6e-3)	(8.9e-4)	(2.3e-3)	(7.3e-4)	(2.1e-3)	(6.6e-4)	(1.9e-3)	(6.2e-4)
		-0.0076	0.1449	-0.0080	0.1311	-0.0077	0.1353	-0.0076	0.1353
		(8.9e-4)	(2.5e-3)	(7.3e-4)	(2.1e-3)	(6.6e-4)	(1.8e-3)	(6.2e-4)	(1.6e-3)
	$V^{(2)}$	0.1347	0.0004	0.1333	0.0008	0.1389	0.0000	0.1392	-0.0008
		(2.9e-3)	(8.3e-4)	(2.4e-3)	(6.3e-4)	(2.1e-3)	(6.1e-4)	(1.8e-3)	(5.3e-4)
		0.0004	0.1043	0.0008	0.0949	0.0000	0.0977	-0.0008	0.1020
		(8.3e-4)	(2.2e-3)	(6.3e-4)	(1.6e-3)	(6.1e-4)	(1.5e-3)	(5.3e-4)	(1.3e-3)

(0.050, 0.050)	$V^{(1)}$	0.1435	-0.0075	0.1408	-0.0097	0.1412	-0.0106	0.1456	-0.0104
		(2.2e-3)	(7.6e-4)	(1.8e-3)	(6.3e-4)	(1.7e-3)	(5.2e-4)	(1.5e-3)	(4.7e-4)
		-0.0075	0.1473	-0.0097	0.1413	-0.0106	0.1389	-0.0104	0.1373
		(7.6e-4)	(2.0e-3)	(6.3e-4)	(1.7e-3)	(5.2e-4)	(1.5e-3)	(4.7e-4)	(1.4e-3)
	$V^{(2)}$	0.1408	-0.0006	0.1383	0.0001	0.1388	0.0001	0.1450	-0.0007
		(2.6e-3)	(6.9e-4)	(1.9e-3)	(5.2e-4)	(1.8e-3)	(4.7e-4)	(1.7e-3)	(4.4e-4)
		-0.0006	0.1109	0.0001	0.0977	0.0001	0.0961	-0.0007	0.1002
		(6.9e-4)	(1.8e-3)	(5.2e-4)	(1.4e-3)	(4.7e-4)	(1.1e-3)	(4.4e-4)	(1.0e-3)
(0.075, 0.075)	$V^{(1)}$	0.1489	-0.0078	0.1468	-0.0086	0.1522	-0.0135	0.1503	-0.0133
		(1.7e-3)	(5.6e-4)	(1.4e-3)	(4.5e-4)	(1.3e-3)	(3.9e-4)	(1.2e-3)	(3.8e-4)
		-0.0078	0.1548	-0.0086	0.1538	-0.0135	0.1382	-0.0133	0.1425
		(5.6e-4)	(1.6e-3)	(4.5e-4)	(1.2e-3)	(3.9e-4)	(1.2e-3)	(3.8e-4)	(1.1e-3)
	$V^{(2)}$	0.1489	0.0015	0.1515	0.0004	0.1466	-0.0005	0.1437	-0.0001
		(2.0e-3)	(5.9e-4)	(1.6e-3)	(4.5e-4)	(1.4e-3)	(3.5e-4)	(1.2e-3)	(3.2e-4)
		0.0015	0.1149	0.0004	0.1180	-0.0005	0.0932	-0.0001	0.0986
		(5.9e-4)	(1.3e-3)	(4.5e-4)	(1.1e-3)	(3.5e-4)	(8.7e-4)	(3.2e-4)	(8.5e-4)
(0.100, 0.100)	$V^{(1)}$	0.1494	-0.0087	0.1584	-0.0196	0.1639	-0.0194	0.1584	-0.0191
		(1.1e-3)	(3.8e-4)	(1.4e-3)	(3.7e-4)	(1.0e-3)	(3.2e-4)	(9.1e-4)	(2.9e-4)
		-0.0087	0.1568	-0.0196	0.1441	-0.0194	0.1397	-0.0191	0.1431
		(3.8e-4)	(1.0e-3)	(3.7e-4)	(1.0e-3)	(3.2e-4)	(8.7e-4)	(2.9e-4)	(7.8e-4)
	$V^{(2)}$	0.1534	0.0009	0.1490	0.0004	0.1479	0.0005	0.1545	0.0001
		(1.3e-3)	(3.9e-4)	(1.3e-3)	(2.9e-4)	(1.1e-3)	(2.4e-4)	(1.0e-3)	(2.2e-4)
		0.0009	0.1239	0.0004	0.0948	0.0005	0.0952	0.0001	0.0937
		(3.9e-4)	(9.0e-4)	(2.9e-4)	(8.7e-4)	(2.4e-4)	(7.7e-4)	(2.2e-4)	(7.0e-4)
(0.200, 0.200)	$V^{(1)}$	0.1478	-0.0089	0.1614	-0.0224	0.1634	-0.0230	0.1608	-0.0226
		(7.6e-4)	(2.6e-4)	(9.0e-4)	(3.1e-4)	(8.8e-4)	(2.9e-4)	(7.7e-4)	(2.9e-4)
		-0.0089	0.1565	-0.0224	0.1451	-0.0230	0.1446	-0.0226	0.1450
		(2.6e-4)	(7.2e-4)	(3.1e-4)	(7.9e-4)	(2.9e-4)	(6.4e-4)	(2.9e-4)	(6.2e-4)
	$V^{(2)}$	0.1508	0.0013	0.1513	0.0002	0.1526	0.0008	0.1527	0.0005
		(8.5e-4)	(2.6e-4)	(8.9e-4)	(1.9e-4)	(7.7e-4)	(1.8e-4)	(7.5e-4)	(1.5e-4)
		0.0013	0.1288	0.0002	0.0914	0.0008	0.0931	0.0005	0.0935
		(2.6e-4)	(7.3e-4)	(1.9e-4)	(8.1e-4)	(1.8e-4)	(7.5e-4)	(1.5e-4)	(7.2e-4)

3.4.3 Comparison with Classical Markov Switching Model

Base on the analysis in the simulation studies so far, our model has successfully detected the switching points in the simulated series and provides good estimation about regression parameters. The accomplishment would not be

Table 3.6: Theoretical average number of change points for different (p, q) scenarios and practical average number of change points based on 500 simulations in each $(p,\ q)$ scenario for $T = 2000, 3000, 4000$ and 5000. Standard errors are shown in parenthesis.

(p, q)	1000	2000	3000	4000	5000
(0.002, 0.002)	2	4	6	8	10
(0.004, 0.004)	4	8	12	16	20
(0.008, 0.008)	8	16	24	32	40
(0.016, 0.016)	16	32	48	64	80
(0.016, 0.032)	22	43	64	84	107
(0.032, 0.032)	32	64	96	128	160
(0.050, 0.050)	50	100	150	200	250
(0.075, 0.075)	75	150	225	300	375
(0.100, 0.100)	100	200	300	400	500
(0.200, 0.200)	200	400	600	800	1000

(a) Theoretical mean change points

(p, q)	2000	3000	4000	5000
(0.002, 0.002)	1.9(0.04)	2.9(0.04)	4.2(0.05)	5.1(0.06)
(0.004, 0.004)	3.2(0.04)	5.0(0.05)	7.0(0.05)	8.8(0.06)
(0.008, 0.008)	6.1(0.05)	9.4(0.06)	12.8(0.07)	16.0(0.07)
(0.016, 0.016)	9.0(0.04)	13.7(0.05)	18.4(0.06)	23.2(0.07)
(0.016, 0.032)	13.3(0.06)	20.1(0.07)	26.9(0.08)	33.8(0.09)
(0.032, 0.032)	17.8(0.06)	26.8(0.07)	35.8(0.08)	45.0(0.08)
(0.050, 0.050)	24.9(0.06)	37.5(0.07)	50.1(0.08)	62.8(0.09)
(0.075, 0.075)	37.7(0.07)	56.9(0.09)	75.9(0.10)	95.0(0.11)
(0.100, 0.100)	68.6(0.12)	103.2(0.15)	137.4(0.17)	172.0(0.19)
(0.200, 0.200)	104.8(0.11)	157.5(0.13)	210.0(0.14)	262.7(0.17)

(b) Practical average number of change points

more convincing without a comparison with the classical Markov switching (MS) regression models, hence the topic of this subsection.

To compare with model (3.7), the same regression construction is chosen for the classical MS model defined as

$$y_t = \alpha_{S_t} + \beta_{S_t} y_{t-1} + \sigma_{S_t}\epsilon, \tag{3.9}$$

where $\epsilon \sim N(0,1)$ and S_t follows a two-state first order Markov chain with $p = P(S_t = 2|S_{t-1} = 1)$ and $q = P(S_t = 1|S_{t-1} = 2)$ for all t. Regression parameters rely on regime status and no prior is assumed for these parameters, so Bayesian method would not be applicable. Final estimation includes two pieces of α, β, and σ for 2 regimes, the Markov transition probabilities and the probabilities of the regime at each time point. The number of estimates are much smaller than that of our model.

Two statistical packages (MatLab and R) written by Perlin (n.d.-a, n.d.-b) are available to implement the ideas in model (3.9). Interested readers may download the packages from the url provided in the reference section. In both packages maximum likelihood estimation (MLE) method is used to estimate the parameters; EM estimation is not available in these packages.

To make the comparison convincing, model (3.9) uses the same data as in Section 3.4 that were generated from model (3.7) resulting in 500 simulated series in each of 10 transition probability scenarios and each of four series lengths ($T = 2000, 3000, 4000$ and 5000). These data are loaded into R or MatLab in the proper format and the estimates recorded include $\hat{\alpha}_k, \hat{\beta}_k, \hat{\sigma}_k, \hat{p}, \hat{q}$ and $\widehat{P}(S_t = k)$ for $k = 1, 2$ and $t = 1, \ldots, T$. We compute the diagnostic statistics such as KL divergence, SSSE, L_2 norm and identification ratio based on the recorded statistics to compare with the statistics in Table 3.3.

Classical MS models assume that the level of parameters is a constant within a certain regime. In a two-regime system, model (3.9) can estimate only two pieces of α's, β's and so on. However, data were generated by a process where the levels of regression parameters are stochastic within each regime, so a two-state MS model is not ideal to capture the variations within the regime. Naturally KL statistics are not anticipated to perform well in this case. Table 3.7a shows that KL statistics are much larger and more volatile than those in Table 3.3a in all scenarios out of all series lengths. The mean KL statistics can be as high as 100 times more than those estimated by our model. Likewise, since L_2 norm measures the difference between the true and the estimated regression coefficients, it would not be difficult to understand why L_2 statistics in Table 3.7c have higher means and standard errors than the corresponding cells in Table 3.3c. SSSE statistics seem to be a controversial metric for the comparison of two models. SSSE can be decomposed into SSSEs by regimes for both models and further decomposed into pieces within a regime if using our model. For a particular regime, SSSE statistics by model (3.9) seem to be the averaging effect of the pieces of SSSEs by our model (3.7) in the same regime. To be more specific, $\hat{\beta}_1$ in model (3.9) possibly represents the average of the $\hat{\beta}_t$'s for those $\hat{S}_t = 1$ under model (3.7). The high values in Table 3.7b are not necessarily the sign of good fit.

Table 3.7: Monte Carlo means of diagnostic statistics for the classical Markov switching model. Standard errors are shown in parenthesis.

(p, q)	2000	3000	4000	5000
(0.002, 0.002)	0.3317(1.6e-02)	0.3618(1.5e-02)	0.4271(1.5e-02)	0.4313(1.4e-02)
(0.004, 0.004)	0.3593(1.3e-02)	0.4271(1.7e-02)	0.4690(1.3e-02)	0.5095(1.7e-02)
(0.008, 0.008)	0.4110(1.5e-02)	0.5213(1.8e-02)	0.5459(1.6e-02)	0.5735(1.7e-02)
(0.016, 0.016)	0.4785(1.4e-02)	0.5618(1.6e-02)	0.5754(1.4e-02)	0.6260(1.5e-02)
(0.016, 0.032)	0.5342(1.6e-02)	0.5934(1.6e-02)	0.6204(1.4e-02)	0.6190(1.4e-02)
(0.032, 0.032)	0.5411(1.4e-02)	0.6011(1.3e-02)	0.6177(1.3e-02)	0.6498(1.2e-02)
(0.050, 0.050)	0.5616(1.2e-02)	0.6087(1.3e-02)	0.6158(1.1e-02)	0.6460(1.1e-02)
(0.075, 0.075)	0.5619(9.9e-03)	0.5887(9.4e-03)	0.6153(8.8e-03)	0.6228(9.2e-03)
(0.100, 0.100)	0.5397(7.7e-03)	0.5581(7.4e-03)	0.5540(6.9e-03)	0.5507(7.0e-03)
(0.200, 0.200)	0.4947(4.7e-03)	0.5000(4.5e-03)	0.5096(4.5e-03)	0.5039(4.5e-03)

(a) Kullback-Leibler Divergence (KLD)

(p, q)	2000	3000	4000	5000
(0.002, 0.002)	0.9856(2.1e-02)	0.9367(1.2e-02)	0.9117(1.3e-02)	0.9177(1.2e-02)
(0.004, 0.004)	0.9229(1.4e-02)	0.8869(1.3e-02)	0.8844(1.2e-02)	0.8515(1.3e-02)
(0.008, 0.008)	0.9025(1.1e-02)	0.8350(1.2e-02)	0.8438(1.4e-02)	0.8139(1.1e-02)
(0.016, 0.016)	0.8613(1.2e-02)	0.8336(1.2e-02)	0.8210(1.2e-02)	0.8086(1.2e-02)
(0.016, 0.032)	0.8695(1.2e-02)	0.8081(1.2e-02)	0.8028(1.2e-02)	0.7796(1.1e-02)
(0.032, 0.032)	0.8250(1.2e-02)	0.7952(1.1e-02)	0.8003(1.1e-02)	0.7760(1.1e-02)
(0.050, 0.050)	0.7986(1.2e-02)	0.7798(1.1e-02)	0.7751(1.0e-02)	0.7395(1.0e-02)
(0.075, 0.075)	0.7955(9.8e-03)	0.7949(1.0e-02)	0.7578(9.2e-03)	0.7651(1.0e-02)
(0.100, 0.100)	0.8267(8.3e-03)	0.8144(8.4e-03)	0.8168(8.2e-03)	0.8252(7.7e-03)
(0.200, 0.200)	0.8680(5.8e-03)	0.8696(5.9e-03)	0.8718(5.6e-03)	0.8655(6.0e-03)

(b) Sum of Squares of Standardized Error (SSSE)

Table 3.7: Monte Carlo means of diagnostic statistics for estimates from traditional Markov Switching Model. Standard errors are shown in parenthesis.

(p, q)	2000	3000	4000	5000
(0.002, 0.002)	0.4169(1.3e-02)	0.4719(1.4e-02)	0.5177(1.1e-02)	0.5221(1.2e-02)
(0.004, 0.004)	0.4750(1.3e-02)	0.5074(1.1e-02)	0.5634(1.1e-02)	0.5717(1.1e-02)
(0.008, 0.008)	0.5006(9.5e-03)	0.5702(9.9e-03)	0.5944(1.1e-02)	0.6352(1.8e-02)
(0.016, 0.016)	0.5573(9.7e-03)	0.6154(1.0e-02)	0.6194(9.2e-03)	0.6731(1.0e-02)
(0.016, 0.032)	0.5914(9.6e-03)	0.6272(9.5e-03)	0.6588(9.5e-03)	0.6694(9.7e-03)
(0.032, 0.032)	0.5983(8.5e-03)	0.6587(8.9e-03)	0.6849(9.6e-03)	0.7166(1.0e-02)
(0.050, 0.050)	0.6381(9.1e-03)	0.6815(8.7e-03)	0.6896(8.9e-03)	0.7191(8.7e-03)
(0.075, 0.075)	0.6503(8.1e-03)	0.6813(8.7e-03)	0.7096(8.2e-03)	0.7203(8.9e-03)
(0.100, 0.100)	0.6355(7.4e-03)	0.6575(7.3e-03)	0.6570(6.8e-03)	0.6599(7.3e-03)
(0.200, 0.200)	0.5932(5.0e-03)	0.5900(4.4e-03)	0.6014(4.8e-03)	0.5959(4.4e-03)

(c) L_2

(p, q)	2000	3000	4000	5000
(0.002, 0.002)	0.6282(1.7e-02)	0.6456(1.6e-02)	0.6331(1.6e-02)	0.6365(1.6e-02)
(0.004, 0.004)	0.6667(1.6e-02)	0.6681(1.6e-02)	0.6379(1.6e-02)	0.6185(1.6e-02)
(0.008, 0.008)	0.6156(1.7e-02)	0.6410(1.5e-02)	0.6551(1.5e-02)	0.6459(1.5e-02)
(0.016, 0.016)	0.6372(1.5e-02)	0.6236(1.5e-02)	0.6302(1.4e-02)	0.6111(1.4e-02)
(0.016, 0.032)	0.6380(1.5e-02)	0.6638(1.4e-02)	0.6403(1.3e-02)	0.6702(1.3e-02)
(0.032, 0.032)	0.6400(1.4e-02)	0.6000(1.3e-02)	0.6112(1.3e-02)	0.5933(1.3e-02)
(0.050, 0.050)	0.6228(1.3e-02)	0.6044(1.2e-02)	0.5942(1.2e-02)	0.6051(1.2e-02)
(0.075, 0.075)	0.6002(1.3e-02)	0.6020(1.2e-02)	0.6000(1.2e-02)	0.5980(1.1e-02)
(0.100, 0.100)	0.6114(1.2e-02)	0.6032(1.2e-02)	0.6104(1.2e-02)	0.6197(1.2e-02)
(0.200, 0.200)	0.6310(1.3e-02)	0.6351(1.3e-02)	0.6211(1.3e-02)	0.6517(1.2e-02)

(d) Identification Ratio (IR)

Table 3.8: Monte Carlo means of estimated $\hat{p}$, $\hat{q}$ by Classical Markov switching regression model. Standard errors are shown in parenthesis.

(p, q)	2000		3000		4000		5000	
	$\hat{p}$	$\hat{q}$	$\hat{p}$	$\hat{q}$	$\hat{p}$	$\hat{q}$	$\hat{p}$	$\hat{q}$
(0.002, 0.002)	0.0695 (6.6e-03)	0.0800 (6.2e-03)	0.0866 (7.6e-03)	0.0731 (4.7e-03)	0.0678 (5.8e-03)	0.0573 (3.8e-03)	0.0743 (6.7e-03)	0.0564 (4.2e-03)
(0.004, 0.004)	0.0735 (6.4e-03)	0.0707 (5.4e-03)	0.0691 (6.5e-03)	0.0510 (3.1e-03)	0.0847 (6.0e-03)	0.0590 (3.7e-03)	0.0810 (6.5e-03)	0.0579 (4.2e-03)
(0.008, 0.008)	0.0725 (5.9e-03)	0.0558 (4.0e-03)	0.0897 (6.2e-03)	0.0600 (3.6e-03)	0.0854 (5.4e-03)	0.0596 (3.5e-03)	0.0836 (5.1e-03)	0.0599 (3.5e-03)
(0.016, 0.016)	0.0859 (5.6e-03)	0.0548 (3.3e-03)	0.0968 (5.9e-03)	0.0614 (2.8e-03)	0.0982 (5.6e-03)	0.0656 (3.4e-03)	0.0999 (5.3e-03)	0.0658 (2.8e-03)
(0.016, 0.032)	0.0965 (5.7e-03)	0.0687 (4.1e-03)	0.1092 (6.4e-03)	0.0714 (3.3e-03)	0.1091 (5.3e-03)	0.0749 (3.2e-03)	0.1239 (7.0e-03)	0.0738 (3.2e-03)
(0.032, 0.032)	0.0964 (4.9e-03)	0.0691 (3.3e-03)	0.1114 (6.1e-03)	0.0820 (3.5e-03)	0.1110 (5.4e-03)	0.0787 (3.5e-03)	0.1235 (6.1e-03)	0.0869 (3.9e-03)
(0.050, 0.050)	0.1227 (6.8e-03)	0.0747 (3.2e-03)	0.1443 (7.5e-03)	0.0830 (3.6e-03)	0.1330 (6.2e-03)	0.0846 (3.7e-03)	0.1362 (6.8e-03)	0.0931 (3.6e-03)
(0.075, 0.075)	0.1373 (6.9e-03)	0.0861 (3.7e-03)	0.1476 (7.6e-03)	0.0949 (3.7e-03)	0.1539 (7.1e-03)	0.1023 (3.8e-03)	0.1593 (8.1e-03)	0.0981 (3.5e-03)
(0.100, 0.100)	0.1506 (8.3e-03)	0.1136 (4.3e-03)	0.1507 (8.5e-03)	0.1149 (3.8e-03)	0.1523 (8.6e-03)	0.1137 (3.8e-03)	0.1667 (9.7e-03)	0.1098 (3.6e-03)
(0.200, 0.200)	0.1289 (7.0e-03)	0.1186 (3.2e-03)	0.1262 (7.3e-03)	0.1148 (3.0e-03)	0.1330 (7.2e-03)	0.1159 (2.6e-03)	0.1418 (8.5e-03)	0.1155 (2.8e-03)

It is not entirely fair to evaluate the fitness of a model by applying it to a process generated by a different model. But one of the major concerns in regime switching modelling is the statistical inference of the regimes. The percentages of the regimes correctly identified are systematically 30% higher in Table 3.3d than those in Table 3.7d, which shows that our model is superior in this regard. I cannot leave this section without mentioning the estimation of transition probabilities. We have argued in Section 3.4.2 why the estimated transition probabilities should be lower and how much they should be. Estimation of p and q in Table 3.8 are higher than those in Table

3.4d. This simulation seems to prove the current suspicion that the classical Markov switching models have a tendency to overestimate the transition probabilities.

3.5 Analysis of EM Algorithm

Although Expectation-Maximization (EM) algorithm has been proven mature in theory, there are many challenges in applications. EM algorithm is model-specific, and for any changes in the model specification, practitioners have to derive a set of new EM estimators. Due to this reason, current statistical software is not well equipped with EM algorithm analysis. At the maximization step, EM algorithm may still rely on other optimization algorithms if the closed form solutions do not exit. In this case, EM algorithm loses its advantage to the classical maximum likelihood estimation method which optimizes the likelihood function directly. EM algorithm has also been proven to converge slowly and need more computation time to reach more "accurate" results.

The application of EM algorithm in our model is an innovative one. For one thing, the solutions to EM algorithm have an explicit form and for the other EM algorithm is used to estimate the nuisance parameters and not directly to the model parameters of primary interest. In this simulation, we will discuss the effect of initial prior values on the performance of EM algorithm, explore the properties and the speed of convergence and give some practical advice on the choice of initial prior values.

We continue to use model (3.7) restated here.

$$y_t = \beta_{0,t} + \beta_{1,t} y_{t-1} + \sigma_t \epsilon_t$$

where $\beta_{i,t}$ and σ_t depend on a two-state Markov chain. To save computation time, we choose series length $T = 1000$. 500 series are generated with prior values $\boldsymbol{g}' = (2.5,\ 1.2)$, $\boldsymbol{\lambda}' = (0.8,\ 1)$, $\boldsymbol{z} = \begin{pmatrix} 0.3 & 0.5 \\ -0.2 & -0.5 \end{pmatrix}$, $\boldsymbol{V}^{(1)} = \boldsymbol{V}^{(2)} = \begin{pmatrix} 0.16 & 0 \\ 0 & 0.16 \end{pmatrix}$ and transition matrix $P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$, where p and q are specified in the following scenarios.

Scenario 1 $p = 0.004,\ q = 0.004$

Scenario 2 $p = 0.01,\ q = 0.01$

Scenario 3 $p = 0.04,\ q = 0.08$

Scenario 4 $p = 0.05,\ q = 0.05$

Scenario 5 $p = 0.1,\ q = 0.1$

Like in Section 3.4, the positions and the number of change points are regulated by the Markov chain transition matrix defined above and therefore stochastic. Again we restrict the time interval between two adjacent change points to be no less than 20 to facilitate visual presentation of model estimates.

Although mathematically, convergence of a function ought to be well explored in the entire domain, there is no easy solution to an exhaustive search as far as multidimensional parameter space is concerned. There are 16 prior parameters to be estimated in this model. However for a statistical question, we can evaluate the practical meaning of the parameters which may restrict their practical range. For example, we are interested in a time series where there are a few switches in a certain time interval. So p or q, the probability of making a transition from state 1(2) to state 2(1) would be reasonably less than 0.5. In addition, it makes little sense to assume extreme variance for prior distribution. Variance and covariance of $\boldsymbol{\beta}_t$ is ratio of prior $\boldsymbol{V}$ and another random variable τ_t whose mean is the product of g and λ. For a fixed $\boldsymbol{V}$, large values of g and λ, may yield a small variance and covariance, and vice versa. For fixed g and λ, larger $\boldsymbol{V}$ yields larger variance and covariance and vice versa. In practice, extreme values are usually avoided and we will take a good balance of $\boldsymbol{V}$, g and λ into consideration. The characteristic of stationary AR(1) model also has restrictions on $\boldsymbol{\beta}$ whose mean is the prior $\boldsymbol{z}$. Thus, I choose the following 6 different sets of priors.

Initial prior 1

$$\boldsymbol{g} = \begin{pmatrix} 2.8 \\ 1.5 \end{pmatrix}, \boldsymbol{\lambda} = \begin{pmatrix} 1 \\ 1.2 \end{pmatrix}, P = \begin{pmatrix} 0.95 & 0.05 \\ 0.05 & 0.95 \end{pmatrix}, \boldsymbol{V}^{(1)} = \boldsymbol{V}^{(2)} = \begin{pmatrix} 0.15 & 0 \\ 0 & 0.15 \end{pmatrix}$$

$$\text{and } \boldsymbol{z} = \begin{pmatrix} 0.2 & 0.4 \\ -0.2 & -0.4 \end{pmatrix}$$

Initial prior 2

$$\boldsymbol{g} = \begin{pmatrix} 0.5 \\ 0.3 \end{pmatrix}, \boldsymbol{\lambda} = \begin{pmatrix} 1 \\ 1.2 \end{pmatrix}, P = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{pmatrix}, \boldsymbol{V}^{(1)} = \boldsymbol{V}^{(2)} = \begin{pmatrix} 0.9 & 0 \\ 0 & 0.9 \end{pmatrix}$$

and $\boldsymbol{z} = \begin{pmatrix} 0.2 & 0.4 \\ -0.2 & -0.4 \end{pmatrix}$

Initial prior 3

$\boldsymbol{g} = \begin{pmatrix} 4 \\ 5 \end{pmatrix}, \boldsymbol{\lambda} = \begin{pmatrix} 3 \\ 4 \end{pmatrix}, P = \begin{pmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{pmatrix}, \boldsymbol{V}^{(1)} = \boldsymbol{V}^{(2)} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$ and

$\boldsymbol{z} = \begin{pmatrix} -1 & -1 \\ 0.5 & 0.5 \end{pmatrix}$

Initial prior 4

$\boldsymbol{g} = \begin{pmatrix} 5 \\ 4 \end{pmatrix}, \boldsymbol{\lambda} = \begin{pmatrix} 3 \\ 4 \end{pmatrix}, P = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{pmatrix}, \boldsymbol{V}^{(1)} = \boldsymbol{V}^{(2)} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$ and

$\boldsymbol{z} = \begin{pmatrix} 1 & -0.9 \\ -1 & 0.8 \end{pmatrix}$

Initial prior 5

$$\boldsymbol{g} = \begin{pmatrix} 0.1 \\ 0.2 \end{pmatrix}, \boldsymbol{\lambda} = \begin{pmatrix} 20 \\ 10 \end{pmatrix}, P = \begin{pmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{pmatrix}, \boldsymbol{V}^{(1)} = \boldsymbol{V}^{(2)} = \begin{pmatrix} 10 & 0 \\ 0 & 10 \end{pmatrix}$$

and $\boldsymbol{z} = \begin{pmatrix} 1 & 2 \\ -1 & -2 \end{pmatrix}$

Initial prior 6

$\boldsymbol{g} = \begin{pmatrix} 3 \\ 4 \end{pmatrix}, \boldsymbol{\lambda} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, P = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}, \boldsymbol{V}^{(1)} = \boldsymbol{V}^{(2)} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and

$$\boldsymbol{z} = \begin{pmatrix} -1 & -2 \\ 2 & 1 \end{pmatrix}$$

Prior 1 is chosen very close to the true prior parameters. Prior 2 slightly decrease $\boldsymbol{g}$, and increase (p, q) and $\boldsymbol{V}$. The true z_1 and z_2 are positive for state 1 and negative for state 2. In prior 1 and prior 2, z_1 and z_2 have the same sign as their true prior counterpart. Prior 3 chooses a little larger $\boldsymbol{g}$ and $\boldsymbol{\lambda}$ and reverses signs for (z_1, z_2) for different states. Prior 4 is similar to prior 3 except that z_2 is negative for state 1 and positive for state 2. Prior 5 is special in very large $\boldsymbol{V}$. Prior 6 is similar to prior 3 and the values are closer to the true prior values.

3.5.1 Analysis of Hyperparameter Estimation

When there are only a couple of change points as in scenario 1, the estimated $g^{(k)}$ may explode as the number of iteration increases depending on the choice of initial priors shown in Figure 3.2a and 3.2b; likewise the estimated $\lambda^{(k)}$ may have a chance to converge to zero shown in Figure 3.2c and 3.2d. In larger (p, q) scenarios, $g^{(k)}$ and $\lambda^{(k)}$ always converge to a constant regardless of the choice of initial prior values. The reason of aberrant estimation of $g^{(k)}$ and $\lambda^{(k)}$ in the low transition probability series may be the fact that there is little information in the data to estimate parameters in a complex multidimensional space. The less the number of change points in a series, the less information EM algorithm can uses to estimate in a multidimensional space, thus the less accurate the estimates. However, in such cases, the multiplication of g and λ converges to to a constant that is reasonably closer to its true prior counterpart. This raises a question of the lack of parameter identifiability, i.e. there is a function of the parameters that data do not yield almost any information about. One possible solution is to reduce the number of parameters. In current model, the prior distribution of σ_t is determined by Gamma(g, λ) and hidden states. If hidden states cannot be avoided, the choice is to reduce the number of parameters associated with the distribution of σ_t. T and Chi-square distributions are determined by one parameter, which may be the candidate choice in this case. However, these prior distributions may defeat the desirable property of conjugation that are provided by Gamma-Normal pair and used in the current model estimation. Thus caution must be taken if one wants to pursue other prior distributions.

Despite the fact that EM algorithm does not estimate g and λ well in low frequent switching series, $\hat{g}$ does converge when the series switches regimes more frequently. If it ever converges, it converges fast and usually within 15 iterations. $\hat{g}_1$ converges to a constant, not necessarily the exact true prior, but reasonably closer to the truth. Figure 3.3 – 3.6 (a) show that solid black (prior 1), dashed grey (prior 2) and long dashed blue (prior 5) lines are much closer to solid red line (true prior values) and these lines become stable within 10 iterations. Figure 3.3 – 3.6 (b) show that dotted magenta (prior 3), dash-dot green (prior 4) and short-long dashed orange (prior 6) lines seem to be closer the the true prior values, but these lines take more iterations to stabilize than the black, grey and blue lines do. In general prior 1, 2, 5 gives better $\hat{g}$ estimates in scenarios 2–5. Similarly EM estimation of $\widehat{\boldsymbol{\lambda}}$ works fine in more frequent switching series, i.e. it converges fast and converges to a value reasonably closer to the true prior value, shown in (c) and (d) of Figure 3.3 – 3.6. All prior estimates become stable eventually, but solid black, dashed grey and long dashed blue lines converge faster than other lines. So initial prior 1, 2 and 5 tend to give better estimates of λ for all scenarios except the first one.

Transition probability in general converges fast for all scenarios. When transition probabilities are very small as in scenario 1, the estimation is very much close to the truth, although prior 3 (dotted magenta) gives a much higher estimates shown in Figure 3.2e. When the true transition probabilities are larger as in scenario 3, 4 and 5, the estimates tend to be lower than the true prior values. This result confirms the previous findings in Section 3.4.2 due to the fact the practical average of number of change points generated in the simulation is always lower than the theoretical counterparts. Thus it is not sensible to conclude that the model tends to underestimate transition probabilities. Although different initial priors may result in different converging values, estimates from initial prior 1, 2, 5 tend to converge to the same value, while initial prior 3, 4 and 6 yield a slightly different converging constant. Prior 1, 2, 5 and 6 are better initial values in a way that these estimates tend to converge faster, within 5 iterations, whereas by initial prior 3 and 4, $\hat{p}$ or $\hat{q}$ takes longer steps to converge, as shown in the

Figure 3.2 – 3.6 (e) and (f).

In the estimation of variance and covariance matrix, the trouble also arises in the low frequent switching scenario. In these cases, $\widehat{\boldsymbol{V}}^{(k)}$ has a tendency to converge to a zero matrix. When a nearly singular matrix occurs at a given iteration, the subsequent estimation of other parameters requires the inverse or the determinant of $\widehat{\boldsymbol{V}}^{(k)}$, whose values may be not a number (NaN) or infinity produced by most of computer software. Thus the entire estimation terminates. In scenario 1, the estimation of covariance matrix is sensitive to initial prior choices. If a covariance matrix converges to zero, it can happen as fast as within 10 iterations shown in Figure 3.2 (g) – (l). Covariance matrices converges in all other scenarios. The estimates via prior 1, 2 and 5 share similarities; whereas those by prior 3, 4 and 6 seem to be close to each other. Again estimation of $\boldsymbol{V}$ become stable faster by prior 1, 2, 5 and 6 than those by initial prior 3, 4, shown in Figure 3.2 – 3.6 (g) – (l).

With no surprise $\widehat{\boldsymbol{z}}$ in general converges and converges fast like other prior estimates do. Even in questionable scenario 1, $\widehat{\boldsymbol{z}}$ converges to reasonable values. What is special in this estimation is that the signs of the estimates are associated with the initial prior choices. Since true $\boldsymbol{z}$ is positive for state 1 and negative for state 2, ideally the estimates of the $\boldsymbol{z}$ should follow the same sign for different states. However if the signs of initial $\boldsymbol{z}$ are chosen the opposite of the true prior for all states, the estimates end up with the wrong signs. For example, in initial prior 6 where $\boldsymbol{z}^{(1)}$ are given negative values and $\boldsymbol{z}^{(2)}$ are given the positive values instead, $\widehat{\boldsymbol{z}}^{(1)}$ eventually stabilizes at around

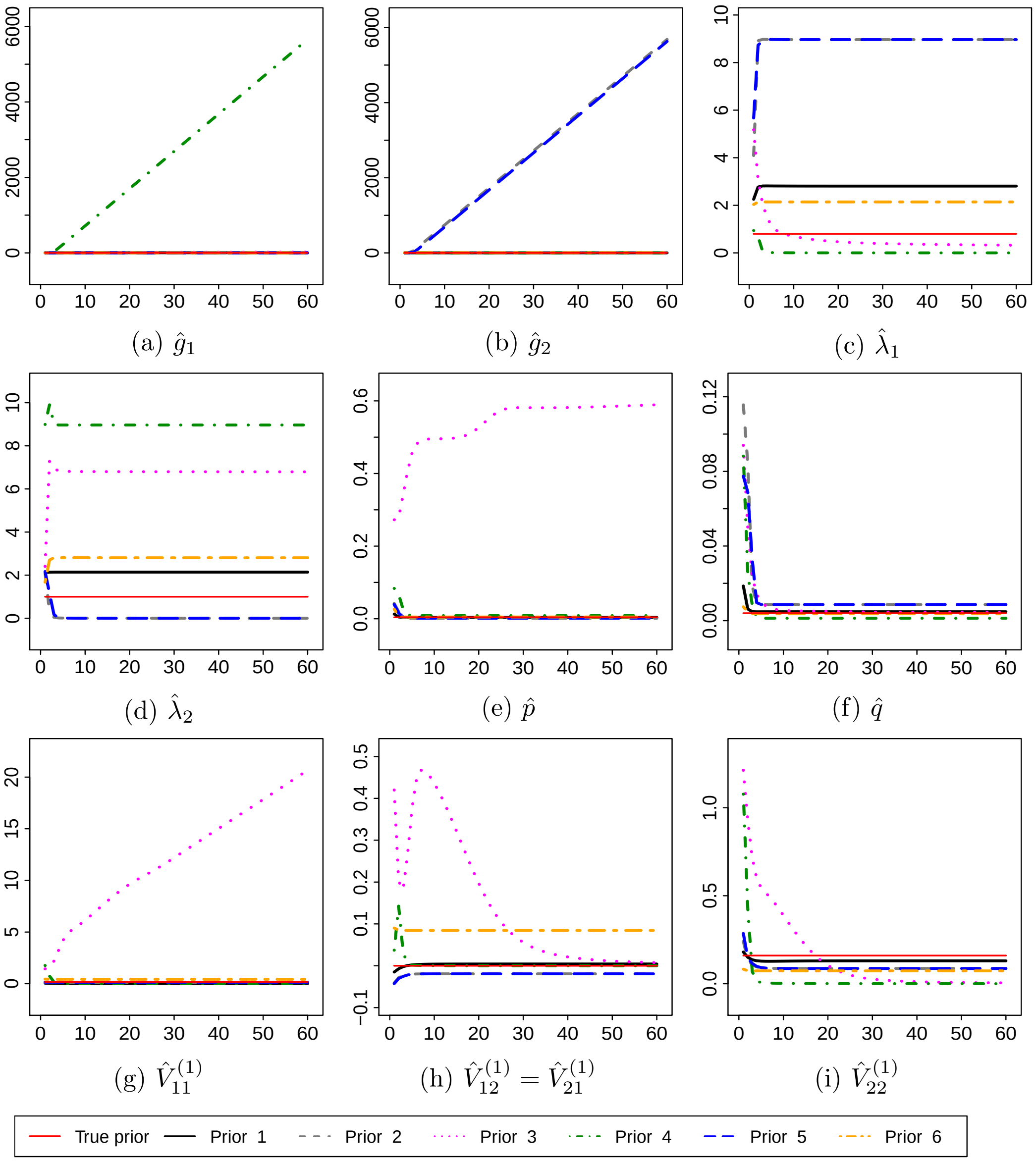
$(-0.34, -0.60)'$ and $\widehat{\boldsymbol{z}}^{(2)}$ comes out about $(0.36, 0.80)'$ shown in Figure 3.3m – 3.3p. In another sign experiment as in prior 4, $z_1^{(k)}$ has the same sign as their true prior counterpart, $z_2^{(k)}$ is given the opposite sign, i.e. negative for state 1 and positive for state 2. Figure 3.3m – 3.3p show that after more number of iterations, $\widehat{\boldsymbol{z}}^{(1)}$ converges to $(-0.34, -0.60)'$ and $\widehat{\boldsymbol{z}}^{(2)}$ to $(0.36, 0.80)'$ as well. On the other hand, if the signs of initial priors are the same as those of their true prior counterpart, the estimation ends up with the right signs and better estimates in a way that these estimates are much closer to the true prior values. In Figure 3.3m – 3.3p , Solid black, dashed grey and long dashed blue lines converge within 5 iterations and those lines are closer to the true value (solid red line). In summary, estimates of $\widehat{\boldsymbol{z}}$ by initial prior 1, 2, 5 tend to converge fast and converge to the values that are closer to the truth, and on the contrary, the estimates of $\widehat{\boldsymbol{z}}$ by initial prior 3, 4, and 6 converge slower and converge to values with the opposite signs of the true prior values for most of the case. These findings are more prominent in more frequent switching series that are generated by scenario 3, 4, and 5, and shown in Figure 3.4m - 3.4p, Figure 3.5m - 3.5p and Figure 3.6m - 3.6p.

EM algorithm has the property to increase the log likelihood for every iteration. Figure 3.2 – 3.6 (q) shows that log likelihood function becomes stable as the number of iterations increases in general except for scenario 1. The discussion of prior estimation in EM algorithm has always been difficult for low frequent switching series, such as in scenario 1 with no exception of log likelihood function. Log likelihood function estimated by prior 3 in scenario

1 is diverging possibly because $\hat{\boldsymbol{g}}$ is diverging. Log likelihood may decrease for some iterations in scenario 2 is still a bit of puzzle shown in Figure 3.3q. Log likelihood by initial prior 1 (solid black), prior 2 (dashed grey), prior 5 (long dashed blue) and prior 7 (short-long dashed orange) converges the fastest within 10 iterations; log likelihood by initial prior 3 (dotted magenta) and 4 (dash-dotted green) converges the slowest within 20 iterations as shown in Figure 3.4q, 3.5q and 3.6q. This may prove a widely acknowledged fact that the EM algorithm finds a local maximum of a latent variable model likelihood. In addition, we observe that there are multiple solutions to the same maximized likelihood value. Which solution is the best, in another word, which initial prior is optimal cannot be determined by evaluating prior estimates alone. I will discuss this issue in the next subsection.

We conclude this subsection by giving some practical advices on how to implement EM algorithm. EM algorithm works better when the data are informative, i.e. more structural changes. Prior estimation is less sensitive to the initial $\boldsymbol{g}$, $\boldsymbol{\lambda}$, P and $\boldsymbol{V}$, but sensitive to the signs of initial $\boldsymbol{z}$. We have seen in many cases that initial prior 1, 2 and 5 produce similar output, initial prior 3, 4, and 6 converge a little slower and may produce another similar output. The major difference among two groups are the signs of $\boldsymbol{z}$. If the initial $\boldsymbol{z}$ has the same sign as the true parameter, the algorithm converges faster and the estimates are likely to be closer to the true values. Running EM algorithm is expensive, especially when there are so many parameters. Each parameter may have its own rate of convergence. My observation is that if

Figure 3.2: The estimated prior parameters $\widehat{\boldsymbol{g}}$, $\widehat{\boldsymbol{\lambda}}$, $\hat{p}$, $\hat{q}$ and $\widehat{\boldsymbol{V}}^{(1)}$ in the first 60 iterations of the EM algorithm from a selected series. The true $(p,\ q) = (0.004,\ 0.004)$



(a) $\hat{g}_1$

(b) $\hat{g}_2$

(c) $\hat{\lambda}_1$

(d) $\hat{\lambda}_2$

(e) $\hat{p}$

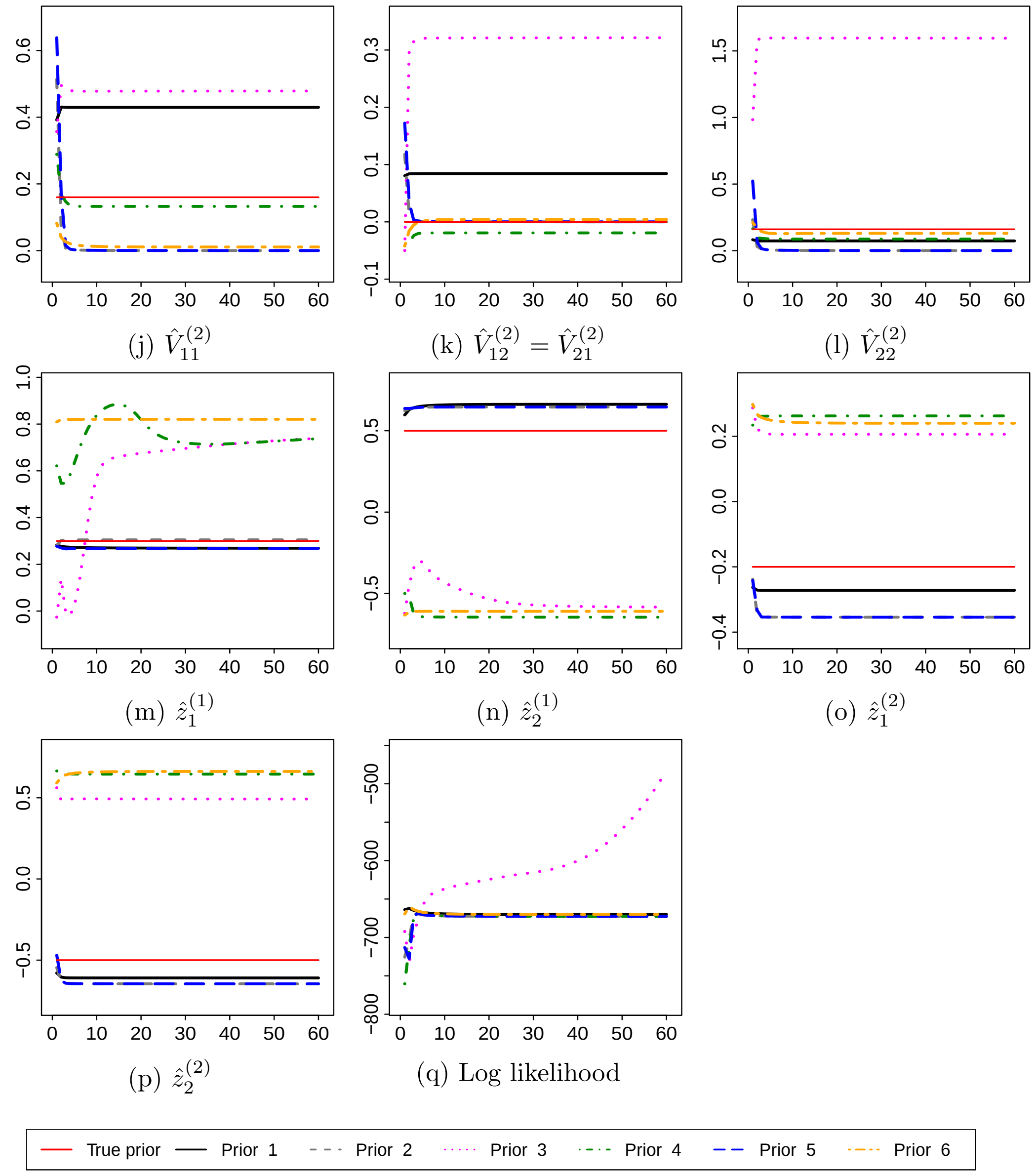
(f) $\hat{q}$

(g) $\hat{V}_{11}^{(1)}$

(h) $\hat{V}_{12}^{(1)} = \hat{V}_{21}^{(1)}$

(i) $\hat{V}_{22}^{(1)}$

Figure 3.2: Continued. The estimated prior parameters $\hat{\boldsymbol{V}}^{(2)}$ and $\hat{\boldsymbol{z}}$ and log likelihood in the first 60 iterations of EM algorithm from a selected series. The true $(p,\ q) = (0.004,\ 0.004)$



(j) $\hat{V}_{11}^{(2)}$

(k) $\hat{V}_{12}^{(2)} = \hat{V}_{21}^{(2)}$

(l) $\hat{V}_{22}^{(2)}$

(m) $\hat{z}_1^{(1)}$

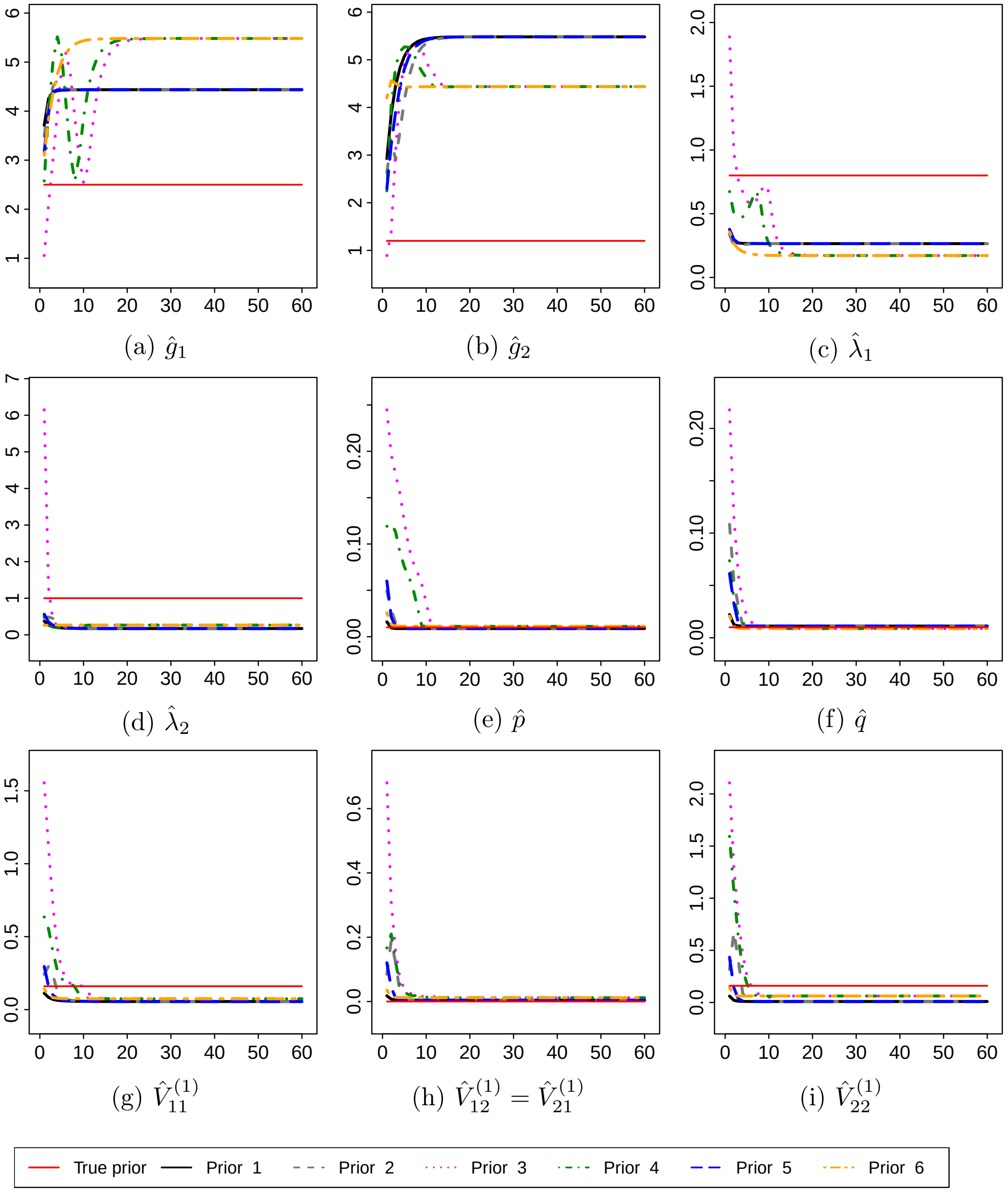
(n) $\hat{z}_2^{(1)}$

(o) $\hat{z}_1^{(2)}$

(p) $\hat{z}_2^{(2)}$

(q) Log likelihood

Figure 3.3: The estimated prior parameters $\widehat{\boldsymbol{g}}$, $\widehat{\boldsymbol{\lambda}}$, $\hat{p}$, $\hat{q}$ and $\widehat{\boldsymbol{V}}^{(1)}$ in the first 60 iterations of the EM algorithm from a selected series. The true $(p,\ q) = (0.01,\ 0.01)$



(a) $\hat{g}_1$

(b) $\hat{g}_2$

(c) $\hat{\lambda}_1$

(d) $\hat{\lambda}_2$

(e) $\hat{p}$

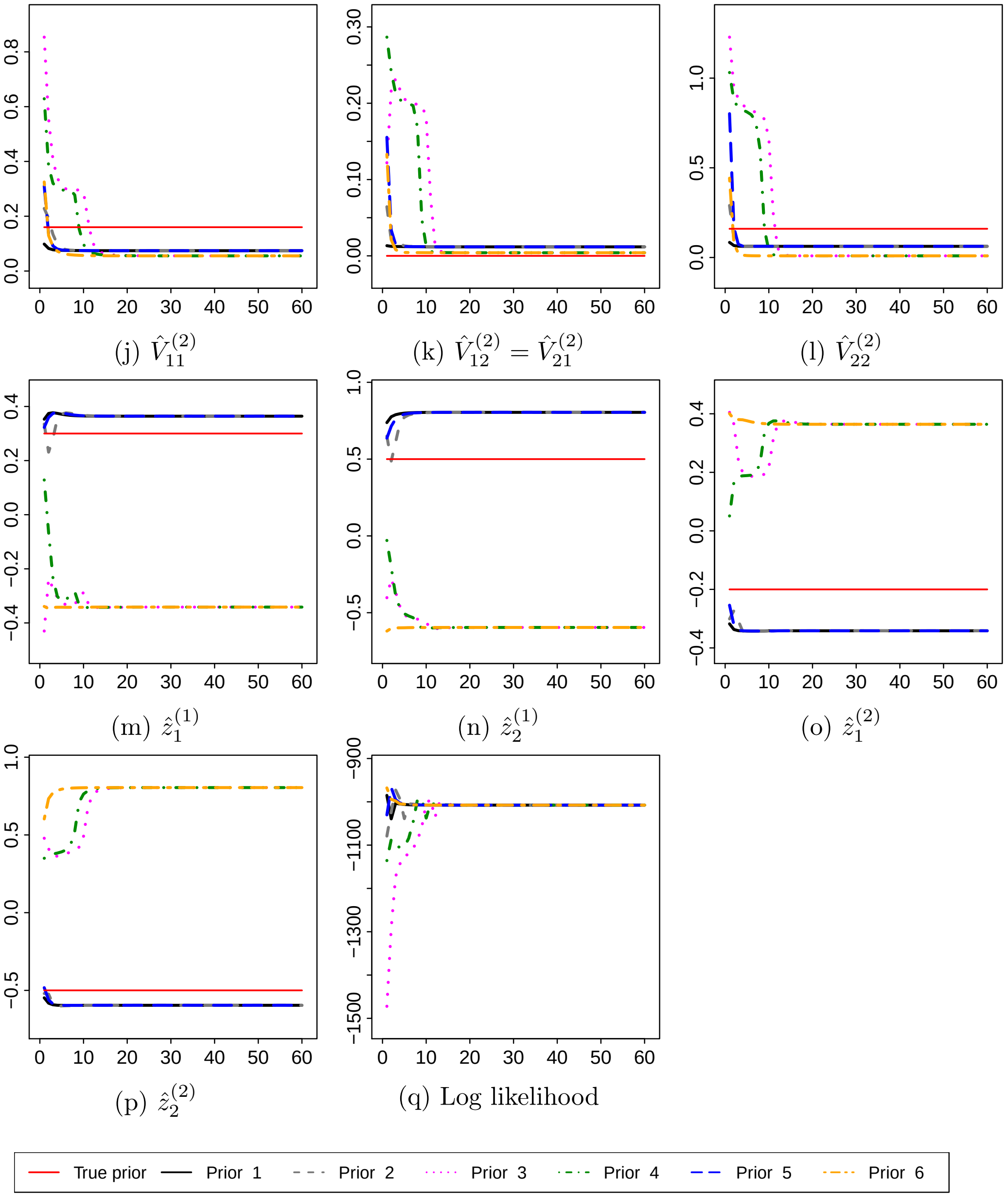
(f) $\hat{q}$

(g) $\hat{V}_{11}^{(1)}$

(h) $\hat{V}_{12}^{(1)} = \hat{V}_{21}^{(1)}$

(i) $\hat{V}_{22}^{(1)}$

Figure 3.3: Continued. The estimated prior parameters $\hat{\boldsymbol{V}}^{(2)}$, $\hat{\boldsymbol{z}}$ and log likelihood in the first 60 iterations of EM algorithm from a selected series. The true $(p,\ q) = (0.01,\ 0.01)$



(j) $\hat{V}_{11}^{(2)}$

(k) $\hat{V}_{12}^{(2)} = \hat{V}_{21}^{(2)}$

(l) $\hat{V}_{22}^{(2)}$

(m) $\hat{z}_1^{(1)}$

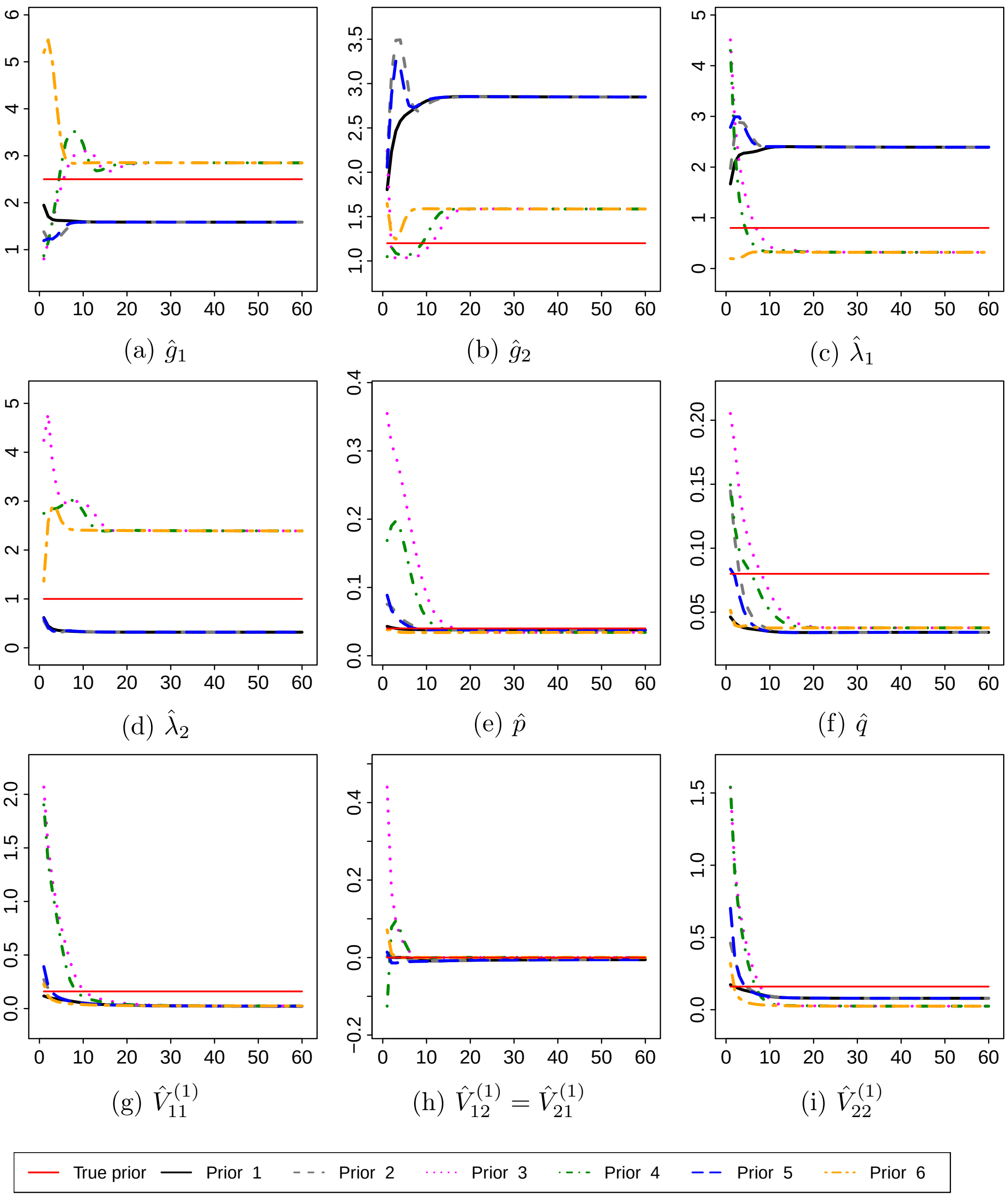
(n) $\hat{z}_2^{(1)}$

(o) $\hat{z}_1^{(2)}$

(p) $\hat{z}_2^{(2)}$

(q) Log likelihood

Figure 3.4: The estimated prior parameters $\widehat{\boldsymbol{g}}$, $\widehat{\boldsymbol{\lambda}}$, $\hat{p}$, $\hat{q}$ and $\widehat{\boldsymbol{V}}^{(1)}$ in the first 60 iterations of the EM algorithm from a selected series. The true $(p,\ q) = (0.04,\ 0.08)$

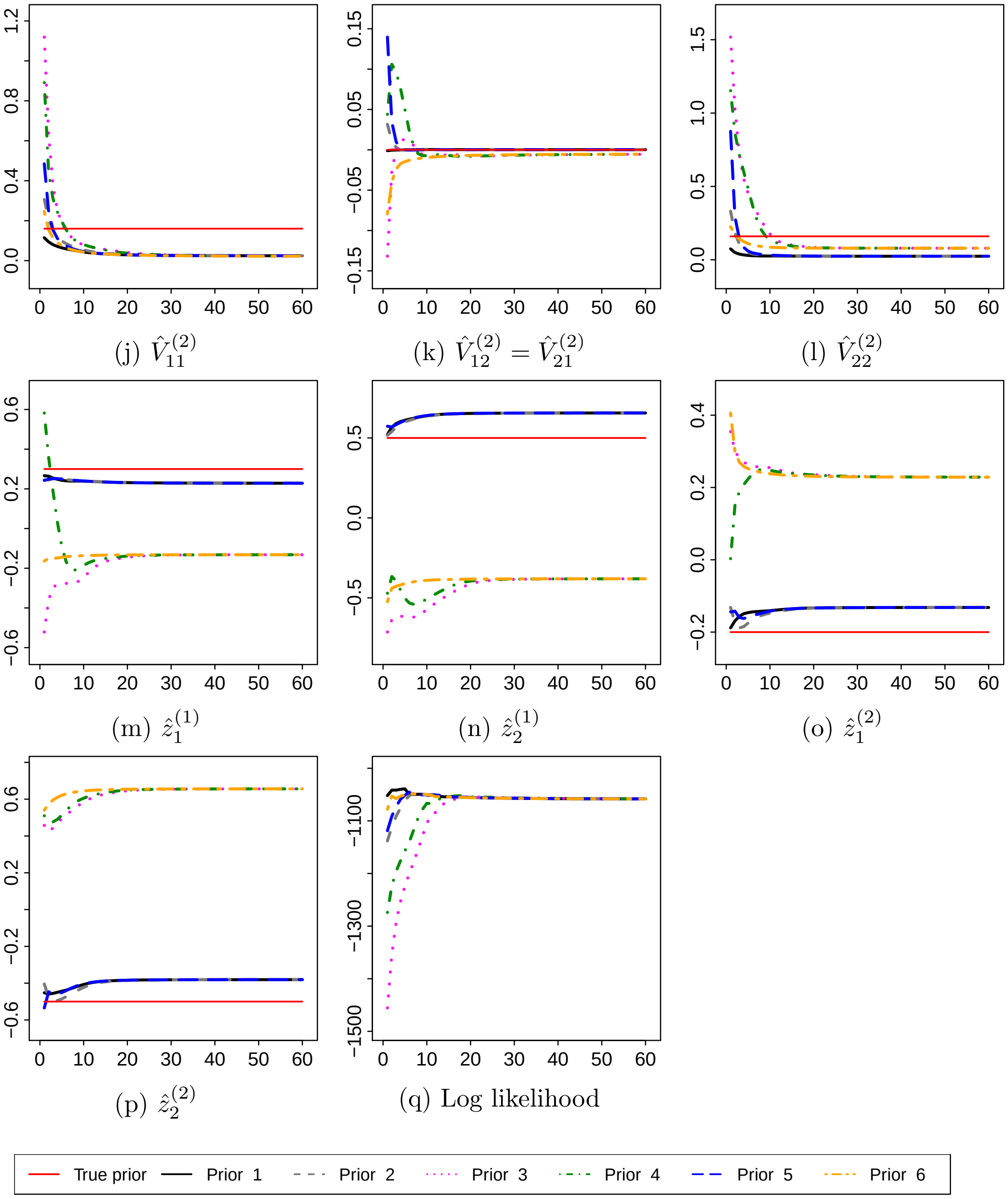


(a) $\hat{g}_1$ (b) $\hat{g}_2$ (c) $\hat{\lambda}_1$

(d) $\hat{\lambda}_2$ (e) $\hat{p}$ (f) $\hat{q}$

(g) $\hat{V}_{11}^{(1)}$ (h) $\hat{V}_{12}^{(1)} = \hat{V}_{21}^{(1)}$ (i) $\hat{V}_{22}^{(1)}$

Figure 3.4: Continued. The estimated prior parameters $\widehat{\boldsymbol{V}}^{(2)}$, $\hat{\boldsymbol{z}}$ and log likelihood in the first 60 iterations of EM algorithm from a selected series. The true $(p,\ q) = (0.04,\ 0.08)$



(j) $\hat{V}_{11}^{(2)}$

(k) $\hat{V}_{12}^{(2)} = \hat{V}_{21}^{(2)}$

(l) $\hat{V}_{22}^{(2)}$

(m) $\hat{z}_1^{(1)}$

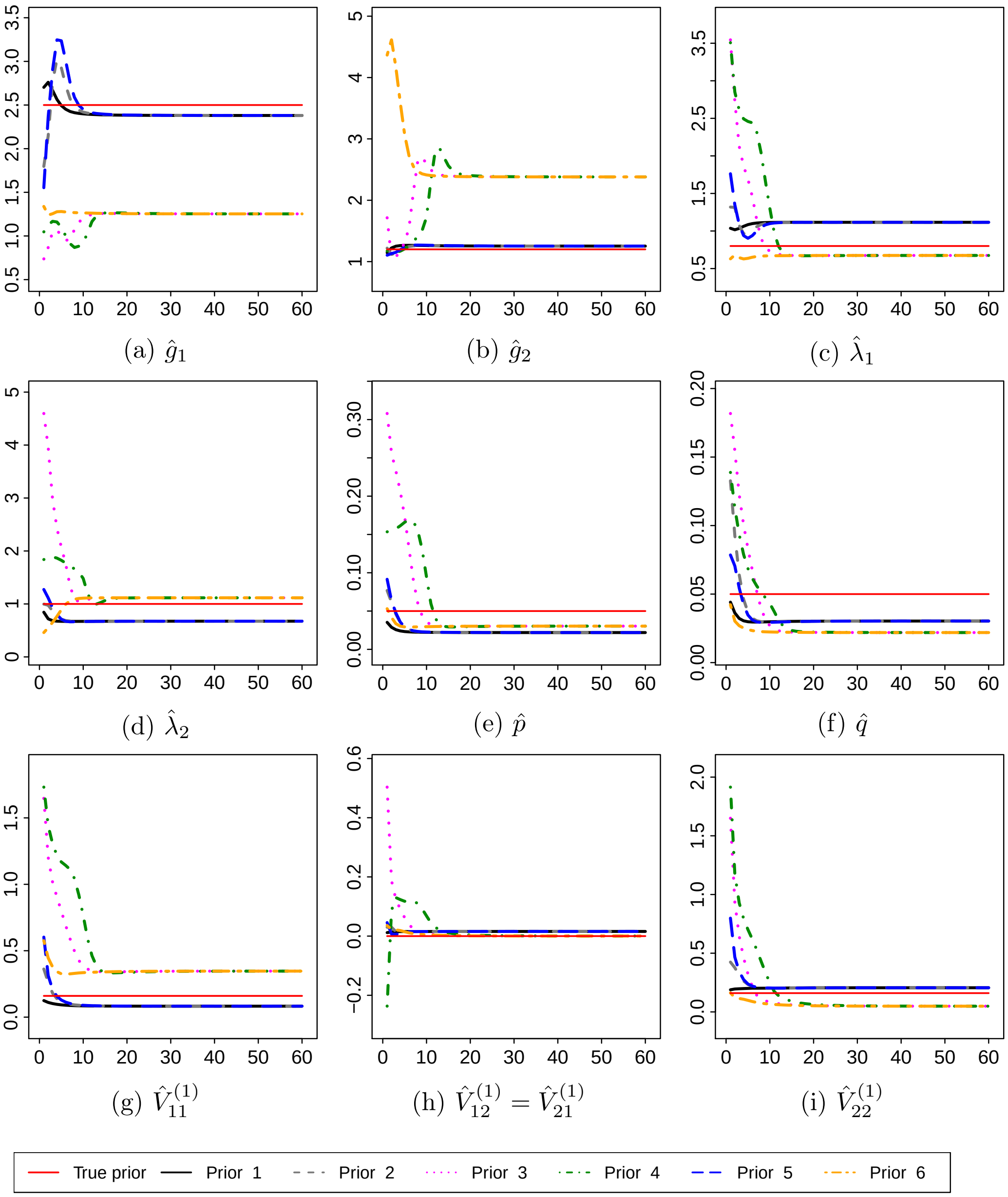
(n) $\hat{z}_2^{(1)}$

(o) $\hat{z}_1^{(2)}$

(p) $\hat{z}_2^{(2)}$

(q) Log likelihood

Figure 3.5: The estimated prior parameters $\widehat{\boldsymbol{g}}$, $\widehat{\boldsymbol{\lambda}}$, $\hat{p}$, $\hat{q}$ and $\widehat{\boldsymbol{V}}^{(1)}$ in the first 60 iterations of the EM algorithm from a selected series. The true $(p,\ q) = (0.05,\ 0.05)$



(a) $\hat{g}_1$

(b) $\hat{g}_2$

(c) $\hat{\lambda}_1$

(d) $\hat{\lambda}_2$

(e) $\hat{p}$

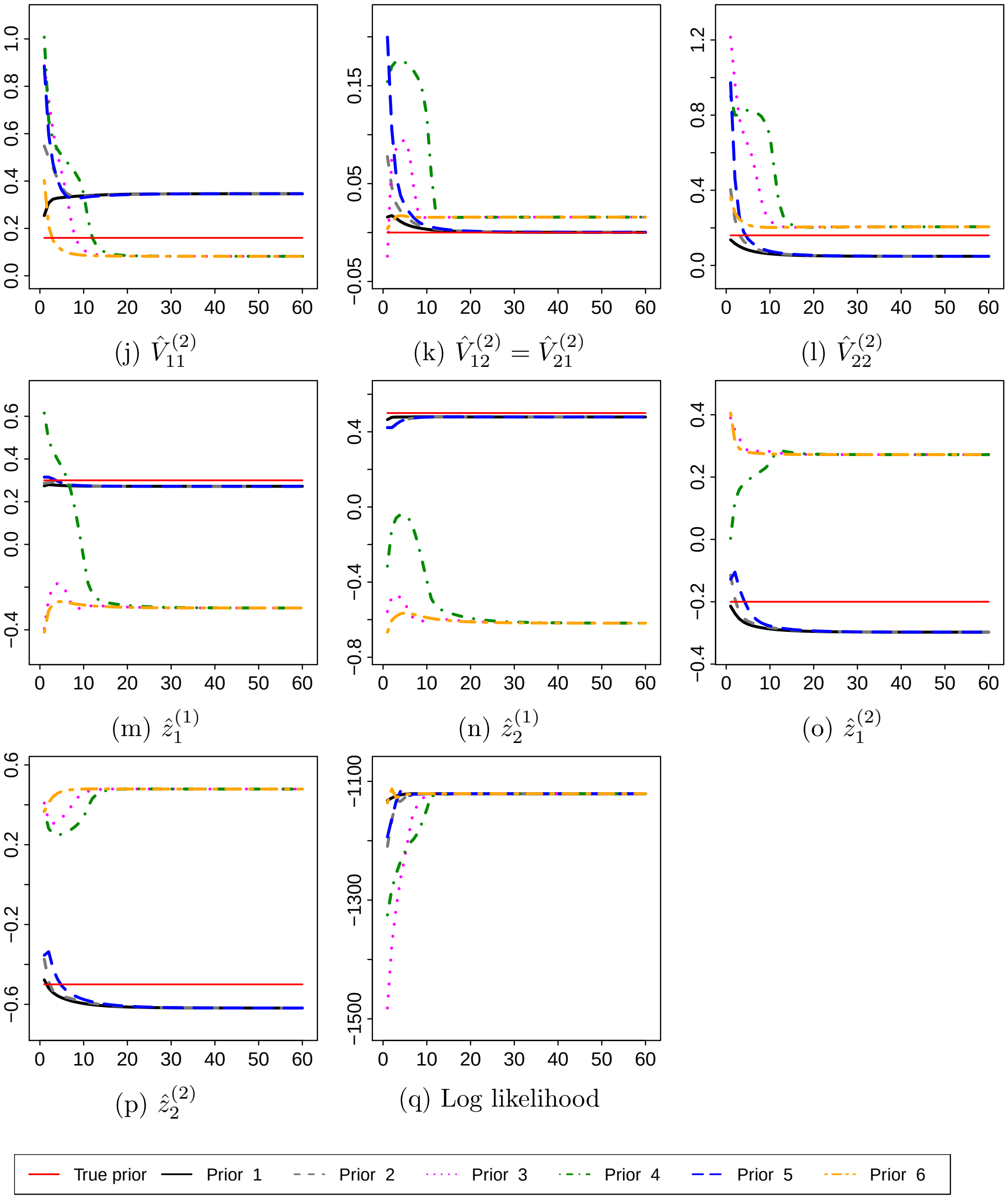
(f) $\hat{q}$

(g) $\hat{V}_{11}^{(1)}$

(h) $\hat{V}_{12}^{(1)} = \hat{V}_{21}^{(1)}$

(i) $\hat{V}_{22}^{(1)}$

Figure 3.5: Continued. The estimated prior parameters $\hat{\boldsymbol{V}}^{(2)}$ and $\hat{\boldsymbol{z}}$ and log likelihood in the first 60 iterations of EM algorithm from a selected series. The true $(p,\ q) = (0.05,\ 0.05)$



(j) $\hat{V}_{11}^{(2)}$

(k) $\hat{V}_{12}^{(2)} = \hat{V}_{21}^{(2)}$

(l) $\hat{V}_{22}^{(2)}$

(m) $\hat{z}_1^{(1)}$

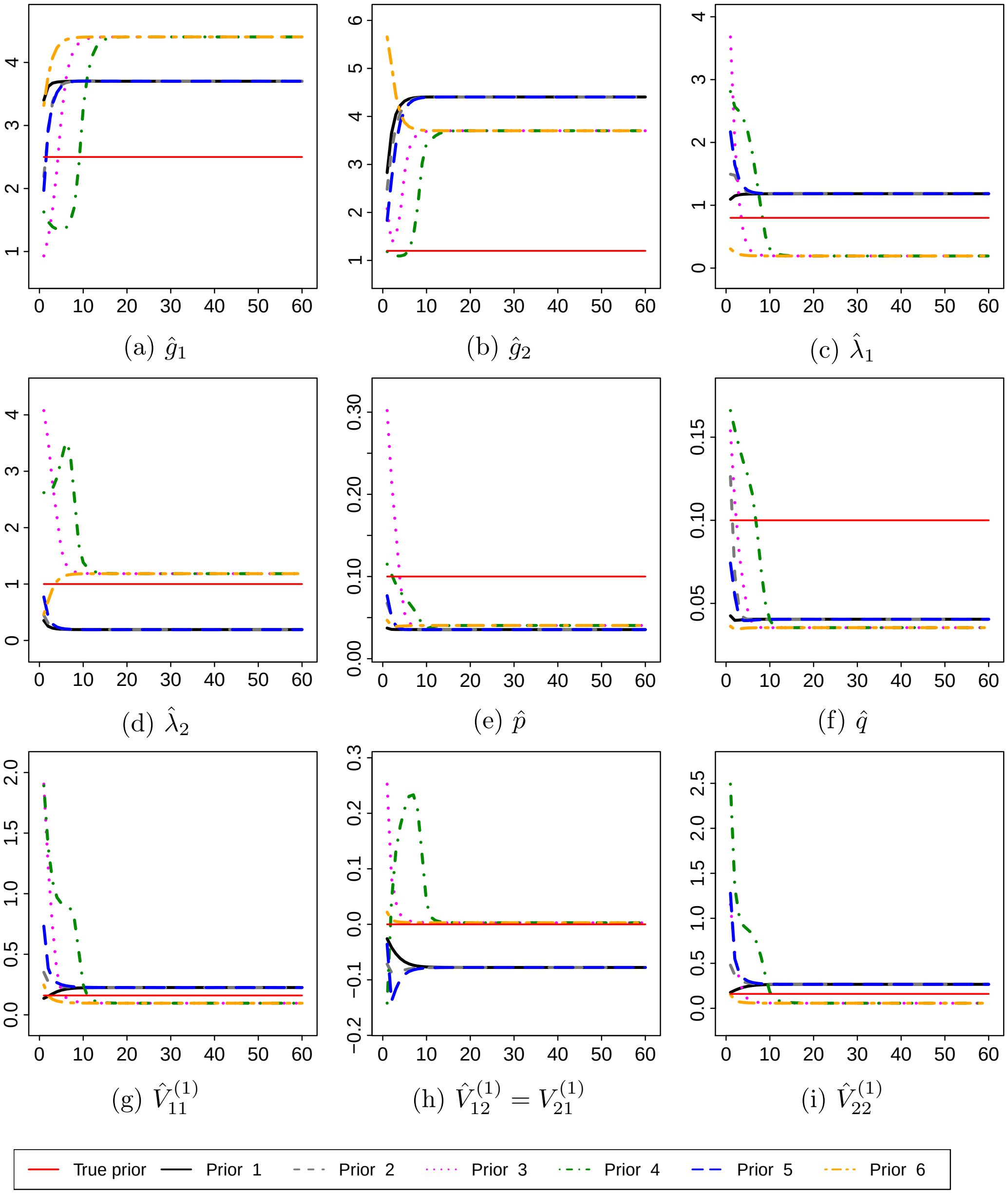
(n) $\hat{z}_2^{(1)}$

(o) $\hat{z}_1^{(2)}$

(p) $\hat{z}_2^{(2)}$

(q) Log likelihood

Figure 3.6: The estimated prior parameters $\widehat{\boldsymbol{g}}$, $\widehat{\boldsymbol{\lambda}}$, $\hat{p}$, $\hat{q}$ and $\widehat{\boldsymbol{V}}^{(1)}$ in the first 60 iterations of the EM algorithm from a selected series. The true $(p,\ q) =$ (0.1 0.1)



(a) $\hat{g}_1$

(b) $\hat{g}_2$

(c) $\hat{\lambda}_1$

(d) $\hat{\lambda}_2$

(e) $\hat{p}$

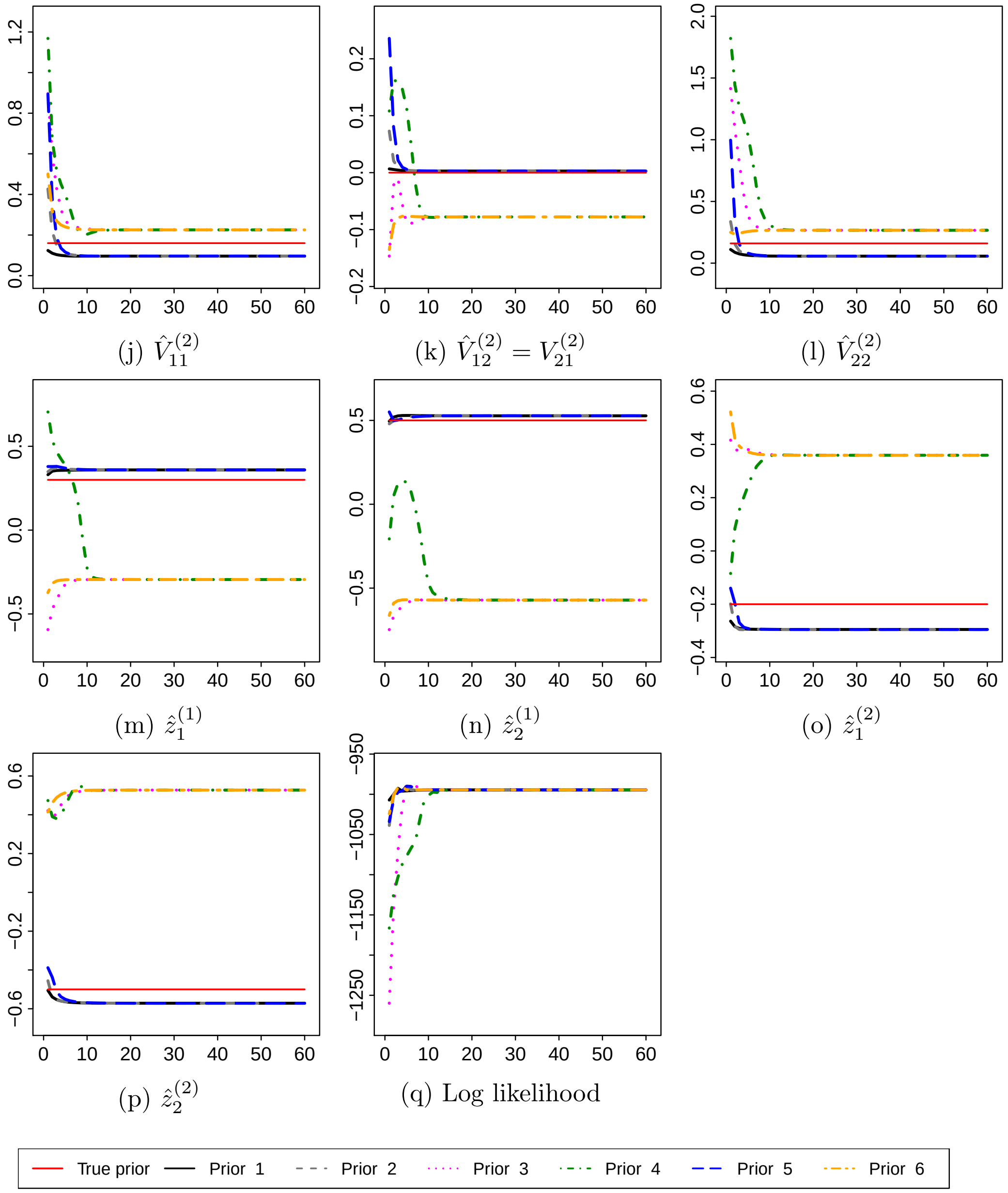
(f) $\hat{q}$

(g) $\hat{V}_{11}^{(1)}$

(h) $\hat{V}_{12}^{(1)} = V_{21}^{(1)}$

(i) $\hat{V}_{22}^{(1)}$

Figure 3.6: Continued. The estimated prior parameters $\hat{\boldsymbol{V}}^{(2)}$, $\hat{\boldsymbol{z}}$ and log likelihood in the first 60 iterations of EM algorithm from a selected series. The true $(p,\ q) = (0.1,\ 0.1)$



(j) $\hat{V}_{11}^{(2)}$

(k) $\hat{V}_{12}^{(2)} = V_{21}^{(2)}$

(l) $\hat{V}_{22}^{(2)}$

(m) $\hat{z}_1^{(1)}$

(n) $\hat{z}_2^{(1)}$

(o) $\hat{z}_1^{(2)}$

(p) $\hat{z}_2^{(2)}$

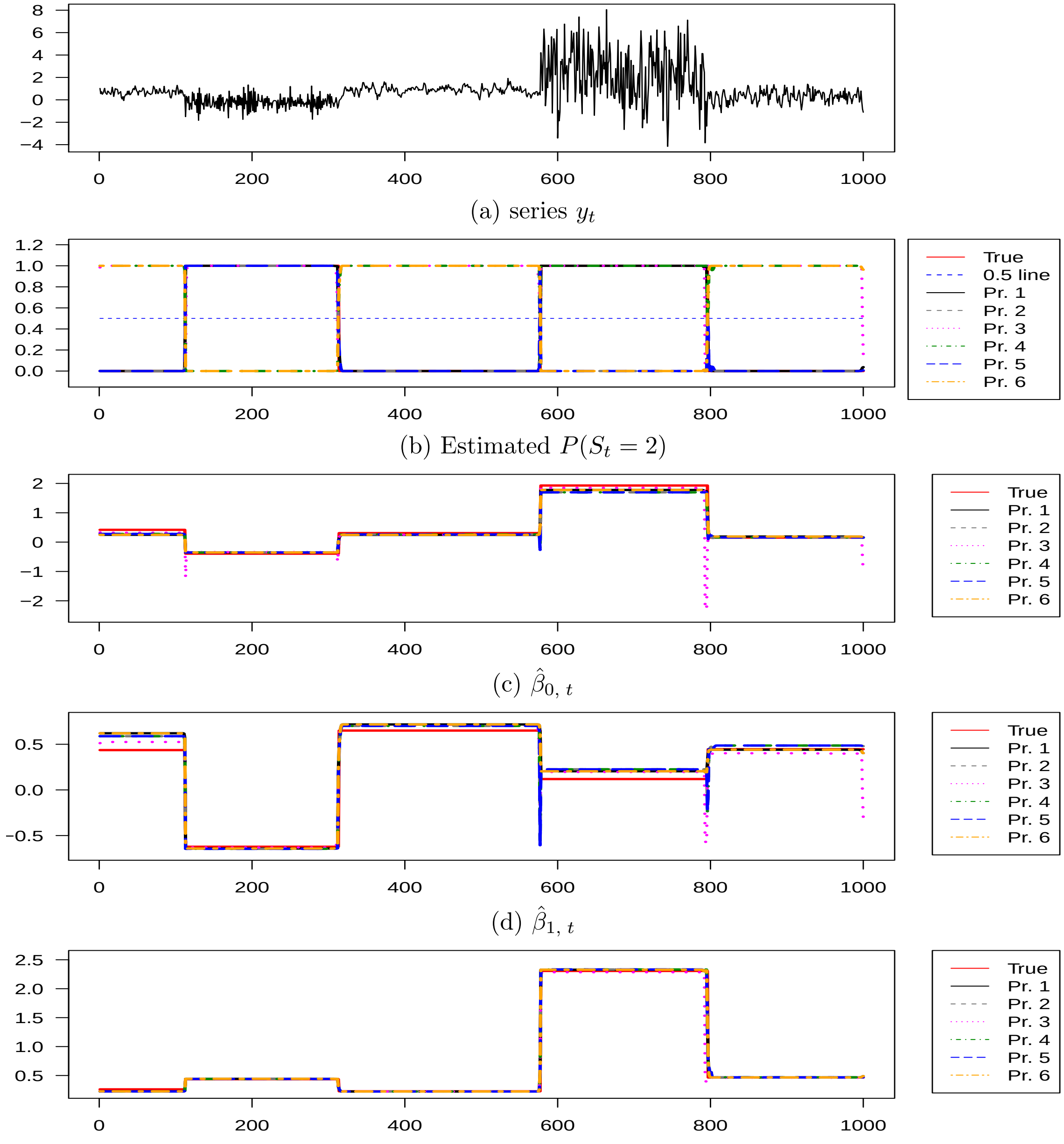
(q) Log likelihood

EM algorithm converges, it converges fast usually within 10 to 15 iterations. Therefore hyperparameters estimated by 15^{th} iteration is chosen to further estimate the regression parameters in the next subsection. In addition, it is still a good practice to begin the estimation with several random initial prior values.

3.5.2 Analysis of Model Parameters

The estimation of hyperparameters is an intermediate step to estimate the model parameters of primary interest, i.e. the probability of hidden state and the regression parameters $\boldsymbol{\beta}_t$ and σ_t at each time point. As mentioned before, we choose to estimate model parameters in each series by EM algorithm with 15 iterations for 6 different sets of initial priors. We begin with the evaluation of the impact of initial priors on the estimated probabilities of the hidden state. By assumption all the series begin with state 1 and at some point in time jump to state 2 and may switch back to state 1 and so forth. Since there are only two states and the estimated probabilities of two different states must sum up to one, it is sufficient to analyze one of the estimates such as the probabilities of state 2. Figure 3.7 – 3.11 (b) show the probabilities of being in state two estimated from 6 different sets of initial priors. Remember that the true probabilities of state 2 (solid red) begin with zero in all figures. 0.5 line (dashed blue) is used to distinguish two states. As we can see from these figures, our model may mislabel the states for some initial prior values.

Figure 3.7: The selected time series and the estimated regression parameters using priors after 15 iterations in EM algorithm. True $(p,\ q) = (0.004, 0.004)$



(a) series y_t

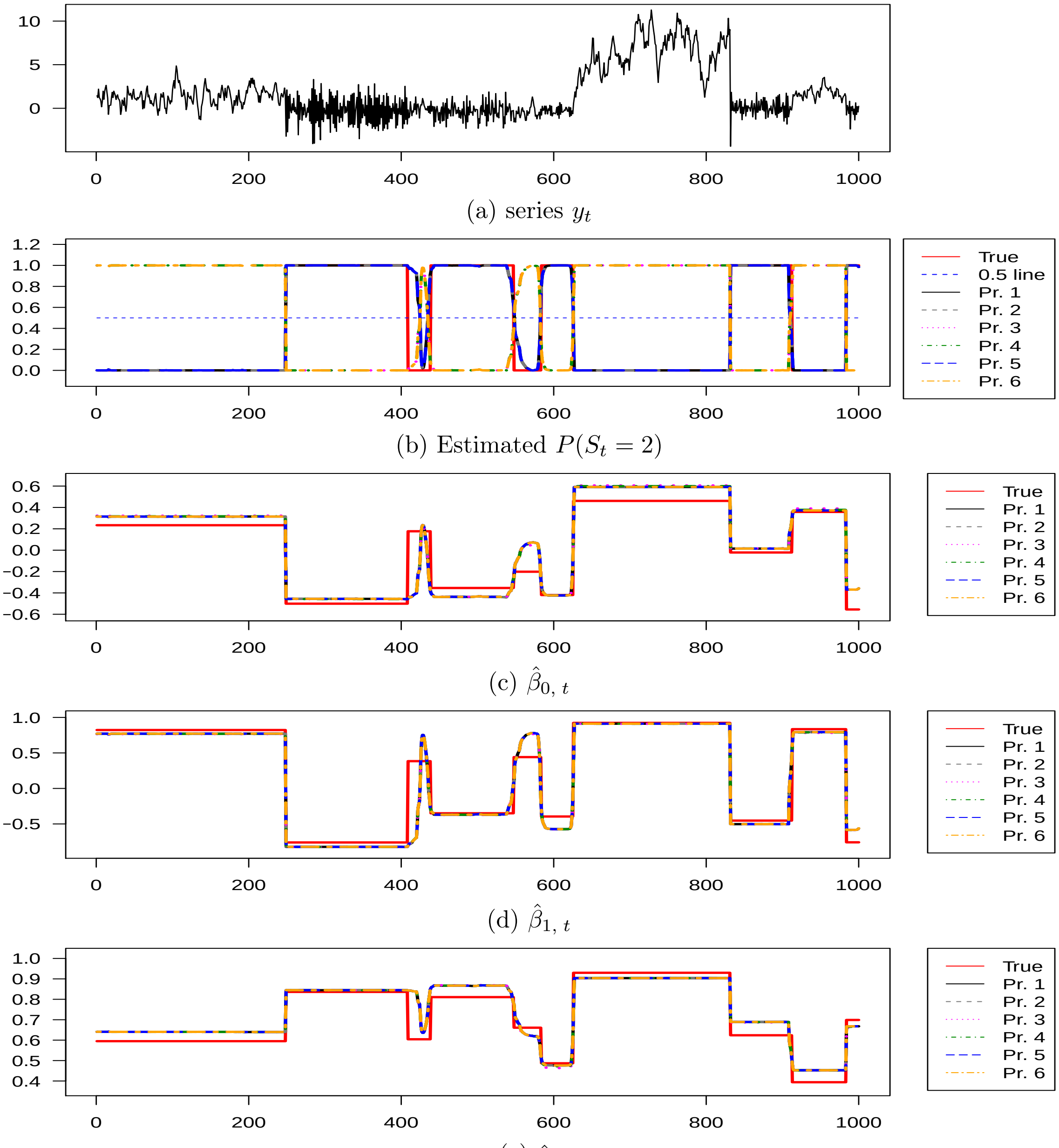
(b) Estimated $P(S_t = 2)$

(c) $\hat{\beta}_{0,\ t}$

(d) $\hat{\beta}_{1,\ t}$

(e) $\hat{\sigma}_t$

Figure 3.8: The selected time series and the estimated regression parameters using priors after 15 iterations in EM algorithm. True $(p,\ q) = (0.01, 0.01)$



(a) series y_t

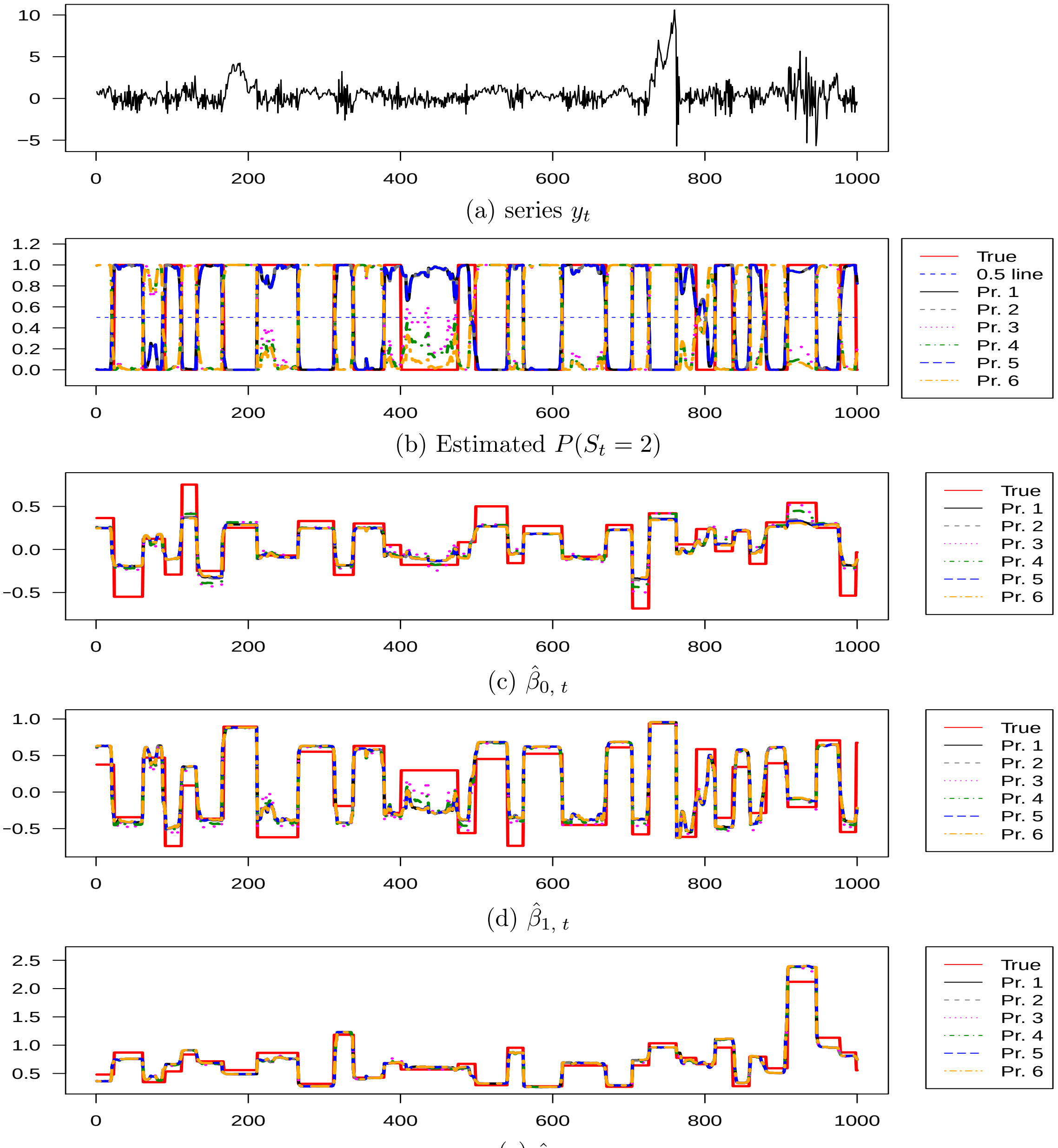
(b) Estimated $P(S_t = 2)$

(c) $\hat{\beta}_{0,\ t}$

(d) $\hat{\beta}_{1,\ t}$

(e) $\hat{\sigma}_t$

Figure 3.9: The selected time series and the estimated regression parameters using priors after 15 iterations in EM algorithm. True $(p,\ q) = (0.04, 0.08)$



(a) series y_t

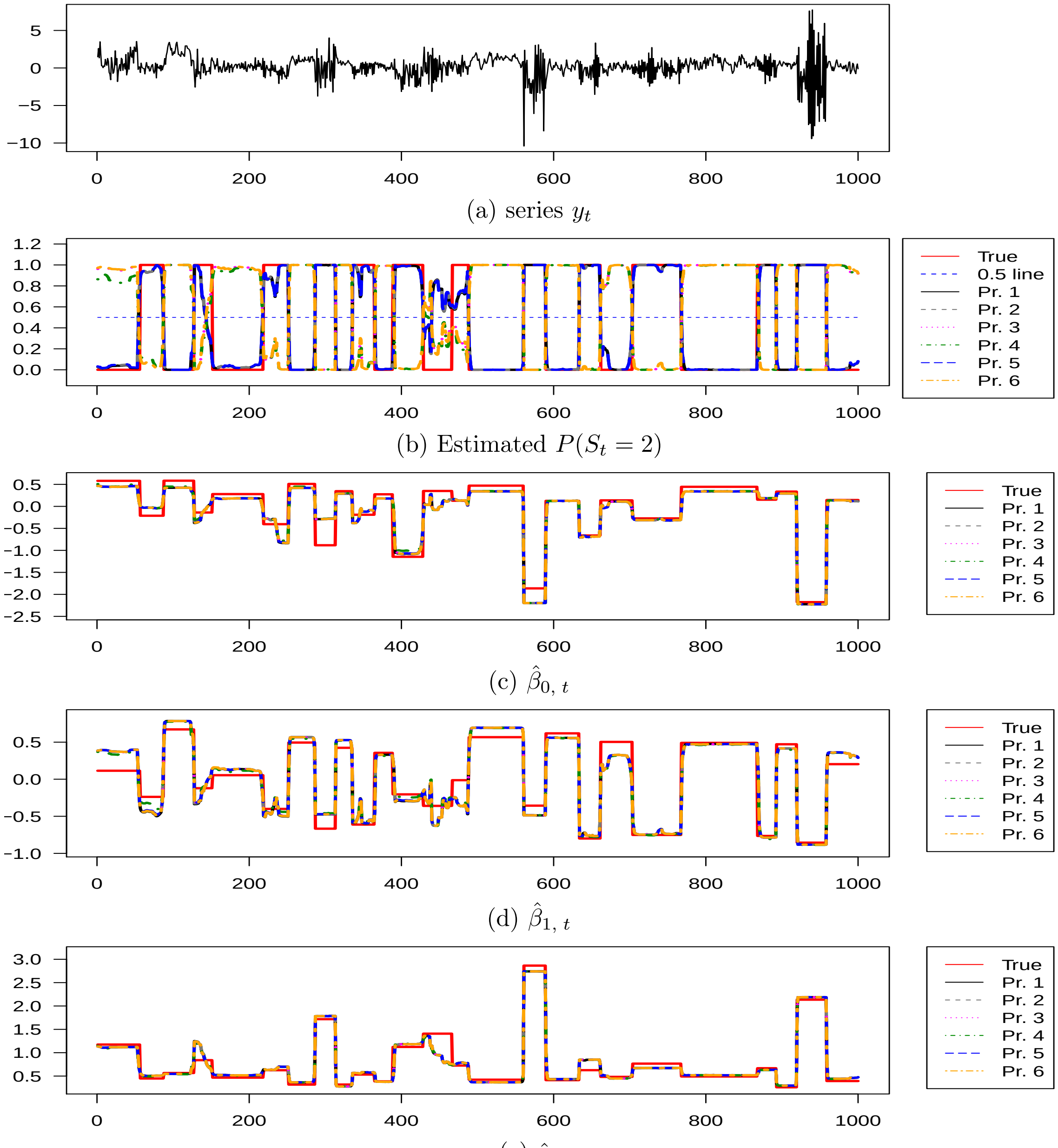
(b) Estimated $P(S_t = 2)$

(c) $\hat{\beta}_{0,\ t}$

(d) $\hat{\beta}_{1,\ t}$

(e) $\hat{\sigma}_t$

Figure 3.10: The selected time series and the estimated regression parameters using priors after 15 iteration in EM algorithm. True $(p,\ q) = (0.05, 0.05)$



(a) series y_t

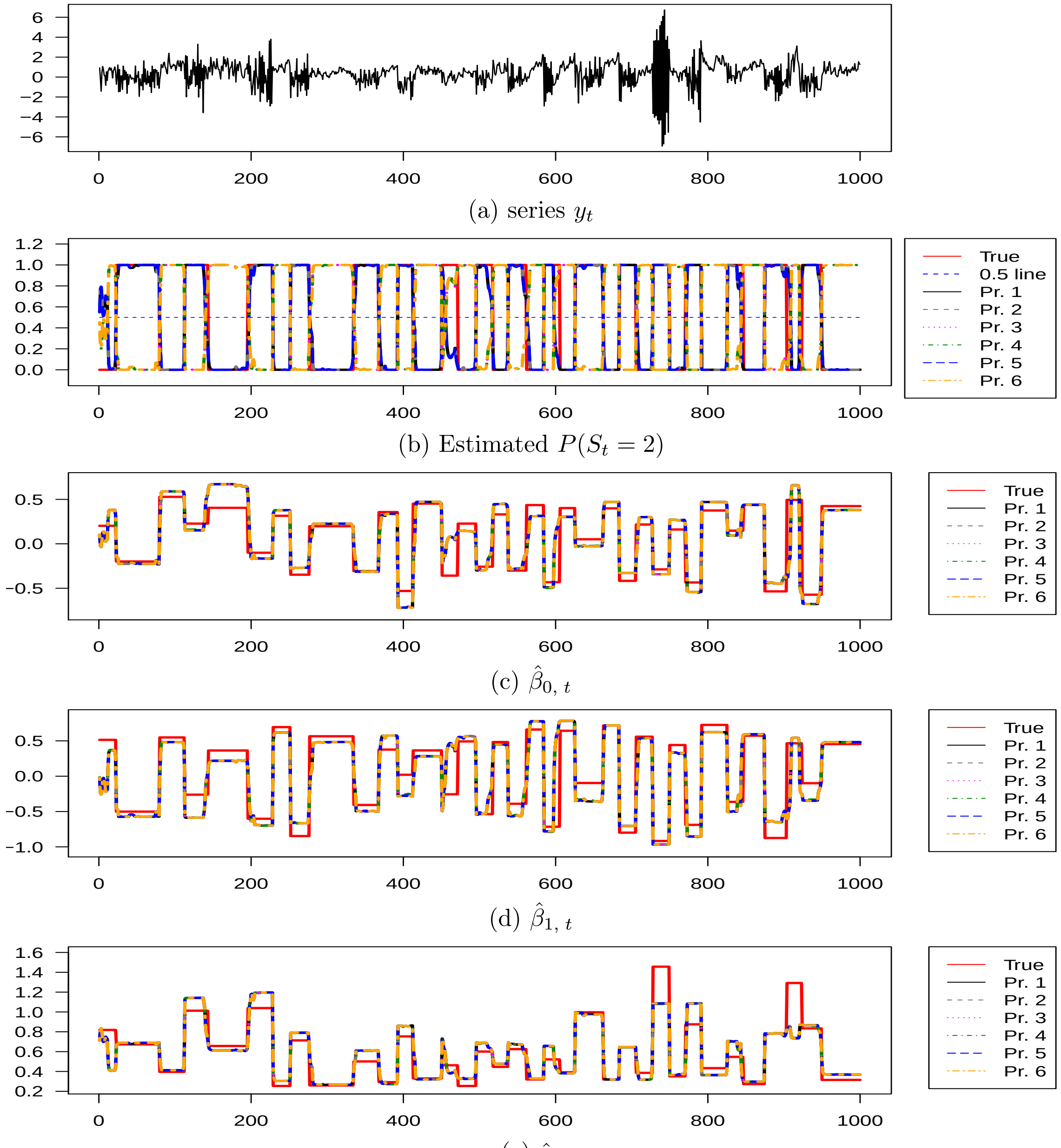
(b) Estimated $P(S_t = 2)$

(c) $\hat{\beta}_{0,\ t}$

(d) $\hat{\beta}_{1,\ t}$

(e) $\hat{\sigma}_t$

Figure 3.11: The selected time series and the estimated regression parameters using priors after 15 iterations in EM algorithm. True $(p,\ q) = (0.1, 0.1)$



(a) series y_t

(b) Estimated $P(S_t = 2)$

(c) $\hat{\beta}_{0,\ t}$

(d) $\hat{\beta}_{1,\ t}$

(e) $\hat{\sigma}_t$

Table 3.9: Monte Carlo mean of diagnostic statistics based on 500 simulations using hyperparameters estimated at the 15^{th} EM iteration for different $(p,\ q)$ scenarios and different initial prior values. Standard errors are shown in parenthesis.

(p, q)	Prior1	Prior2	Prior3	Prior4	Prior5	Prior6
(0.004, 0.004)	0.0155 (5.5e-04)	0.0171 (5.7e-04)	0.0367* (2.6e-03)	0.0228 (1.2e-03)	0.0164 (5.7e-04)	0.0155** (5.5e-04)
(0.010, 0.010)	0.0286 (6.2e-04)	0.0296 (6.8e-04)	0.0412 (1.5e-03)	0.0349 (9.4e-04)	0.0291 (6.4e-04)	0.0285 (6.1e-04)
(0.040, 0.080)	0.0757 (8.8e-04)	0.0756 (8.8e-04)	0.0823 (1.2e-03)	0.0785 (1.0e-03)	0.0756 (8.9e-04)	0.0757 (8.9e-04)
(0.050, 0.050)	0.0730 (9.4e-04)	0.0730 (9.4e-04)	0.0793 (1.3e-03)	0.0769 (1.1e-03)	0.0730 (9.4e-04)	0.0730 (9.4e-04)
(0.100, 0.100)	0.0942 (9.2e-04)	0.0942 (9.2e-04)	0.0959 (9.9e-04)	0.0968 (1.1e-03)	0.0942 (9.2e-04)	0.0944 (9.3e-04)

(a) Kullback-Leibler (KL) divergence

(p, q)	Prior1	Prior2	Prior3	Prior4	Prior5	Prior6
(0.004, 0.004)	0.9840 (7.8e-04)	0.9787 (1.1e-03)	0.9348* (3.9e-03)	0.9672 (1.7e-03)	0.9825 (8.8e-04)	0.9837** (7.7e-04)
(0.010, 0.010)	0.9598 (7.8e-04)	0.9579 (9.4e-04)	0.9318 (2.4e-03)	0.9485 (1.3e-03)	0.9592 (8.3e-04)	0.9597 (7.8e-04)
(0.040, 0.080)	0.8706 (8.1e-04)	0.8705 (8.1e-04)	0.8578 (1.5e-03)	0.8655 (1.1e-03)	0.8705 (8.1e-04)	0.8705 (8.1e-04)
(0.050, 0.050)	0.8762 (8.1e-04)	0.8760 (8.1e-04)	0.8640 (1.5e-03)	0.8705 (1.2e-03)	0.8761 (8.1e-04)	0.8761 (8.1e-04)
(0.100, 0.100)	0.8350 (7.1e-04)	0.8349 (7.1e-04)	0.8309 (1.0e-03)	0.8325 (9.8e-04)	0.8350 (7.1e-04)	0.8349 (7.1e-04)

(b) Sum of Squares of Standardized Error (SSSE)

(p, q)	Prior1	Prior2	Prior3	Prior4	Prior5	Prior6
(0.004, 0.004)	0.0771 (1.4e-03)	0.0804 (1.5e-03)	0.1104* (3.7e-03)	0.0872 (1.7e-03)	0.0790 (1.4e-03)	0.0769** (1.3e-03)
(0.010, 0.010)	0.1130 (1.4e-03)	0.1139 (1.5e-03)	0.1290 (2.2e-03)	0.1192 (1.6e-03)	0.1137 (1.5e-03)	0.1127 (1.4e-03)
(0.040, 0.080)	0.1875 (1.2e-03)	0.1870 (1.2e-03)	0.1938 (1.5e-03)	0.1902 (1.3e-03)	0.1870 (1.2e-03)	0.1873 (1.2e-03)
(0.050, 0.050)	0.1868 (1.2e-03)	0.1867 (1.2e-03)	0.1933 (1.5e-03)	0.1910 (1.4e-03)	0.1866 (1.2e-03)	0.1870 (1.2e-03)
(0.100, 0.100)	0.2135 (1.2e-03)	0.2134 (1.2e-03)	0.2152 (1.3e-03)	0.2163 (1.3e-03)	0.2134 (1.2e-03)	0.2138 (1.2e-03)

(c) L_2 norm

Table 3.9: Continued

(p, q)	Prior1	Prior2	Prior3	Prior4	Prior5	Prior6
(0.004, 0.004)	0.9869 (2.5e-03)	0.8997 (7.8e-03)	0.2656* (1.1e-02)	0.4177 (1.6e-02)	0.9393 (6.6e-03)	0.0120** (2.6e-03)
(0.010, 0.010)	0.9790 (2.4e-03)	0.9323 (6.0e-03)	0.2426 (9.5e-03)	0.5095 (1.6e-02)	0.9592 (4.7e-03)	0.0165 (1.4e-03)
(0.040, 0.080)	0.9588 (1.1e-03)	0.9592 (1.1e-03)	0.0773 (4.6e-03)	0.6070 (1.9e-02)	0.9588 (1.1e-03)	0.0410 (1.1e-03)
(0.050, 0.050)	0.9608 (1.0e-03)	0.9608 (1.0e-03)	0.0833 (5.5e-03)	0.6920 (1.8e-02)	0.9603 (1.1e-03)	0.0392 (1.0e-03)
(0.100, 0.100)	0.9464 (1.0e-03)	0.9463 (1.0e-03)	0.0592 (2.1e-03)	0.7835 (1.5e-02)	0.9463 (1.0e-03)	0.0540 (1.0e-03)

(d) Identification Ratio (IR)

For example, in Figure 3.7b, by initial prior 4 and 6, the series is estimated to start with state 2 and then at the time point where the series is supposed to change from state 1 to state 2 it switches to state 1. The same is true for the second change point, i.e. the transition position is almost correctly detected, but from the wrong state to another wrong state. State misidentification occurs for other scenarios of (p, q) as well. Although our model may mislabel the states by certain initial priors, the estimation of regression parameters, such as $\widehat{\boldsymbol{\beta}}_t$ and $\hat{\sigma}_t$ in Figure 3.7 - 3.11 (c)- (e) is not affected as much by the choice of initial prior values. For example, in Figure 3.10c, solid red line is true $\beta_{0,t}$, all other lines estimated by different initial priors tend to overlap with each other and the estimation is close enough to the true parameter values. The same is true for estimation of $\beta_{1,t}$ and σ_t for all scenarios.

Figure 3.7 - 3.11 only present the results of a few selected time series. To understand the large scale impact of initial prior values on the estimation of

model parameters, we still resort to the diagnostics in Monte Carlo simulation. Table 3.9 computes Kullback-Leibler (KL) divergence, sum of squares of standardized error (SSSE), L_2 norm and identification ratio (IR) to compare the goodness of fit for each set of initial prior values in every scenario. Star $*$ in Table 3.9 indicates that in scenario $(0.004, 0.004)$, the model is unable to produce the estimates by initial prior 3 for series 213, 294, 347, and 443 among 500 series, thus these series are eliminated for diagnostic analysis. Diagnostics labelled by double star $**$ exclude the series 141, 213, 247, 294, 347 and 443 that fail to produce the results by initial prior 6. It is clear that in Table 3.9 prior 3 and prior 4 produce the largest KLD's (a) and L_2 norm (c) and the smallest SSSE (b) for every (p, q) scenario. IR statistics are the worst by prior 3, 4 and 6, which confirms the previous results of mislabelling. In summary Monte Carlo simulation further confirms the results of the analysis in selected individual series that prior 1, 2 and 5 are better initial values and the signs of prior mean $\boldsymbol{z}$ are crucial in the choice of good initial priors.

Chapter 4

Real Data Analysis

4.1 Unemployment Rate

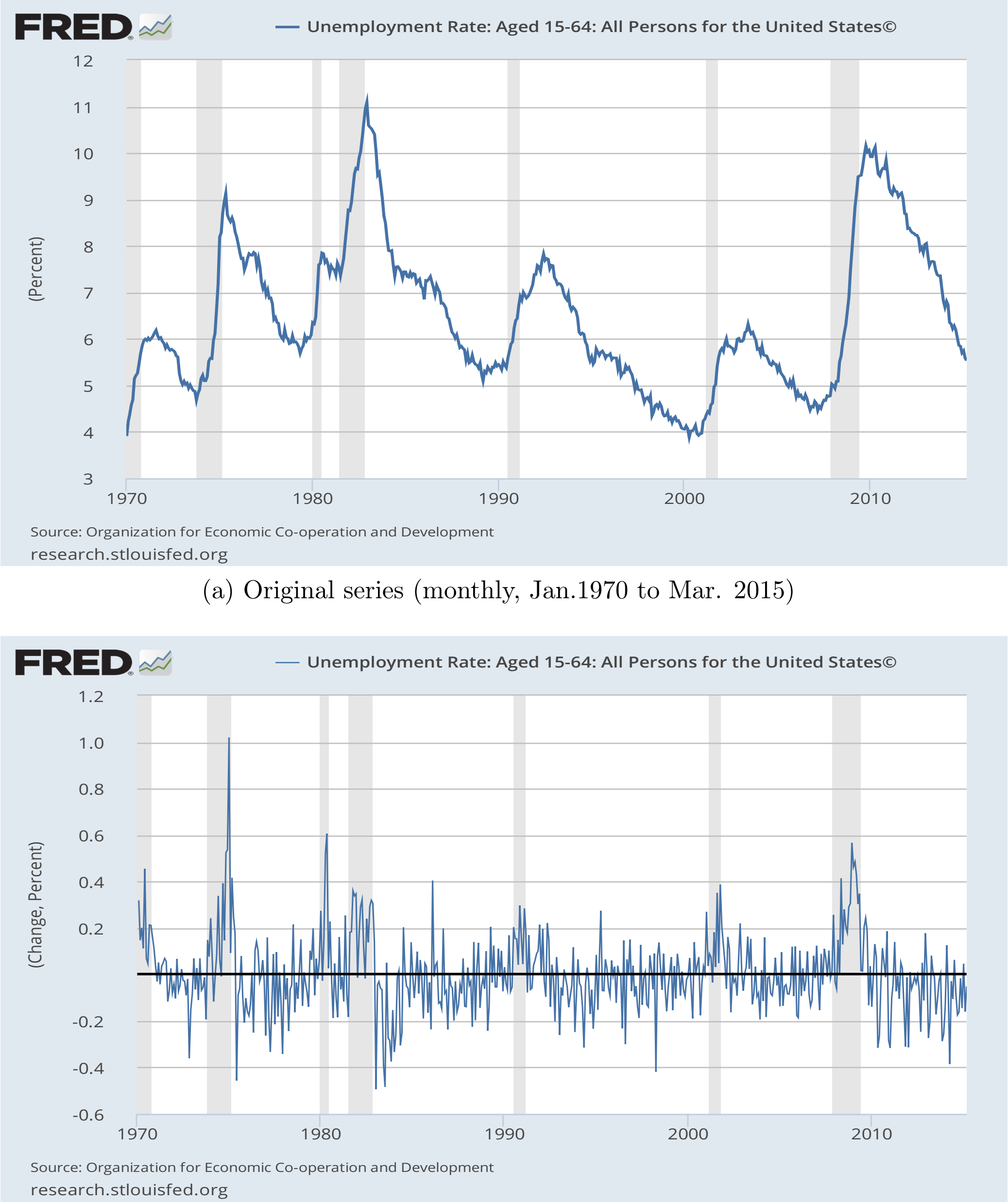
We will apply model (2.1) to several economic time series with various autoregressive orders and series lengths. We begin with the data "Unemployment Rate: Aged 15-64: All Persons for the United States" available from the website of Federal Reserve Bank of St. Louis at `https://research.stlouisfed.org/fred2/series/LRUN64TTUSM156S`. Unemployment rate is an important statistical indicator to measure the strength of the job market and economic health. Such time series have been analyzed constantly in econometric research and business cycle analysis. Figure 4.1a presents the monthly unemployment rate in percent from January 1970 to March 2015. The shaded areas indicate economic recession periods identified by The National Bureau of Economic Research (NBER). Table 4.1 lists the beginning and the ending months that the US economy is in recession according to NBER statistics from 1970 to 2015. The shortest recession period lasts only 4 months and the longest lasts 16 months. The periods between the end of a recession and the beginning of the next recession are the economic growth periods that may be called "expansion" in economic terms. The fact that expansion periods are usually longer than recession periods is a typical example of asymmetry between regimes in a time series analysis.

Table 4.1: NBER recession periods from January 1970 to December 2015

Recession periods	Starting date	Ending date
1	1970-01 (1970:Q1)	1970-11 (1970:Q4)
2	1973-12 (1973:Q4)	1975-03 (1975:Q1)
3	1980-02 (1980:Q1)	1980-07 (1980:Q3)
4	1981-08 (1981:Q3)	1982-11 (1982:Q4)
5	1990-08 (1990:Q3)	1991-03 (1991:Q1)
6	2001-04 (2001:Q2)	2001-11 (2001:Q4)
7	2008-01 (2008:Q1)	2009-06 (2009:Q2)

The shaded vertical bars in Figure 4.1a are equivalent to the above 7 recession periods. Clearly this time series is non-stationary. The unemployment rate goes up when the economy is in recession and comes down in expansion. In recession period 2 (the second shaded bar), the level of unemployment rate rises from 4.7 percent to 9.2 percent; whereas in recession period 4 (the fourth shade bar) it begins with 7.3 percent and reaches as high as 11.3 percent. To make this series stationary, at least within the periods of recession or expansion, we take the difference. Thus the new series fitting for the model analysis is $y_t = y'_t - y'_{t-1}$ where y'_t is the original unemploy-

Figure 4.1: Unemployment rate series, NBER recessions in shaded areas



(a) Original series (monthly, Jan.1970 to Mar. 2015)

(b) Change: $y_t = y'_t - y'_{t-1}$ (monthly, Feb.1970 to Mar. 2015) ment rate at time t. As can be seen from Figure 4.1b, the transformed series roughly scatters around 0. Obviously, overall difference in recession periods are much higher than that in expansion periods. For example, the difference in recession period 5 seems to be lower than that in recession period 7. Without background recession areas, it would be very difficult to visualize the pattern change at the transition between recession and expansion. One of the main tasks of this study is to find the periods of recession and expansion by analyzing unemployment rate time series alone.

A two-regime model $y_t = \beta_{0,t} + \beta_{1,t} y_{t-1} + \sigma_t \epsilon_t$ is chosen to estimate the series in Figure 4.1b, where $\epsilon_t \sim N(0,1)$. In practice, the EM algorithm is run a limited number of times. After several trial runs, the initial prior parameters are chosen to be $g^{(1)} = 2$, $g^{(2)} = 3$, $\lambda^{(1)} = 2$, $\lambda^{(2)} = 1$, $\boldsymbol{z}^{(1)\prime} = (0.2, 0.4)$, $\boldsymbol{z}^{(2)\prime} = (-0.2, -0.4)$, $\boldsymbol{V}^{(1)} = \boldsymbol{V}^{(2)} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$, and $P = \begin{pmatrix} 0.99 & 0.01 \\ 0.01 & 0.99 \end{pmatrix}$. After 25 iterations, the estimated hyperparameters are $\hat{g}^{(1)} = 39.61$, $\hat{g}^{(2)} = 4.05$, $\hat{\lambda}^{(1)} = 0.93$, $\hat{\lambda}^{(2)} = 5.03$, $\hat{\boldsymbol{z}}^{(1)\prime} = (-0.0354, -0.1034)$, $\hat{\boldsymbol{z}}^{(2)\prime} = (0.0884, 0.2585)$, $\hat{\boldsymbol{V}}^{(1)} = \begin{pmatrix} 0.0035 & -0.0028 \\ -0.0028 & 0.4768 \end{pmatrix}$, $\hat{\boldsymbol{V}}^{(2)} = \begin{pmatrix} 0.2144 & -0.0328 \\ -0.0328 & 0.3436 \end{pmatrix}$, and $\hat{P} = \begin{pmatrix} 0.98 & 0.02 \\ 0.05 & 0.95 \end{pmatrix}$. The smoothing estimation of $\beta_{i,t}$ and σ_t are shown in Figure 4.2 (b) – (d) in solid black, accompanied by 95% confidence intervals in dashed green lines and the regime estimation in (a).

Figure 4.2a shows that the estimated probabilities of series being in state

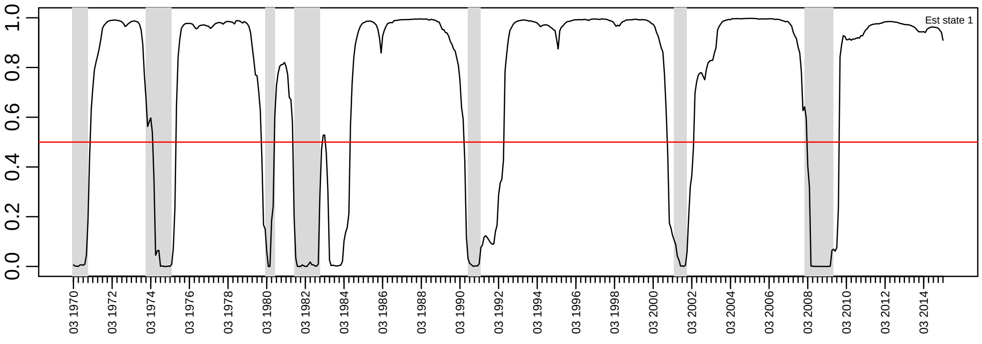
1 begins around zero from March 1970, which means state 1 is impossible to represent any economic status this period belongs to. Since the year 1970 is in recession measured by NBER, we may infer that state 1 represents economic expansion and state 2 represents economic recession. The simple criterion to identify the state is to use the 0.5 threshold. The US economy is in recession (state 2) if the estimated $P(S_t = 1) < 0.5$ according to our model. The estimated recession periods are shown in Table 4.2. The beginning and the ending dates of some of the estimated recession periods may differ a few months from the NBER statistics. Whenever our model successfully detects the changes, the estimated $P(S_t = 1)$ has to experience a gradual transition from 1 to 0 or 0 to 1. If $P(S_t)$ is estimated to be around 0.5 in the process of transition, the regime the time series belongs to at time t is actually undetermined. The simple 0.5 threshold may cause the slight difference in Table 4.1 and Table 4.2. In general, our model estimation of recession periods by a single series agrees very well with NBER statistics, which take many economic factors into consideration.

Table 4.2: Estimated recession periods for unemployment rate series

Periods	Starting date	Ending date
1	1970-03	1971-01
2	1974-05	1975-06
3	1979-12	1980-07
4	1981-08	1983-01
5	1990-06	1992-06
6	2000-12	2002-04
7	2008-03	2009-10

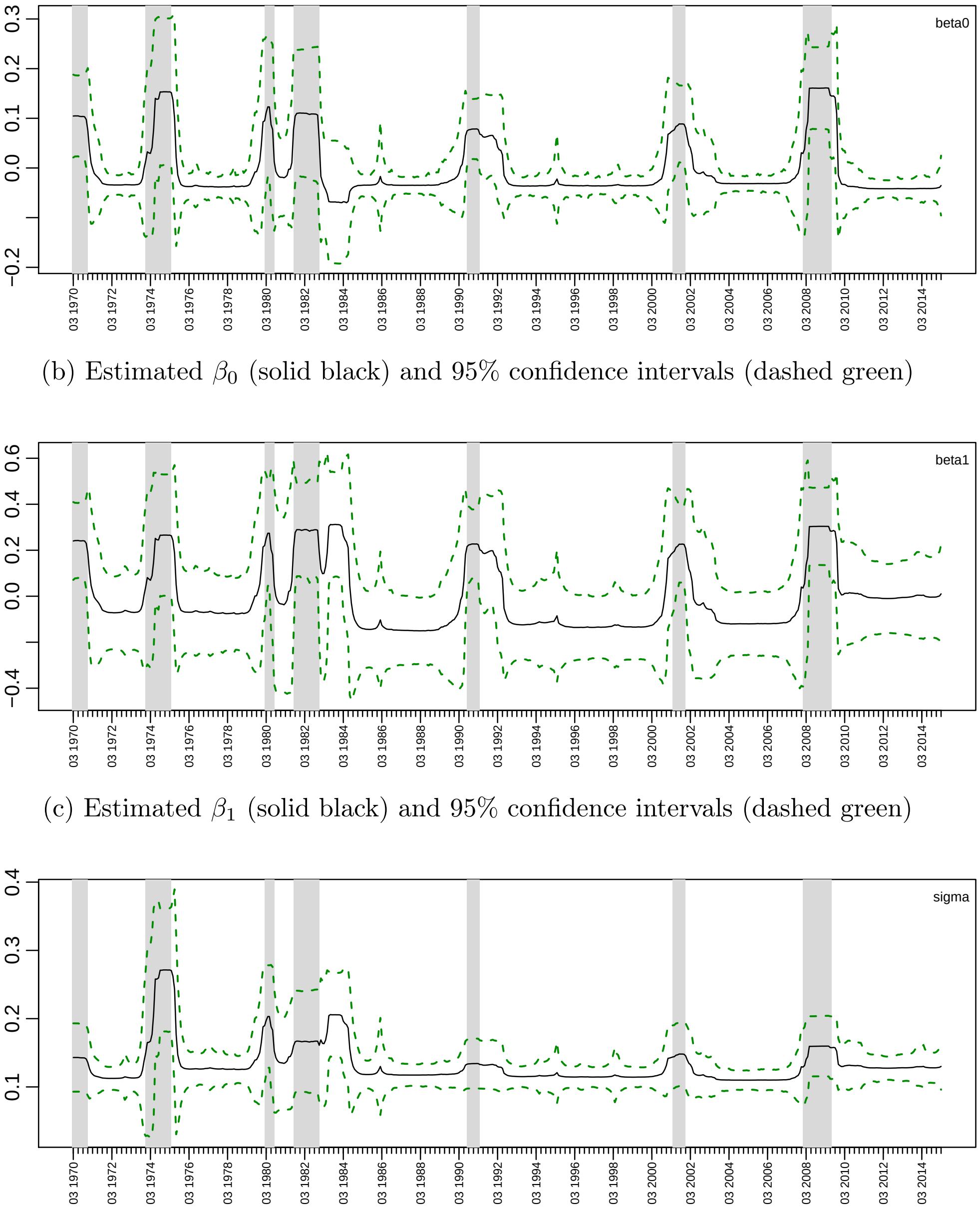
Estimation of other model parameters takes the regimes into consideration automatically – no manual work is needed to assign regimes in order to write down the estimated equations like classical MS models. In fitted equation $\hat{y}_t = \hat{\beta}_{0,t} + \hat{\beta}_{1,t}y_{t-1}$, $\hat{\beta}_{i,t}$ at every time point t is shown in Figure 4.2 (b) and (c). $\hat{\beta}_{0,t}$'s are not a constant for all estimated recession (or expansion) periods, neither are $\hat{\beta}_{1,t}$'s. In Figure 4.2 (d), $\hat{\sigma}_t$ behaves similarly. The confidence intervals (dashed green line) shows that the estimation is reasonable. Confidence intervals are wider around the transition points and narrower inside the recession (expansion) periods, indicating that the estimation is less certain when the model is in the process of detecting a switch. Finally, the probability of making a transition from expansion to recession is estimated to be 0.02 and from recession to expansion is estimated to be slightly higher at 0.05. These estimates are reasonable for the current series length and the number of switches estimated.

Figure 4.2: Estimated model parameters for unemployment rate series



(a) Estimated $P(S_t = 1)$ (solid black)

Figure 4.2: Continued



(b) Estimated β_0 (solid black) and 95% confidence intervals (dashed green)

(c) Estimated β_1 (solid black) and 95% confidence intervals (dashed green)

(d) Estimated σ (solid black) and 95% confidence intervals (dashed green)

4.2 Industrial Production: Manufacturing

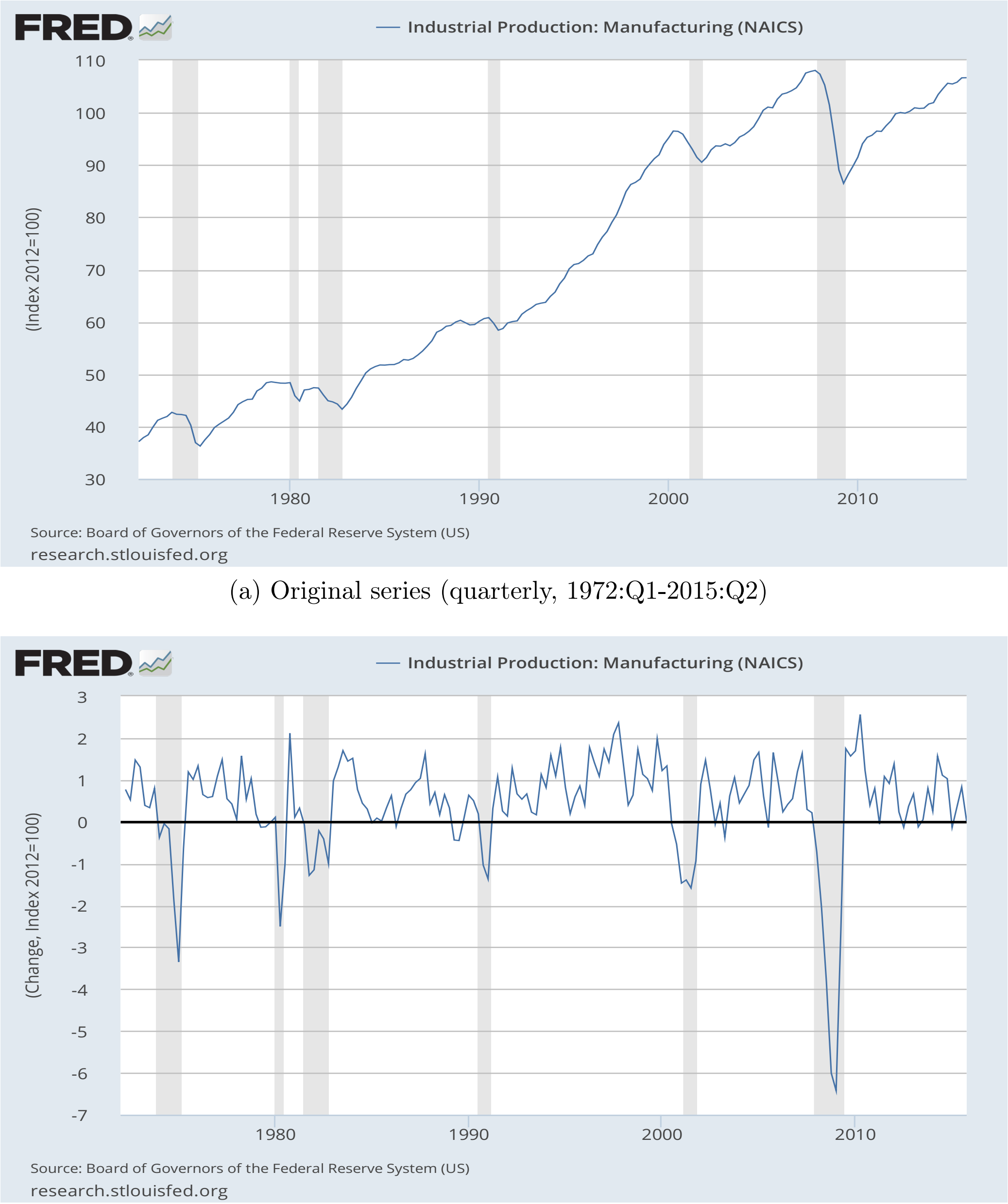
The second economic series we will analyze is that of industrial production: manufacturing series available from the website of Federal Reserve Bank of St. Louis at `https://research.stlouisfed.org/fred2/series/IPGMFSQ`. This series is very short with 176 quarterly observations from 1972:Q1 to 2015:Q4, shown in Figure 4.3a. Table 4.1 also shows NBER recession periods by quarters from 1972 to 2015. Series in (a) has an increasing trend and drops slightly during the recession periods. To make the series stationary, we difference the series by $y'_t - y'_{t-1}$ (y'_t is the original series) resulting in Figure 4.3b. The differences center around zero and seem to have larger negative values in shaded areas and small positive values in non-shaded areas.

We choose to fit the following model to the quarterly data:

$$y_t = \beta_{0,t} + \beta_{1,t}y_{t-1} + \beta_{2,t}y_{t-2} + \beta_{3,t}y_{t-3} + \beta_{4,t}y_{t-4} + \sigma_t\epsilon_t$$

where y_t is the difference in Figure 4.3b and $\epsilon_t \sim N(0,1)$. Let $K = 2$ representing two regimes in the economy (recession and expansion). Since $\boldsymbol{\beta}_t$ is a 5-dimensional vector, the prior mean is also a 5-dimensional vector and the prior variance is a 5×5 matrix. There are 46 hyperparameters to be estimated in total. Comparing with the number of observations available for modelling, the number of parameters to be estimated in this model is huge. After 60 iterations of the EM algorithm, the hyperparameters are estimated to be $\hat{g}^{(1)} = 538.59$, $\hat{g}^{(2)} = 127.95$, $\hat{\lambda}^{(1)} = 0.0029$, $\hat{\lambda}^{(2)} = 0.0047$, $\hat{p} = 0.04$,

Figure 4.3: Industrial production: manufacturing series, NBER recessions are shown in shaded areas



Source: Board of Governors of the Federal Reserve System (US)
research.stlouisfed.org

(a) Original series (quarterly, 1972:Q1-2015:Q2)

Source: Board of Governors of the Federal Reserve System (US)
research.stlouisfed.org

(b) Change: $y_t = y'_t - y'_{t-1}$ (quarterly, 1972:Q2-2015:Q2)

$\hat{q} = 0.19$, $\hat{\boldsymbol{z}}^{(1)} = (0.4927, 0.4335, -0.0931, 0.0327, -0.0467)'$,

$\hat{\boldsymbol{z}}^{(2)} = (-0.3827, 0.7727, -0.5576, 0.1365, -0.4260)'$,

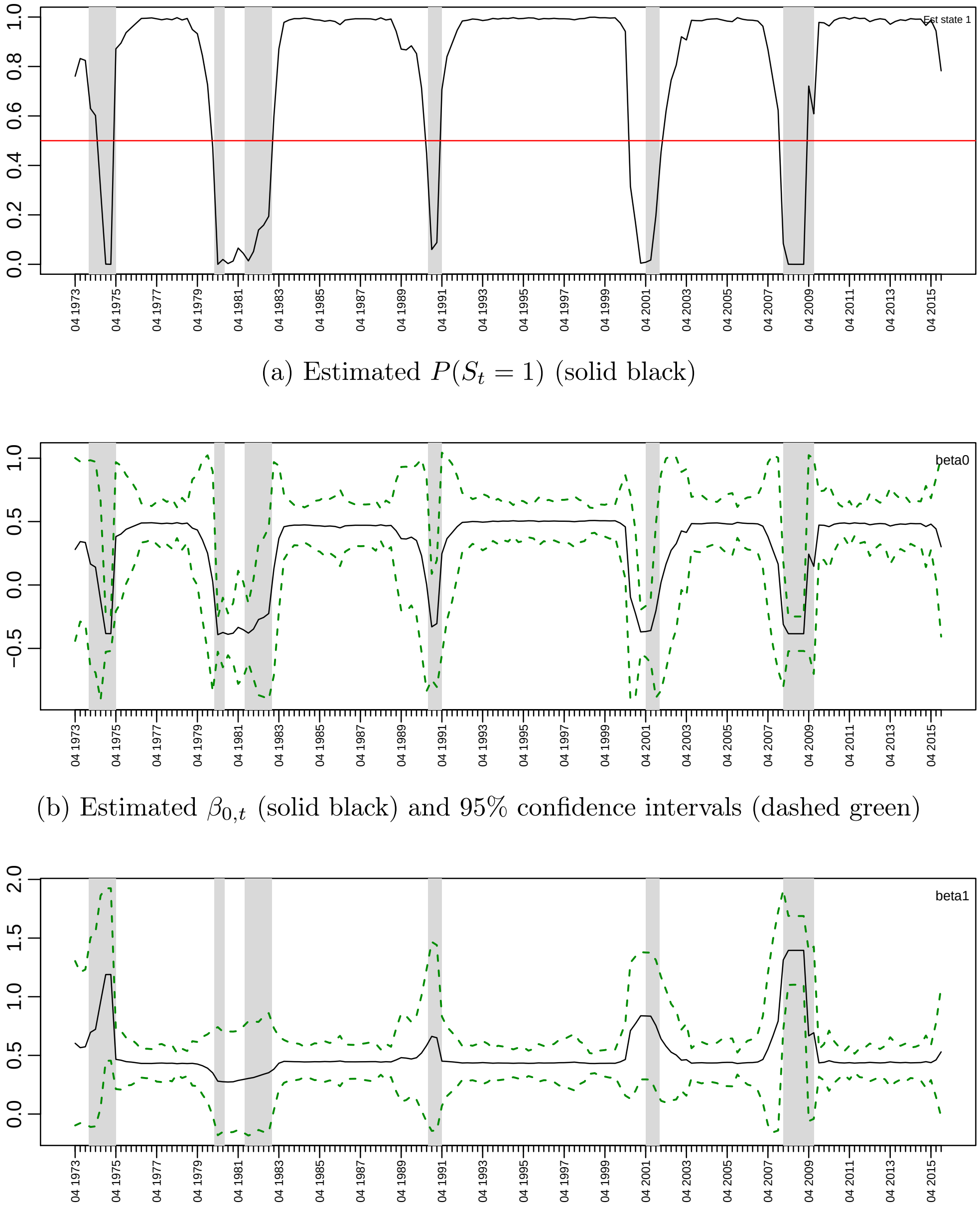
$$\hat{\boldsymbol{V}}^{(1)} = \begin{pmatrix} 0.0066 & -0.0019 & -0.0003 & 7.2e-5 & 0.0005 \\ -0.0019 & 0.0036 & -0.0010 & 0.0001 & -0.0007 \\ -0.0003 & -0.0010 & 0.0026 & -0.0010 & -1.2e-6 \\ 7.2e-5 & 0.0001 & -0.0010 & 0.0024 & -0.0009 \\ 0.0005 & -0.0007 & -1.2e-6 & -0.0009 & 0.0389 \end{pmatrix},$$

and

$$\hat{\boldsymbol{V}}^{(2)} = \begin{pmatrix} 0.0058 & 0.0025 & 0.0009 & 2.7e-5 & 0.0003 \\ 0.0025 & 0.3026 & -0.0025 & -0.0008 & -0.0004 \\ 0.0009 & -0.0025 & 0.0069 & -0.0018 & 0.0013 \\ 2.7e-5 & -0.0008 & -0.0018 & 0.0050 & -0.0016 \\ 0.0003 & -0.0004 & 0.0013 & -0.0016 & 0.0052 \end{pmatrix}.$$

Figure 4.4a shows the estimated probabilities of regimes being in state 1. We still use the criterion that the series is in regime 1 if the estimated $P(S_t = 1) > 0.5$. The estimates in Figure 4.4a begin high close to 0.85 in expansion period (non-shaded area). Thus we may name regime 1 as expansion and regime 2 as recession. By 0.5 threshold, the estimated recession periods from Figure 4.4a are 1974:Q3 - 1975:Q1, 1980:Q1 - 1982:Q4, 1990:Q3 - 1991:Q1, 2000:Q3 - 2002:Q1 and 2008:Q1 - 2009:Q1. Compar-

Figure 4.4: Estimated model parameters of industrial production series

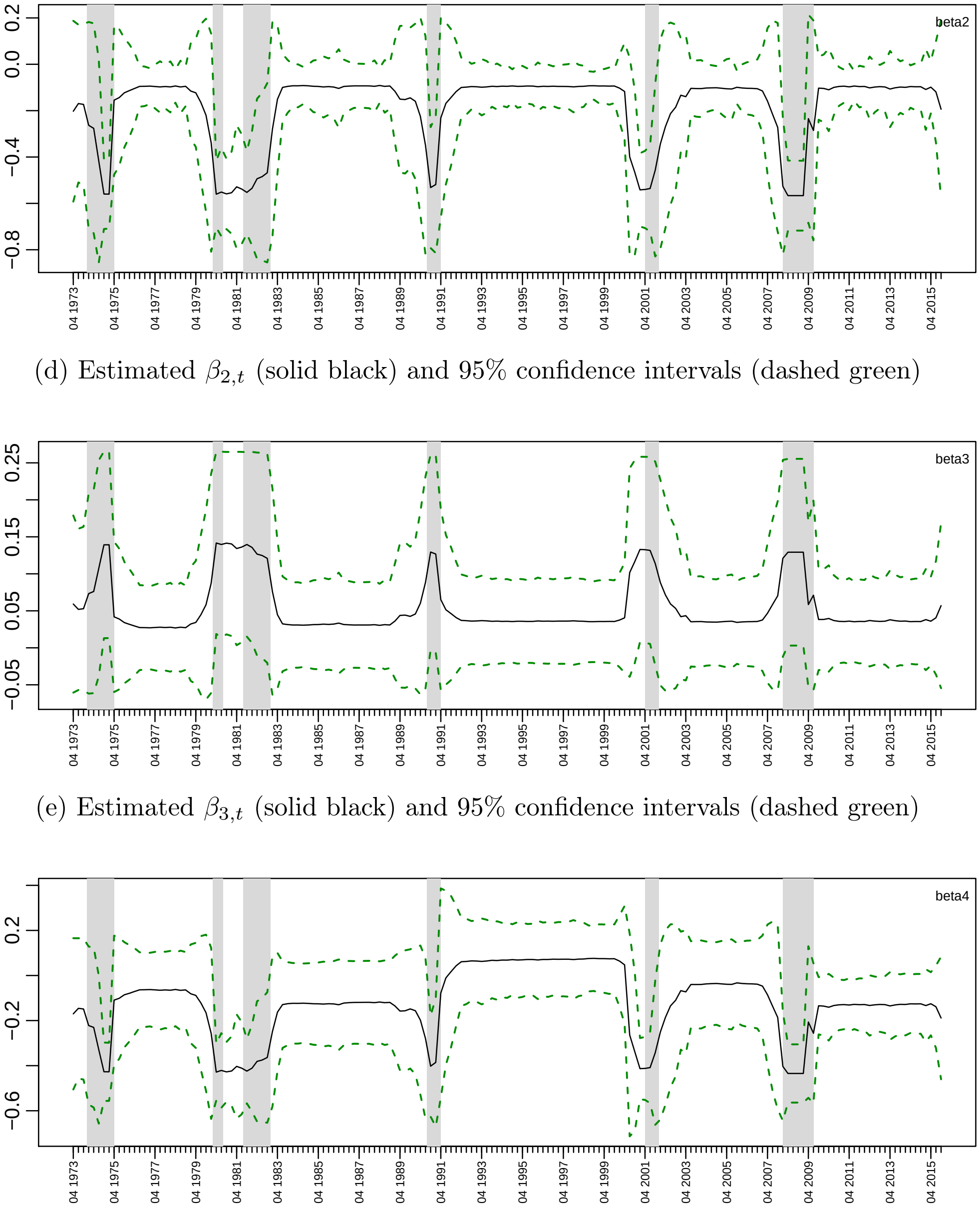


(a) Estimated $P(S_t = 1)$ (solid black)

(b) Estimated $\beta_{0,t}$ (solid black) and 95% confidence intervals (dashed green)

(c) Estimated $\beta_{1,t}$ (solid black) and 95% confidence intervals (dashed green)

Figure 4.4: Continued

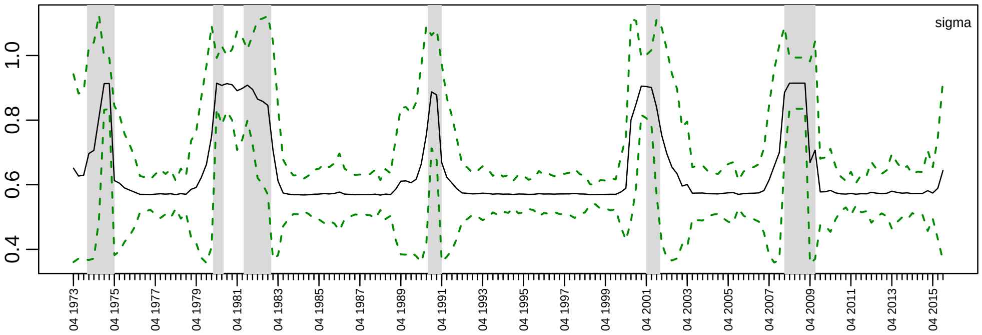


(d) Estimated $\beta_{2,t}$ (solid black) and 95% confidence intervals (dashed green)

(e) Estimated $\beta_{3,t}$ (solid black) and 95% confidence intervals (dashed green)

(f) Estimated $\beta_{4,t}$ (solid black) and 95% confidence intervals (dashed green)

Figure 4.4: Continued



(g) Estimated σ_t (solid black) and 95% confidence intervals (dashed green)

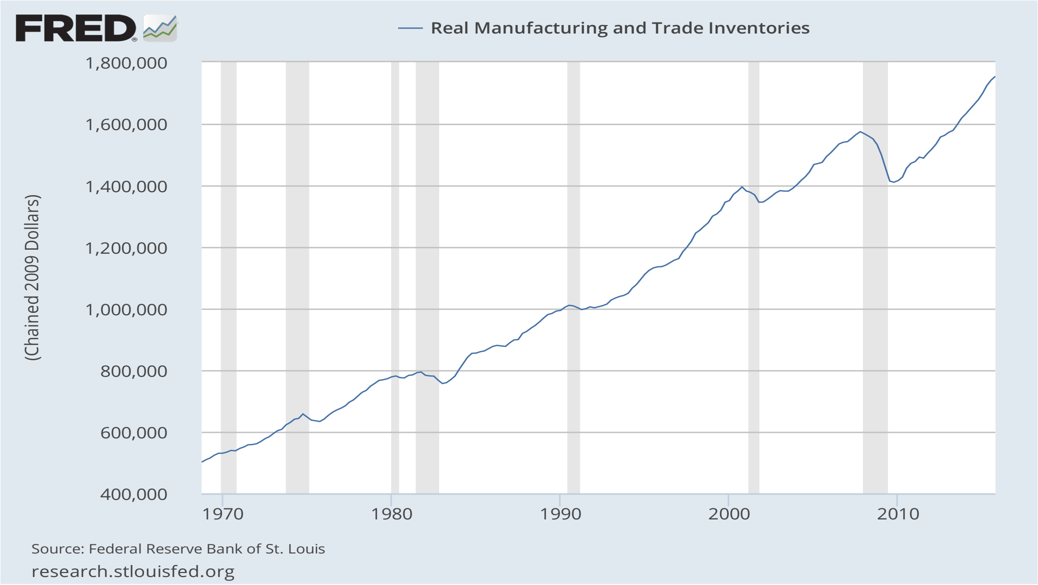
ing with Table 4.1, the first estimated recession period begins two quarters later than NBER statistics. The second estimated recession period joins two recession periods 1980:Q1-1980:Q3 and 1981:Q3-1981:Q4 in NBER record into one recession period. Since there are only 3 observations in between the two recession periods, little information is known to estimate two regime changes in three quarters. The third and fifth estimated recession periods coincide with NBER statistics and the fourth recession period is estimated to be longer, beginning 3 quarters earlier and ending one quarter later than the NBER record. In general our estimation of recession periods agrees well with the NBER statistics. Other estimates including $\hat{\beta}_{0,t}$, $\hat{\beta}_{1,t}$, $\hat{\beta}_{2,t}$, $\hat{\beta}_{3,t}$, $\hat{\beta}_{4,t}$ and $\hat{\sigma}_t$ and their confidence intervals are shown in Figure 4.4 (b) to (g) respectively. 95% confidence intervals show that $\beta_{0,t}$ and $\beta_{1,t}$ are significantly different from zero in each regime and $\beta_{2,t}$, $\beta_{3,t}$ and $\beta_{4,t}$ in recession periods are significantly different from zero. In addition, the estimated β_i's are not necessarily the same in the same regime. For example, in expansion period (1991:Q2 - 2000:Q2) $\beta_{4,t}$ is estimated to be around 0.1 on average; whereas in 2009:Q2 - 2015:Q2, it is estimated to be -0.1 on average.

4.3 Real Manufacturing and Trade Inventory

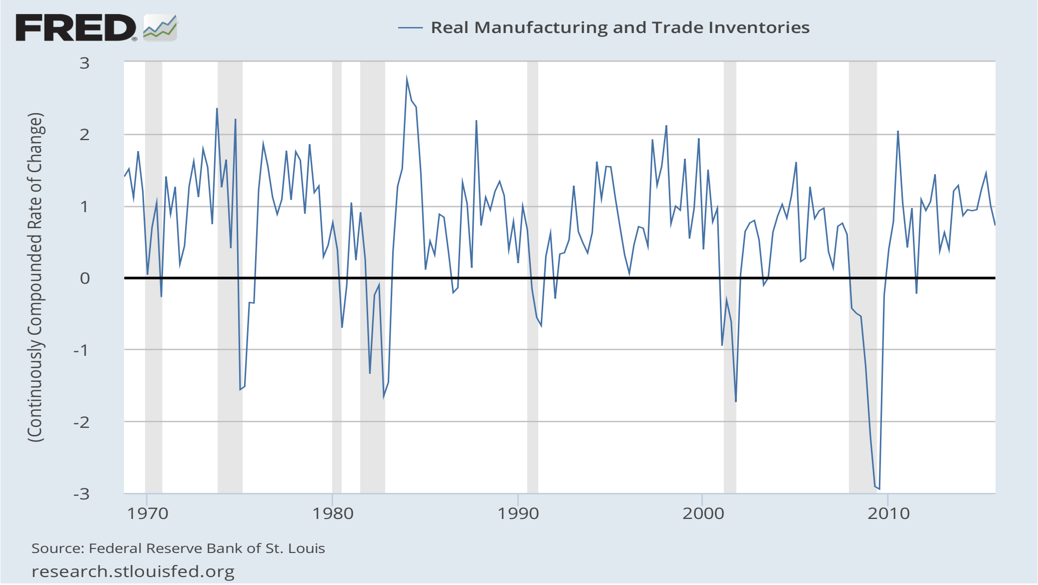
The third and the last economic series we will analyze is that of real manufacturing and trade inventory from 1968:Q4 to 2015:Q4. It can be downloaded from the website of Federal Reserve Bank of St. Louis at `https://research.stlouisfed.org/fred2/series/INVCQRMTSPL`. This series is also short with only 189 quarterly observations. Figure 4.5a shows that this series increases from 460,330 to 1,754,238 and drops slightly at the recession periods indicated in the shaded areas. It is a convention to transform the aggregated values by continuously compounded rate of change to make the series stationary. Continuously compounded rate of change can be computed as $y_t = 100 \times \ \ln \frac{y'_t}{y'_{t-1}}$, shown in Figure 4.5b where y'_t is the original inventory series. y_t ranges between -3 and 3. When series is in recession period, the values tends to be more negative and vice versa.

To choose a proper model, I have tried models with various autoregressive orders from 1 to 4. Only AR(4) model is able to successfully find all recession periods, other models either detect no regime switching or detect the change points partially . For quarterly economic time series, it is usually reasonable to consider annual effects in the AR(4) model. Therefore we use the following

Figure 4.5: Real manufacturing and trade inventory series, NBER recessions are shown in shaded areas



(a) Original series (quarterly, 1968:Q4 - 2015:Q4)



(b) Continuously compounded rate of change (quarterly, 1969:Q1-2015:Q4) model for the manufacturing and trade inventory series:

$$y_t = \beta_{0,t} + \beta_{1,t}y_{t-1} + \beta_{2,t}y_{t-2} + \beta_{3,t}y_{t-3} + \beta_{4,t}y_{t-4} + \sigma_t\epsilon_t$$

where $\epsilon_t \sim N(0,1)$ and y_t is the continuously compounded rate of change. After a careful selection of initial prior values and running 40 iterations in EM algorithm, the estimated hyperparameters are as follows: $\hat{g}^{(1)} = 19.05$, $\hat{g}^{(2)} = 24.21$, $\hat{\lambda}^{(1)} = 0.10$, $\hat{\lambda}^{(2)} = 0.47$, $\hat{p} = 0.05$, $\hat{q} = 0.30$, $\hat{\boldsymbol{z}}^{(1)\prime} = (0.6503, 0.2506, 0.1263, 0.0840, -0.1841)$,
$\hat{\boldsymbol{z}}^{(2)\prime} = (-1.1808, 0.1167, -0.3183, 0.5983, -0.5424)$,

$$\hat{\boldsymbol{V}}^{(1)} = \begin{pmatrix} 0.0144 & -0.0056 & 0.0022 & -0.0013 & -0.0027 \\ -0.0056 & 0.0327 & -0.0134 & -0.0005 & -0.0067 \\ -0.0022 & -0.0013 & 0.0327 & -0.0024 & -0.0001 \\ -0.0013 & -0.0005 & -0.0024 & 0.0050 & -0.0013 \\ -0.0027 & -0.0067 & -0.0001 & -0.0013 & 0.0195 \end{pmatrix},$$

and

$$\hat{\boldsymbol{V}}^{(2)} = \begin{pmatrix} 4.4641 & 0.9154 & -0.3267 & 0.0103 & -0.1364 \\ 0.9154 & 4.9120 & 0.0498 & 0.0381 & 0.0579 \\ -0.3267 & 0.0498 & 0.5340 & -0.0118 & 0.0451 \\ 0.0103 & 0.0381 & -0.0118 & 0.07440 & -0.0352 \\ -0.1364 & 0.0579 & 0.0451 & -0.0352 & 0.3937 \end{pmatrix}.$$

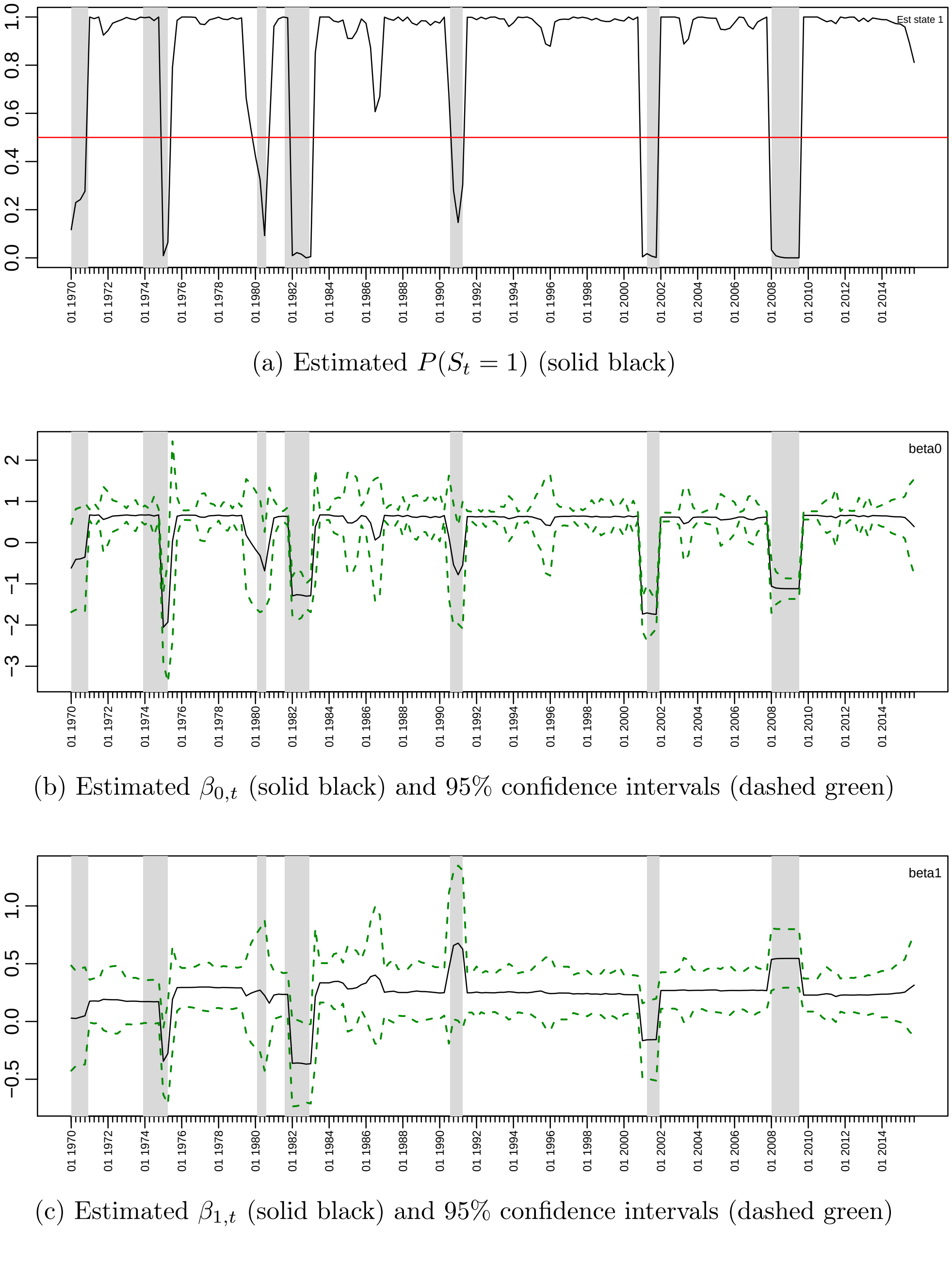
Figure 4.6a shows the estimated probabilities of series being in regime 1. The estimated probabilities begins low in 1970:Q1 and rise up quickly to 1 in 1970:Q4. Since NBER labels period 1969:Q4 – 1970:Q4 as recessio, We may induce that regime 1 represents economic expansion and regime 2 represents economic recession. Our model predicts 7 recession periods listed in more details in Table 4.3 by comparing estimated $P(S_t = 1)$ (Figure 4.4a) with threshold 0.5.

Table 4.3: Estimated recession periods from manufacturing and trade inventory series

Recession periods	Starting date	Ending date
1	1970:Q1	1970:Q4
2	1975:Q1	1975:Q2
3	1980:Q1	1980:Q3
4	1982:Q1	1983:Q1
5	1990:Q4	1991:Q2
6	2001:Q1	2001:Q4
7	2008:Q1	2009:Q3

Seven recession periods are clearly detected by the model. The second estimated recession period seems to begin and end later and shorter than

Figure 4.6: Estimated model parameters of manufacturing and trade inventory series

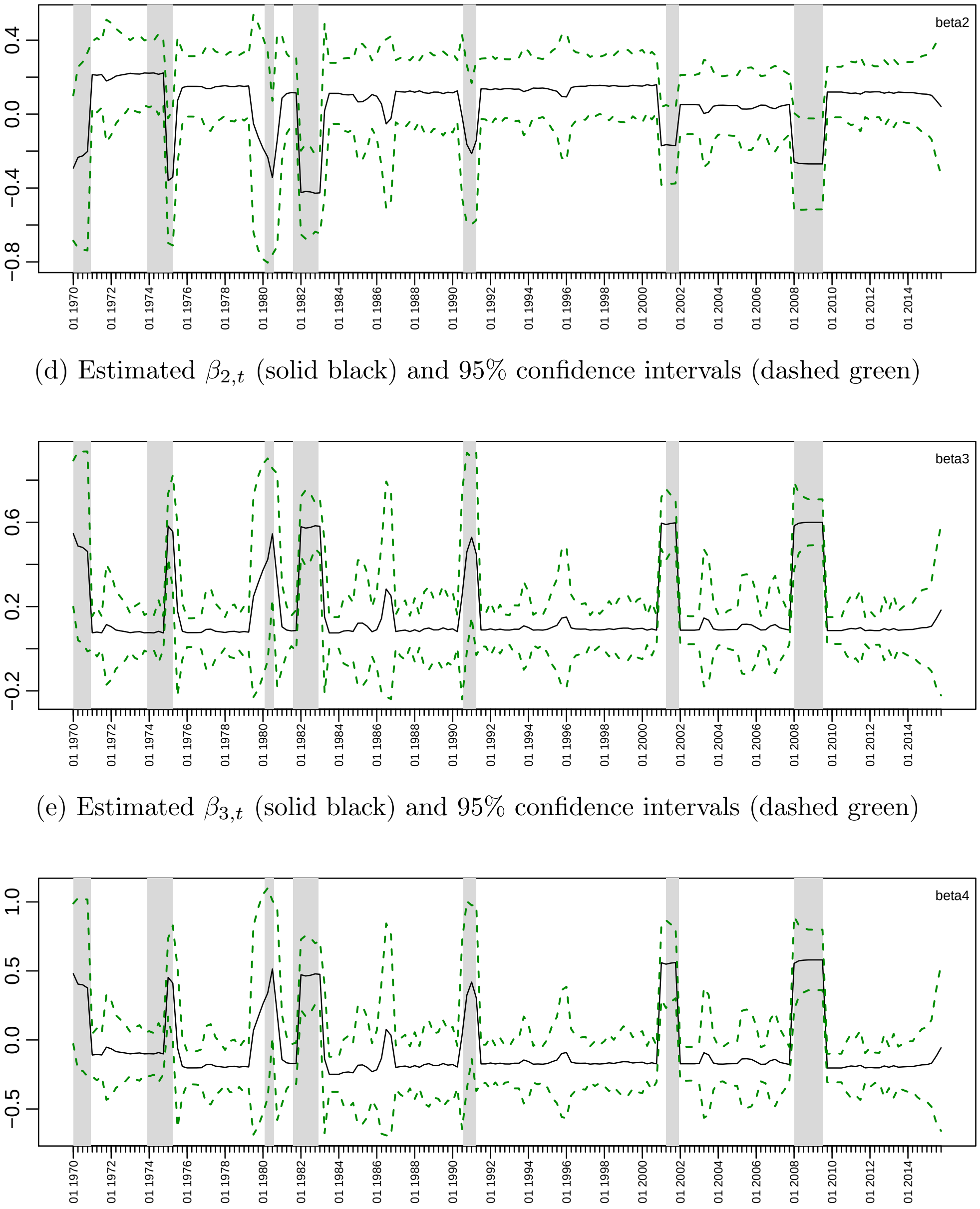


(a) Estimated $P(S_t = 1)$ (solid black)

(b) Estimated $\beta_{0,t}$ (solid black) and 95% confidence intervals (dashed green)

(c) Estimated $\beta_{1,t}$ (solid black) and 95% confidence intervals (dashed green)

Figure 4.6: Continued

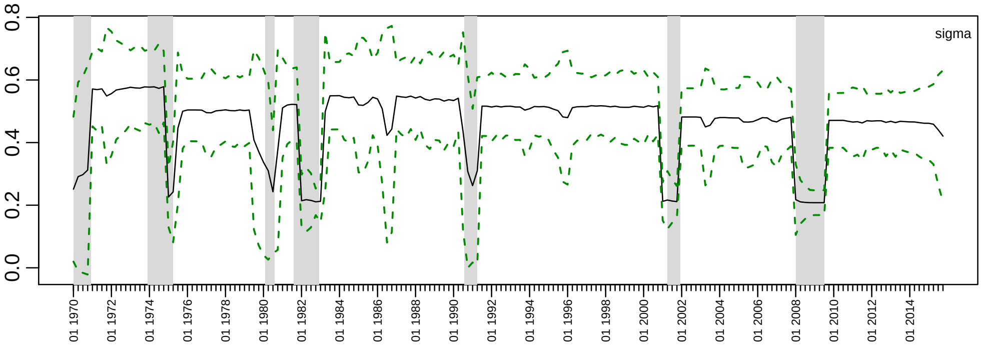


(d) Estimated $\beta_{2,t}$ (solid black) and 95% confidence intervals (dashed green)

(e) Estimated $\beta_{3,t}$ (solid black) and 95% confidence intervals (dashed green)

(f) Estimated $\beta_{4,t}$ (solid black) and 95% confidence intervals (dashed green)

Figure 4.6: Continued



(g) Estimated σ_t (solid black) and 95% confidence intervals (dashed green)

NBER record in Table 4.1. The fourth estimated recession period begins two quarter later and end one quarter later. The rest of the estimated recession periods are very similar to NBER statistics. The discrepancy is inevitable, since our estimation is based on single economic series, whereas NBER statistics are the results of collective consideration. Estimation of regression parameters and their confidence intervals are shown in Figure 4.6 (b) – (g). Clearly the estimates in different regimes are quite different, and they are not piece-wise constants even within the same regimes. The analysis of parameter estimation would be very similar to those in Section 4.1 and 4.2 and thus the discussion is omitted here. The proposed model has wide application in detecting regime changes in economic series analysis.

4.4 Comparison with Classical Markov Switching Models

In this section, we compare Hamilton's classical regime switching model with our model and show that our model is more sensitive to regime changes and work better for more complicated models. To achieve the desirable comparison, we re-analyze data series in Section 4.1 - 4.3 in classical MS models with the same number of regimes, i.e. $K = 2$ and with the same order in the autoregressive equation. For unemployment rate series we choose

$$y_t = \beta_{0,k} + \beta_{1,k} y_{t-1} + \epsilon_t.$$

For industrial production series and real manufacturing and trade inventory series we choose the following regression equation:

$$y_t = \beta_{0,k} + \beta_{1,k} y_{t-1} + \beta_{2,k} y_{t-2} + \beta_{3,k} y_{t-3} + \beta_{4,k} y_{t-4} + \epsilon_t.$$

In both equations, $\epsilon_t \sim N(0, \sigma_k^2)$ and k is the k^{th} regime. A MatLab package written by Perlin (n.d.-a) is used to conduct the statistical analysis to the transformed series in Figure 4.1b, 4.3b and 4.5b. This package is downloadable from the url provided in the reference. A typical maximum likelihood estimation (MLE) method is used to estimate the model parameters. No hyperparameters or EM algorithm are needed in this model.

Figure 4.7 shows the estimated smoothing probabilities in regime 1, i.e.

estimated $P(S_t = 1)$ for each time t for series from Figure 4.1b, 4.3b and 4.5b. The classical regime switching model also uses Bayesian classifier to classify the regime status according to the estimated smoothing probabilities. If the estimated $P(S_t = 1) > 0.5$ at time t, the series is at regime 1; otherwise the series is at regime 2. To determine which regime represents recession or expansion, a comparison is made between the beginning few regime estimates with the beginning white or shaded area in each plot in Figure 4.7. The shaded areas are again the recession periods reported by NBER. If the series begins with regime 1 in a white area, regime 1 is chosen for expansion periods; if the regime begins with regime 1 in a shaded area, regime 1 represents recession periods. Henceforth, regime 1 represents recession in unemployment rate series, and regime 1 is chosen to represent expansion periods for industrial production series and inventory series. It is obvious to understand such choices from Figure 4.7 (a)- (c).

The estimated equation for unemployment rate series (difference) in recession periods is

$$\begin{aligned}\hat{y}_t = -&0.02575 - 0.0238 y_{t-1}, \quad \text{with} \quad \hat{\sigma}_{\text{recession}} = 0.04823 \\ &(0.0386) \quad (0.0088)\end{aligned}$$

And the estimated equation in expansion periods is

$$\hat{y}_t = 0.6274 - 0.05389 y_{t-1}, \quad \text{with} \quad \hat{\sigma}_{\text{expansion}} = 0.01402$$

$$(0.029) \quad (0.0778)$$

The values in parenthesis are the standard errors of corresponding estimators such as $\hat{\beta}_0$ and $\hat{\beta}_1$ in each regime. β_1 is not significant different from zero for expansion periods at 5% level. According to Figure 4.7a, classical MS model also successfully detects recession periods in unemployment rate data. The estimated recession periods are presented in Table 4.4. Recession periods 1, 5 and 6 are estimated to be closer to NBER report in Table 4.1 in classical MS model. Estimation of period 7 is much wider than our model estimation in Table 4.2 and NBER record too. Estimation of other recession periods is very similar to our model estimation. The transition probability from recession to expansion is estimated to be 0.08 (higher than our model estimate 0.05) and from expansion to recession to be 0.02 (the same as our model estimate).

Table 4.4: The estimated recession periods in Classical MS model for unemployment rate series.

Recession periods	Starting date	Ending date
1	1970-03	1970-12
2	1974-04	1975-07
3	1980-01	1980-07
4	1981-07	1984-06
5	1990-07	1991-03
6	2001-06	2002-01
7	2008-03	2011-01

Classical MS model behaves a little different for industrial production series. The number of recession periods is much less than our model estimation in Section 4.2. Comparing the estimated probabilities with 0.5 in Figure 4.7b, the estimated recession periods are 1974:Q4-1975:Q1, 1982:Q2, 2001:Q1, 2001:Q3 and 2008:Q1-2009:Q1. The estimated probability in recession is 0.57 in 2001:Q1 and 0.52 in 2001:Q3, not a strong indication that the regime is in recession. These probabilities can be also interpreted as the fuzziness in the expansion regime. Thus it is reasonable to draw a conclusion that there are only 3 non-controversial estimated recession periods: 1974:Q4-1975:Q1, 1982:Q2 and 2008:Q1-2009:Q1.

The estimation of regression coefficients relies on the identification of the regime. For expansion (regime 1) periods 1973:Q2-1974:Q3, 1975:Q2-1982:Q1, 1982:Q3 - 2007:Q4 and 2009:Q2 - 2015:Q4, the fitted equation is

$$\begin{array}{cccccc} \hat{y}_t = & 0.4010 & + 0.5362 y_{t-1} & - 0.1927 y_{t-2} & + 0.1069 y_{t-3} & - 0.1202 y_{t-4} \\ & (0.0679) & (0.077) & (0.0970) & (0.0813) & (0.0612) \end{array}$$

where $\hat{\sigma}_1 = 0.4087$. For recession (regime 2) periods mentioned above, the fitted equation is

$$\begin{array}{cccccc} \hat{y}_t = & 0.49 & + 1.3297 y_{t-1} & + 0.2708 y_{t-2} & - 0.7682 y_{t-3} & - 0.7232 y_{t-4} \\ & (0.1577) & (0.4109) & (0.8502) & (0.8512) & (0.3612) \end{array}$$

where $\hat{\sigma}_2 = 1.9449$. The standard errors are given in parenthesis. The probability of switching from expansion to recession is estimated to be 0.04

and from recession to expansion is 0.37. It seems that the model is over parametrized, because the only significant coefficients are the ones for y_{t-1} in both regimes by constructing 95% confidence interval via standard errors. When a simpler model MS(2)-AR(1) is run and tested, the regression coefficients behave better and transition probabilities become a little smaller. Nevertheless, there is no improvement in the estimation of regimes. The estimation of the smoothed probabilities of the regime in this simpler model looks very similar to Figure 4.7b thus not reported here.

The application of classical MS model on real manufacturing and trade inventory series shows that in expansion periods (regime 1) the estimated equation is

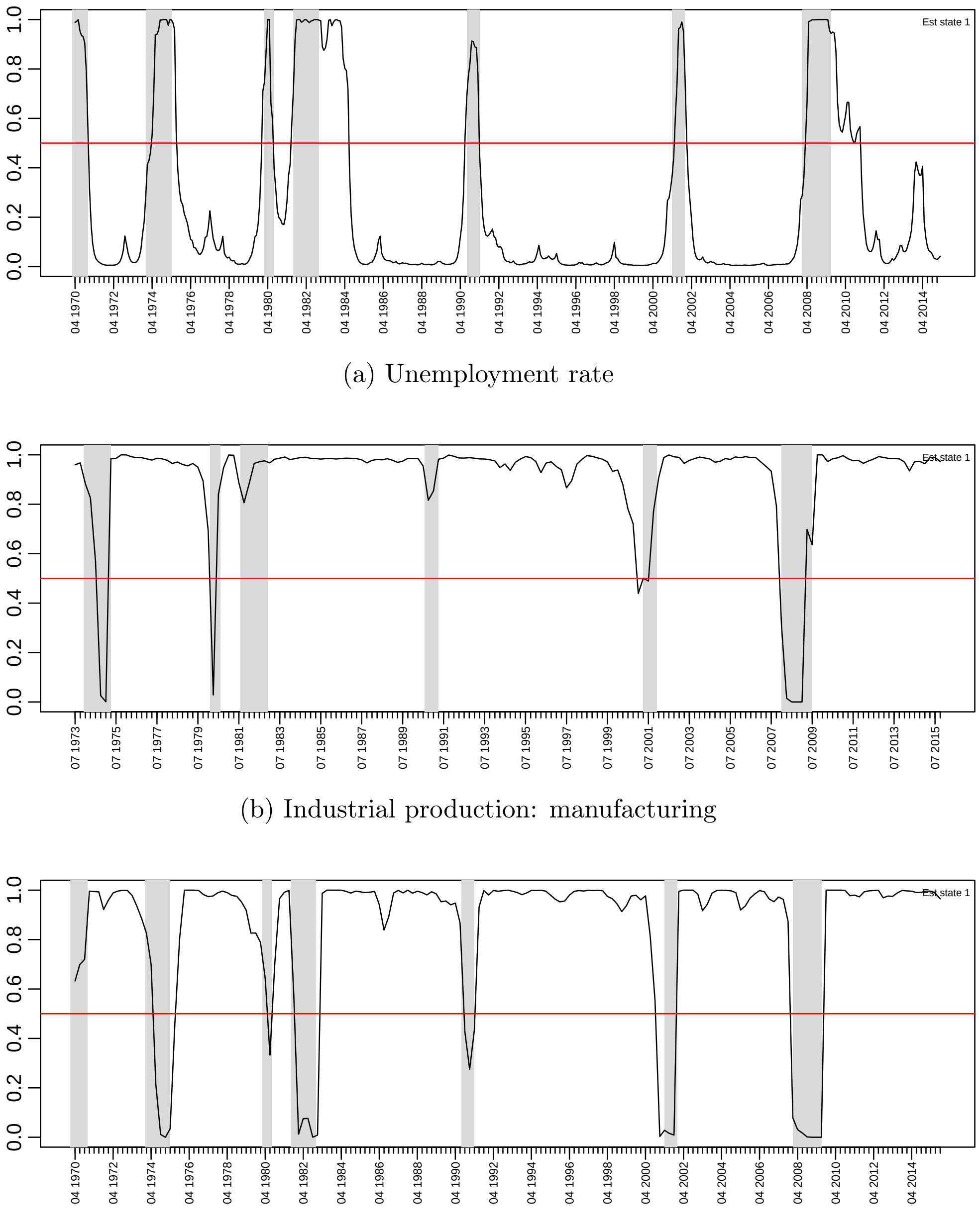
$$\begin{array}{llllll} \hat{y}_t = & 0.6493 & - 0.294y_{t-1} & + 0.1148y_{t-2} & + 0.0918y_{t-3} & - 0.2188y_{t-4} \\ & (0.0773) & (0.0788) & (0.0756) & (0.0745) & (0.0616) \end{array}$$

with estimated standard deviation $\hat{\sigma}_1 = 0.2812$ (0.038) and in recession periods, the estimation is

$$\begin{array}{llllll} \hat{y}_t = & -1.2544 & - 0.1634y_{t-1} & + 0.044y_{t-2} & + 0.408y_{t-3} & + 0.6775y_{t-4} \\ & (0.3411) & (0.2117) & (0.2271) & (0.2155) & (0.2629) \end{array}$$

with estimated standard deviation $\hat{\sigma}_2 = 0.5876$ (0.1880). The coefficients for y_{t-4} are significant for both regimes confirming the validity of auto regression

Figure 4.7: Estimation of smoothed probabilities of the series being in state 1 from the classical MS model for three real data series in Figure 4.1b, 4.3b and 4.5b



(a) Unemployment rate

(b) Industrial production: manufacturing

(c) Real manufacturing and trade inventory of order 4. The transition probability from expansion to recession is estimated to be 0.05 and from recession to expansion is 0.25. However this model fails to detect the first recession period in the first shaded area shown in Figure 4.7c.

In summary, comparing Figure 4.2a, 4.4a and 4.6a with Figure 4.7 (a) – (c), our model seems to detect more recession periods and the estimation of the recession periods are closer to NBER dating scheme than the classical MS model. Besides, in a two-regime classical MS model, there are only two sets of estimated parameters for each regime. However, the estimation of model parameters in our model is not restricted by the number of regimes. Our model gives more flexibility to estimate the parameters even in the same regime. First the values of model parameters are estimated at each time point and thus not a constant within a regime. Second, no manual work of identifying the regime is needed to write the fitted equation. The transition probabilities between the regimes in our model tend to be estimated no more than classical MS estimates in most of the cases. As also shown in the simulation study, the classical MS model has a tendency to overestimate the transition probabilities between the regimes. So our model estimation of transition probabilities is more accurate. Finally our model seems to work better in more complex modeling environment.

Chapter 5

Conclusions

In this dissertation we proposed a new stochastic regime switching model (2.1) with innovative assumptions (A1) and (A2) where the prior distribution of regression parameters depends on a finite state Markov chain S_t as well as the chain status at previous time, i.e. S_{t-1}. Prior distribution of regression parameters is no longer piece-wise constant within the same regime; but rather is regenerated at every switch toward a particular regime. Thus the posterior means and variances are time-varying throughout the process, accompanied by the estimation of regime status. In particular this system can automatically detect unknown number of switches, switching location and estimate the switching magnitude for every regression parameter. Under Bayesian inference framework, the posterior means and standard errors all have closed form solutions.

One major element in statistical inference of this model is the definition and derivation of the recursive weights. Forward, backward and smoothing recursive weights are derived on demand for three popular estimation methods, i.e. forward filtering, backward filtering and smoothing methods. Posterior means and variances of the parameters are thus estimated in the aforementioned context. Hyperparameters in this model are estimated by Expectation-Maximization (EM) procedure. To increase the computational efficiency, Bounded Complexity Mixture (BCMIX) algorithm is derived and implemented in simulation studies. This approximation scheme is proven to have lower computational complexity, yet comparable to Bayes estimation in statistical efficiency. Simulation studies have shown that regime switching identification ratio is as high as 99% for low switching series regardless of series length and with more frequent switching series, the estimation errors are smaller for longer series. Our simulation studies also show that classical Markov switching models have a tendency to overestimate the transition probabilities than our model estimates.

Finally, our model has successfully applied to three important economic time series: unemployment rate, industrial production and manufacturing and trade inventory. In comparison with classical MS model, our model is more sensitive in detecting regime changes. The dating of recession periods by our model may serve as a benchmark for NBER historical records and can provide valuable information for economic policy making.

Appendix A

Theorem 1. *Derive* $f(y_t|J_t^{(k)} = t, \mathcal{F}_{t-1})$

Proof.

$$
\begin{aligned}
& f(y_t|J_t^{(S_t)} = t, S_t = k, \mathcal{F}_{t-1}) \\
& = \iint f(y_t, \boldsymbol{\beta}_t, \tau_t | J_t^{(k)} = t, \mathcal{F}_{t-1}) \mathrm{d}\boldsymbol{\beta}_t \mathrm{d}\tau_t \\
& = \iint f(y_t|\boldsymbol{\beta}_t, \tau_t, J_t^{(k)} = t, \mathcal{F}_{t-1}) f(\boldsymbol{\beta}_t|\tau_t, J_t^{(k)} = t, \mathcal{F}_{t-1}) \ \mathrm{d}\boldsymbol{\beta}_t f(\tau_t|J_t^{(k)} = t, \mathcal{F}_{t-1}) \ \mathrm{d}\tau_t \\
& = \iint \pi^{-\frac{1}{2}} \tau_t^{\frac{1}{2}} \exp\Big\{ -\tau_t(y_t - \boldsymbol{x}_t' \boldsymbol{\beta}_t)^2 \Big\} \pi^{-\frac{d}{2}} \tau_t^{\frac{d}{2}} \big|\boldsymbol{V}^{(k)}\big|^{-\frac{1}{2}} \\
& \quad \exp\Bigg\{ -\tau_t \Big(\boldsymbol{\beta}_t' \big(\boldsymbol{V}^{(k)}\big)^{-1} \boldsymbol{\beta}_t - 2\boldsymbol{\beta}_t' \big(\boldsymbol{V}^{(k)}\big)^{-1} \boldsymbol{z}^{(k)} + (\boldsymbol{z}^{(k)})' \big(\boldsymbol{V}^{(k)}\big)^{-1} \boldsymbol{z}^{(k)} \Big) \Bigg\} \ \mathrm{d}\boldsymbol{\beta}_t \cdot \\
& \quad f(\tau_t|J_t^{(k)} = t, \mathcal{F}_{t-1}) \ \mathrm{d}\tau_t \\
& = \iint \pi^{-\frac{1}{2}} \tau_t^{\frac{1}{2}} \pi^{-\frac{d}{2}} \tau_t^{\frac{d}{2}} \big|\boldsymbol{V}^{(k)}\big|^{-\frac{1}{2}} \exp\Big\{ -\tau_t \Big(\boldsymbol{\beta}_t' (\boldsymbol{x}_t \boldsymbol{x}_t') \boldsymbol{\beta}_t - 2\boldsymbol{\beta}_t' \boldsymbol{x}_t y_t + y_t^2 \Big) \Big\} \\
& \quad \cdot \exp\Bigg\{ -\tau_t \Big(\boldsymbol{\beta}_t' \big(\boldsymbol{V}^{(k)}\big)^{-1} \boldsymbol{\beta}_t - 2\boldsymbol{\beta}_t' \big(\boldsymbol{V}^{(k)}\big)^{-1} \boldsymbol{z}^{(k)} + (\boldsymbol{z}^{(k)})' \big(\boldsymbol{V}^{(k)}\big)^{-1} \boldsymbol{z}^{(k)} \Big) \Bigg\} \ \mathrm{d}\boldsymbol{\beta}_t \\
& \quad \cdot f(\tau_t|J_t^{(k)} = t, \mathcal{F}_{t-1}) \ \mathrm{d}\tau_t \\
& = \iint \pi^{-\frac{1}{2}} \tau_t^{\frac{1}{2}} \pi^{-\frac{d}{2}} \tau_t^{\frac{d}{2}} \big|\boldsymbol{V}^{(k)}\big|^{-\frac{1}{2}} \exp\Bigg\{ -\tau_t \Bigg(\boldsymbol{\beta}_t' \Big(\boldsymbol{x}_t \boldsymbol{x}_t' + \big(\boldsymbol{V}^{(k)}\big)^{-1} \Big) \boldsymbol{\beta}_t
\end{aligned}
$$

$$- 2\boldsymbol{\beta}_t'\Big(\boldsymbol{x}_t y_t + \big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}^{(k)}\Big) + y_t^2 + (\boldsymbol{z}^{(k)})'\big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}^{(k)}\Big)\Big\}\ \mathrm{d}\boldsymbol{\beta}_t f(\tau_t|J_t^{(k)} = t, \mathcal{F}_{t-1})\ \mathrm{d}\tau_t$$

$$= \iint \pi^{-\frac{1}{2}}\tau_t^{\frac{1}{2}}\pi^{-\frac{d}{2}}\tau_t^{\frac{d}{2}}\big|\boldsymbol{V}^{(k)}\big|^{-\frac{1}{2}}\exp\Big\{ - \tau_t\Big(\boldsymbol{\beta}_t'\big(\boldsymbol{V}_{t,t}^{(k)}\big)^{-1}\boldsymbol{\beta}_t - 2\boldsymbol{\beta}_t'\big(\boldsymbol{V}_{t,t}^{(k)}\big)^{-1}\boldsymbol{z}_{t,t}^{(k)} +$$

$$\big(\boldsymbol{z}_{t,t}^{(k)}\big)'\big(\boldsymbol{V}_{t,t}^{(k)}\big)^{-1}\boldsymbol{z}_{t,t}^{(k)} - \big(\boldsymbol{z}_{t,t}^{(k)}\big)'\big(\boldsymbol{V}_{t,t}^{(k)}\big)^{-1}\boldsymbol{z}_{t,t}^{(k)} + y_t^2 + (\boldsymbol{z}^{(k)})'\big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}^{(k)}\Big)\Big\}\ \mathrm{d}\boldsymbol{\beta}_t$$

$$\cdot f(\tau_t|J_t^{(k)} = t, \mathcal{F}_{t-1})\ \mathrm{d}\tau_t$$

$$= \int \pi^{-\frac{1}{2}}\tau_t^{\frac{1}{2}}\big|\boldsymbol{V}^{(k)}\big|^{-\frac{1}{2}}\big|\boldsymbol{V}_{t,t}^{(k)}\big|^{\frac{1}{2}}\exp\Big\{ - \tau_t\Big(- \big(\boldsymbol{z}_{t,t}^{(k)}\big)'\big(\boldsymbol{V}_{t,t}^{(k)}\big)^{-1}\boldsymbol{z}_{t,t}^{(k)} + y_t^2 + (\boldsymbol{z}^{(k)})'\big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}^{(k)}\Big)\Big\}$$

$$\frac{1}{\Gamma(g^{(k)})(\lambda^{(k)})^{g^{(k)}}}\tau_t^{g^{(k)}-1}\exp\Big\{ - \frac{\tau_t}{\lambda^{(k)}}\Big\}\ \mathrm{d}\tau_t$$

$$= \pi^{-\frac{1}{2}}\big|\boldsymbol{V}^{(k)}\big|^{-\frac{1}{2}}\big|\boldsymbol{V}_{t,t}^{(k)}\big|^{\frac{1}{2}}\frac{\Gamma(g_{t,t}^{(k)})(\lambda_{t,t}^{(k)})^{g_{t,t}^{(k)}}}{\Gamma(g^{(k)})(\lambda^{(k)})^{g^{(k)}}}$$

□

Theorem 2. *Derive* $f(y_t|J_{t-1}^{(k)} = i, \mathcal{F}_{t-1})$

Proof.

$$f(y_t|J_{t-1}^{(S_{t-1})} = i, S_{t-1} = k, \mathcal{F}_{t-1})$$

$$= \iint f(y_t, \boldsymbol{\beta}_t, \tau_t|J_{t-1}^{(k)} = i, \mathcal{F}_{t-1})\mathrm{d}\boldsymbol{\beta}_t\mathrm{d}\tau_t$$

$$= \iint f(y_t|\boldsymbol{\beta}_t, \tau_t, J_{t-1}^{(k)} = i, \mathcal{F}_{t-1})f(\boldsymbol{\beta}_t|\tau_t, J_{t-1}^{(k)} = i, \mathcal{F}_{t-1})\ \mathrm{d}\boldsymbol{\beta}_t f(\tau_t|J_{t-1}^{(k)} = i, \mathcal{F}_{t-1})\ \mathrm{d}\tau_t$$

$$= \iint \pi^{-\frac{1}{2}}\tau_t^{\frac{1}{2}}\ \exp\Big\{ - \tau_t(y_t - \boldsymbol{x}_t'\boldsymbol{\beta}_t)^2\Big\}\pi^{-\frac{d}{2}}\tau_t^{\frac{d}{2}}\big|\boldsymbol{V}_{i,t-1}^{(k)}\big|^{-\frac{1}{2}}$$

$$\exp\Big\{ - \tau_t\Big(\boldsymbol{\beta}_t'\big(\boldsymbol{V}_{i,t-1}^{(k)}\big)^{-1}\boldsymbol{\beta}_t - 2\boldsymbol{\beta}_t'\big(\boldsymbol{V}_{i,t-1}^{(k)}\big)^{-1}\boldsymbol{z}_{i,t-1}^{(k)} + (\boldsymbol{z}_{i,t-1}^{(k)})'\big(\boldsymbol{V}_{i,t-1}^{(k)}\big)^{-1}\boldsymbol{z}_{i,t-1}^{(k)}\Big)\Big\}$$

$$\mathrm{d}\boldsymbol{\beta}_t f(\tau_t|J_{t-1}^{(k)} = i, \mathcal{F}_{t-1})\ \mathrm{d}\tau_t$$

$$= \iint \pi^{-\frac{1}{2}}\tau_t^{\frac{1}{2}}\pi^{-\frac{d}{2}}\tau_t^{\frac{d}{2}}\big|\boldsymbol{V}_{i,t-1}^{(k)}\big|^{-\frac{1}{2}}\exp\Big\{ - \tau_t\Big(\boldsymbol{\beta}_t'(\boldsymbol{x}_t\boldsymbol{x}_t')\boldsymbol{\beta}_t - 2\boldsymbol{\beta}_t'\boldsymbol{x}_t y_t + y_t^2\Big)\Big\}$$

$$\exp\Big\{-\tau_t\Big(\boldsymbol{\beta}_t'\big(\boldsymbol{V}_{i,t-1}^{(k)}\big)^{-1}\boldsymbol{\beta}_t-2\boldsymbol{\beta}_t'\big(\boldsymbol{V}_{i,t-1}^{(k)}\big)^{-1}\boldsymbol{z}_{i,t-1}^{(k)}+(\boldsymbol{z}_{i,t-1}^{(k)})'\big(\boldsymbol{V}_{i,t-1}^{(k)}\big)^{-1}\boldsymbol{z}_{i,t-1}^{(k)}\Big)\Big\}$$
$$\mathrm{d}\boldsymbol{\beta}_t f(\tau_t|J_{t-1}^{(k)}=i,\mathcal{F}_{t-1})\ \mathrm{d}\tau_t$$

To further simplify, we need to prove that

$$\boldsymbol{V}_{i,t}^{(k)}=\Big(\boldsymbol{x}_t\boldsymbol{x}_t'+\big(\boldsymbol{V}_{i,t-1}^{(k)}\big)^{-1}\Big)^{-1}\ \text{ and }\ \boldsymbol{z}_{i,t}^{(k)}=\big(\boldsymbol{V}_{i,t}^{(k)}\big)^{-1}\Big(\boldsymbol{x}_ty_t+\big(\boldsymbol{V}_{i,t-1}^{(k)}\big)^{-1}\boldsymbol{z}_{i,t-1}^{(k)}\Big)$$

Proof:

$$\begin{aligned}\Big(\boldsymbol{x}_t\boldsymbol{x}_t'+\big(\boldsymbol{V}_{i,t-1}^{(k)}\big)^{-1}\Big)^{-1}&=\Big(\boldsymbol{x}_t\boldsymbol{x}_t'+\sum_{r=i}^{t-1}\boldsymbol{x}_r\boldsymbol{x}_r'+\big(\boldsymbol{V}^{(k)}\big)^{-1}\Big)^{-1}\\&=\Big(\sum_{r=i}^{t}\boldsymbol{x}_r\boldsymbol{x}_r'+\big(\boldsymbol{V}^{(k)}\big)^{-1}\Big)^{-1}=\boldsymbol{V}_{i,t}^{(k)}\end{aligned}$$

$$\begin{aligned}\big(\boldsymbol{V}_{i,t}^{(k)}\big)^{-1}\Big(\boldsymbol{x}_ty_t+\big(\boldsymbol{V}_{i,t-1}^{(k)}\big)^{-1}\boldsymbol{z}_{i,t-1}^{(k)}\Big)&=\big(\boldsymbol{V}_{i,t}^{(k)}\big)^{-1}\Big(\boldsymbol{x}_ty_t+\sum_{r=i}^{t-1}\boldsymbol{x}_ry_r+\big(\boldsymbol{V}^{(k)}\big)^{-1}\Big)\\&=\big(\boldsymbol{V}_{i,t}^{(k)}\big)^{-1}\Big(\sum_{r=i}^{t}\boldsymbol{x}_ry_r+\big(\boldsymbol{V}^{(k)}\big)^{-1}\Big)=\boldsymbol{z}_{i,t}^{(k)}\end{aligned}$$

Continue with derivation of $f(y_t|J_{t-1}^{(k)}=i,\mathcal{F}_{t-1})$

$$\begin{aligned}=\iint&\pi^{-\frac{1}{2}}\tau_t^{\frac{1}{2}}\pi^{-\frac{d}{2}}\tau_t^{\frac{d}{2}}\big|\boldsymbol{V}_{i,t-1}^{(k)}\big|^{-\frac{1}{2}}\exp\Big\{-\tau_t\Big(\boldsymbol{\beta}_t'\big(\boldsymbol{V}_{i,t}^{(k)}\big)^{-1}\boldsymbol{\beta}_t-2\boldsymbol{\beta}_t'\big(\boldsymbol{V}_{i,t}^{(k)}\big)^{-1}\boldsymbol{z}_{i,t}^{(k)}\\&+(\boldsymbol{z}_{i,t}^{(k)})'\big(\boldsymbol{V}_{i,t}^{(k)}\big)^{-1}\boldsymbol{z}_{i,t}^{(k)}\Big)\Big\}\cdot\exp\Big\{-\tau_t\Big(-(\boldsymbol{z}_{i,t}^{(k)})'\big(\boldsymbol{V}_{i,t}^{(k)}\big)^{-1}\boldsymbol{z}_{i,t}^{(k)}\\&+(\boldsymbol{z}_{i,t-1}^{(k)})'\big(\boldsymbol{V}_{i,t-1}^{(k)}\big)^{-1}\boldsymbol{z}_{i,t-1}^{(k)}+y_t^2\Big)\Big\}\ \mathrm{d}\boldsymbol{\beta}_t f(\tau_t|J_{t-1}^{(k)}=i,\mathcal{F}_{t-1})\ \mathrm{d}\tau_t\end{aligned}$$

$$\begin{aligned}&= \pi^{-\frac{1}{2}} \big|\boldsymbol{V}_{i,t-1}^{(k)}\big|^{-\frac{1}{2}} \big|\boldsymbol{V}_{i,t}^{(k)}\big|^{\frac{1}{2}} \int \tau_t^{\frac{1}{2}} \exp\Big\{ -\tau_t \Big(-(\boldsymbol{z}_{i,t}^{(k)})'\big(\boldsymbol{V}_{i,t}^{(k)}\big)^{-1}\boldsymbol{z}_{i,t}^{(k)} + (\boldsymbol{z}_{i,t-1}^{(k)})'\big(\boldsymbol{V}_{i,t-1}^{(k)}\big)^{-1}\boldsymbol{z}_{i,t-1}^{(k)} \\ &\quad + y_t^2\Big)\Big\} \frac{1}{\Gamma\big(g_{i,t-1}^{(k)}\big)\big(\lambda_{i,t-1}^{(k)}\big)^{g_{i,t-1}^{(k)}}} \tau_t^{g_{i,t-1}^{(k)}-1} \exp\Big\{ -\frac{\tau_t}{\lambda_{i,t-1}^{(k)}}\Big\} \, \mathrm{d}\tau_t\end{aligned}$$

Similarly, we also need to show that

$$g_{i,t}^{(k)} = g_{i,t-1}^{(k)} + \frac{1}{2}$$

and

$$\frac{1}{\lambda_{i,t}^{(k)}} = \frac{1}{\lambda_{i,t-1}^{(k)}} - (\boldsymbol{z}_{i,t}^{(k)})'\big(\boldsymbol{V}_{i,t}^{(k)}\big)^{-1}\boldsymbol{z}_{i,t}^{(k)} + (\boldsymbol{z}_{i,t-1}^{(k)})'\big(\boldsymbol{V}_{i,t-1}^{(k)}\big)^{-1}\boldsymbol{z}_{i,t-1}^{(k)} + y_t^2$$

Proof:

$$\begin{aligned} g_{i,t-1}^{(k)} + \frac{1}{2} &= \frac{t-1-i+1}{2} + g^{(k)} + \frac{1}{2} \\ &= \frac{t-i+1}{2} + g^{(k)} = g_{i,t}^{(k)} \end{aligned}$$

$$\begin{aligned} &\frac{1}{\lambda_{i,t-1}^{(k)}} - (\boldsymbol{z}_{i,t}^{(k)})'\big(\boldsymbol{V}_{i,t}^{(k)}\big)^{-1}\boldsymbol{z}_{i,t}^{(k)} + (\boldsymbol{z}_{i,t-1}^{(k)})'\big(\boldsymbol{V}_{i,t-1}^{(k)}\big)^{-1}\boldsymbol{z}_{i,t-1}^{(k)} + y_t^2 \\ &= \frac{1}{\lambda^{(k)}} - (\boldsymbol{z}_{i,t-1}^{(k)})'\big(\boldsymbol{V}_{i,t-1}^{(k)}\big)^{-1}\boldsymbol{z}_{i,t-1}^{(k)} + \sum_{r=i}^{t-1} y_r^2 + (\boldsymbol{z}^{(k)})'\big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}^{(k)} \\ &\quad - (\boldsymbol{z}_{i,t}^{(k)})'\big(\boldsymbol{V}_{i,t}^{(k)}\big)^{-1}\boldsymbol{z}_{i,t}^{(k)} + (\boldsymbol{z}_{i,t-1}^{(k)})'\big(\boldsymbol{V}_{i,t-1}^{(k)}\big)^{-1}\boldsymbol{z}_{i,t-1}^{(k)} + y_t^2 \\ &= \frac{1}{\lambda^{(k)}} - (\boldsymbol{z}_{i,t}^{(k)})'\big(\boldsymbol{V}_{i,t}^{(k)}\big)^{-1}\boldsymbol{z}_{i,t}^{(k)} + \sum_{r=i}^{t} y_r^2 + (\boldsymbol{z}^{(k)})'\big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}^{(k)} \end{aligned}$$

$$= \frac{1}{\lambda_{i,t}^{(k)}}$$

Continue with derivation of $f(y_t|J_{t-1}^{(k)} = i, \mathcal{F}_{t-1})$ again.

$$= \pi^{-\frac{1}{2}} \big|\boldsymbol{V}_{i,t-1}^{(k)}\big|^{-\frac{1}{2}} \big|\boldsymbol{V}_{i,t}^{(k)}\big|^{\frac{1}{2}} \frac{1}{\Gamma\big(g_{i,t-1}^{(k)}\big)\big(\lambda_{i,t-1}^{(k)}\big)^{g_{i,t-1}^{(k)}}} \int \tau_t^{g_{i,t}^{(k)}-1} \exp\Big\{-\frac{\tau_t}{\lambda_{i,t}^{(k)}}\Big\}\,\mathrm{d}\tau_t$$
$$= \pi^{-\frac{1}{2}} \big|\boldsymbol{V}_{i,t-1}^{(k)}\big|^{-\frac{1}{2}} \big|\boldsymbol{V}_{i,t}^{(k)}\big|^{\frac{1}{2}} \frac{\Gamma\big(g_{i,t}^{(k)}\big)\big(\lambda_{i,t}^{(k)}\big)^{g_{i,t}^{(k)}}}{\Gamma\big(g_{i,t-1}^{(k)}\big)\big(\lambda_{i,t-1}^{(k)}\big)^{g_{i,t-1}^{(k)}}}$$

□

Theorem 3. *Show the variance of posterior $\boldsymbol{\beta}_t$ and σ_t in forwarding filtering estimation.*

Proof. By steiner's rule

$$Var(\boldsymbol{\beta}_t|\mathcal{F}_t) = E\big[\boldsymbol{\beta}_t\boldsymbol{\beta}_t'|\mathcal{F}_t\big] - E\big[\boldsymbol{\beta}|\mathcal{F}_t\big]\Big(E\big[\boldsymbol{\beta}|\mathcal{F}_t\big]\Big)'$$

$$\begin{aligned}
E\big[\boldsymbol{\beta}_t\boldsymbol{\beta}_t'|\mathcal{F}_t\big] &= \int_{\boldsymbol{\beta}_t} \boldsymbol{\beta}_t\Big(\int_{\tau_t} f(\boldsymbol{\beta}_t, \tau_t|\mathcal{F}_t)d\tau_t\Big)\boldsymbol{\beta}_t' d\boldsymbol{\beta}_t \\
&= \int_{\boldsymbol{\beta}_t} \boldsymbol{\beta}_t\Big(\int_{\tau_t} \sum_{k=1}^{K}\sum_{i=1}^{t} f(\boldsymbol{\beta}_t|\tau_t, J_t^{(k)} = i, \mathcal{F}_t) f(\tau_t|J_t^{(k)}, \mathcal{F}_t) f(J_t^{(k)} = i|\mathcal{F}_t) d\tau_t\Big)\boldsymbol{\beta}_t' d\boldsymbol{\beta}_t \\
&= \sum_{k=1}^{K}\sum_{i=1}^{t} \int_{\tau_t}\Big(\int_{\boldsymbol{\beta}_t} \boldsymbol{\beta}_t f(\boldsymbol{\beta}_t|\tau_t, J_t^{(k)} = i, \mathcal{F}_t)\boldsymbol{\beta}_t' d\boldsymbol{\beta}_t\Big) f(\tau_t|J_t^{(k)} = i, \mathcal{F}_t)\xi_{it}^{(k)} d\tau_t
\end{aligned}$$

It turns out that

$$\int_{\boldsymbol{\beta}_t} \boldsymbol{\beta}_t f(\boldsymbol{\beta}_t|\tau_t, J_t^{(k)} = i, \mathcal{F}_t)\boldsymbol{\beta}_t' d\boldsymbol{\beta}_t = E\big[\boldsymbol{\beta}_t\boldsymbol{\beta}_t'|\tau_t, J_t^{(k)} = i, \mathcal{F}_t\big]$$
$$= Var(\boldsymbol{\beta}_t|\tau_t, J_t^{(k)}, \mathcal{F}_t) + E(\boldsymbol{\beta}_t|\tau_t, J_t^{(k)} = i, \mathcal{F}_t)\big[E(\boldsymbol{\beta}_t|\tau_t, J_t^{(k)} = i, \mathcal{F}_t)\big]'$$

By equation (2.11),

$$E\big[\boldsymbol{\beta}_t\boldsymbol{\beta}_t'|\tau_t, J_t^{(k)} = i, \mathcal{F}_t\big] = \frac{\boldsymbol{V}_{it}^{(k)}}{2\tau_t} + \boldsymbol{z}_{it}^{(k)}\boldsymbol{z}_{it}^{(k)\prime}$$

and also by (2.14),

$$E\big[\boldsymbol{\beta}_t\boldsymbol{\beta}_t'|\mathcal{F}_t\big] = \sum_{k=1}^{K}\sum_{i=1}^{t}\xi_{it}^{(k)}\bigg(\frac{\boldsymbol{V}_{it}^{(k)}}{2}\int\frac{1}{\Gamma\big(g_{it}^{(k)}\big)\lambda_{it}^{(k)}}\tau_t^{g_{it}^{(k)}-1-1}\exp\bigg\{\frac{\tau_t}{\lambda_{it}^{(k)}}\bigg\}d\tau_t + \boldsymbol{z}_{it}^{(k)}\boldsymbol{z}_{it}^{(k)\prime}\bigg)$$
$$= \sum_{k=1}^{K}\sum_{i=1}^{t}\Bigg(\frac{\boldsymbol{V}_{it}^{(k)}}{2\lambda_{it}^{(k)}\big(g_{it}^{(k)} - 1\big)} + \boldsymbol{z}_{it}^{(k)}(\boldsymbol{z}_{it}^{k})'\Bigg)\xi_{it}^{(k)}$$

By equation (2.26),

$$Var(\boldsymbol{\beta}_t|\mathcal{F}_t) = \sum_{k=1}^{K}\sum_{i=1}^{t}\Bigg(\frac{\boldsymbol{V}_{it}^{(k)}}{2\lambda_{it}^{(k)}\big(g_{it}^{(k)} - 1\big)} + \boldsymbol{z}_{it}^{(k)}(\boldsymbol{z}_{it}^{k})'\Bigg)\xi_{it}^{(k)}$$
$$- \sum_{k=1}^{K}\sum_{i=1}^{t}\boldsymbol{z}_{it}^{(k)}\xi_{it}^{(k)}\Bigg(\sum_{k=1}^{K}\sum_{i=1}^{t}\boldsymbol{z}_{it}^{(k)}\xi_{it}^{(k)}\Bigg)'$$

□

Proof.

$$Var(\sigma_t|\mathcal{F}_t) = E[\sigma_t^2|\mathcal{F}_t] - \big(E[\sigma_t|\mathcal{F}_t]\big)^2$$

where

$$E[\sigma_t^2|\mathcal{F}_t] = \int \sigma_t^2 f(\sigma_t^2|\mathcal{F}_t)d\sigma_t^2 = \int \sum_{k=1}^{K}\sum_{i=1}^{t} \sigma_t^2 f(\sigma_t^2|J_t^{(k)} = i, \mathcal{F}_t)\xi_{it}^{(k)} d\sigma_t^2$$
$$= \sum_{k=1}^{K}\sum_{i=1}^{t} E\big[\sigma_t^2|J_t^{(k)} = i, \mathcal{F}_t\big]\xi_{it}^{(k)} = \sum_{k=1}^{K}\sum_{i=1}^{t} \frac{\xi_{it}^{(k)}}{2\lambda_{it}^{(k)}(g_{it}^{(k)} - 1)}$$

and by equation (2.14) and Theorem 7 in Appendix A,

$$Var(\sigma_t|\mathcal{F}_t) = \sum_{k=1}^{K}\sum_{i=1}^{t} \frac{\xi_{it}^{(k)}}{2\lambda_{it}^{(k)}\big(g_{it}^{(k)} - 1\big)}$$
$$- \Big(\sum_{k=1}^{K}\sum_{i=1}^{t} \xi_{it}^{(k)}\big(2\lambda_{it}^{(k)}\big)^{-\frac{1}{2}} \frac{\Gamma\big(g_{it}^{(k)} - \frac{1}{2}\big)}{\Gamma\big(g_{it}^{(k)}\big)}\Big)^2$$

□

Theorem 4. *Show that* $\frac{f_{i,t}^{(k)} f_{t+1,j}^{(k)}}{f_{0,0}^{(k)}} \cdot \frac{h_{i,t}^{(k)} h_{t+1,j}^{(k)}}{h_{0,0}^{(k)}}$ *can be simplified to* $\frac{\phi_{0,0}^{(k)}\phi_{i,j}^{(k)}}{\phi_{i,t}^{(k)}\phi_{t+1,j}^{(k)}} f_{i,j}^{(k)} h_{i,j}^{(k)}$*.*

Proof. First

$$\frac{f_{i,t}^{(k)} f_{t+1,j}^{(k)}}{f_{0,0}^{(k)}}$$
$$= \pi^{-\frac{d}{2}}\tau_t^{\frac{d}{2}}\big|\boldsymbol{V}_{i,t}^{(k)}\big|^{-\frac{1}{2}}\exp\Big\{ - \tau_t\Big(\boldsymbol{\beta}_t'\big(\boldsymbol{V}_{i,t}^{(k)}\big)^{-1}\boldsymbol{\beta}_t - 2\boldsymbol{\beta}_t'\big(\boldsymbol{V}_{i,t}^{(k)}\big)^{-1}\boldsymbol{z}_{i,t}^{(k)} + \big(\boldsymbol{z}_{i,t}^{(k)}\big)'\big(\boldsymbol{V}_{i,t}^{(k)}\big)^{-1}\boldsymbol{z}_{i,t}^{(k)}\Big)\Big\}$$
$$\pi^{-\frac{d}{2}}\tau_t^{\frac{d}{2}}\big|\boldsymbol{V}_{t+1,j}^{(k)}\big|^{-\frac{1}{2}}\exp\Big\{ - \tau_t\Big(\boldsymbol{\beta}_t'\big(\boldsymbol{V}_{t+1,j}^{(k)}\big)^{-1}\boldsymbol{\beta}_t - 2\boldsymbol{\beta}_t'\big(\boldsymbol{V}_{t+1,j}^{(k)}\big)^{-1}\boldsymbol{z}_{t+1,j}^{(k)}$$
$$+ \big(\boldsymbol{z}_{t+1,j}^{(k)}\big)'\big(\boldsymbol{V}_{t+1,j}^{(k)}\big)^{-1}\boldsymbol{z}_{t+1,j}^{(k)}\Big)\Big\}\pi^{\frac{d}{2}}\tau_t^{-\frac{d}{2}}\big|\boldsymbol{V}^{(k)}\big|^{\frac{1}{2}}\exp\Big\{\tau_t\Big(\boldsymbol{\beta}_t'\big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{\beta}_t - 2\boldsymbol{\beta}_t'\big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}^{(k)}$$
$$+ \big(\boldsymbol{z}^{(k)}\big)'\big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}^{(k)}\Big)\Big\}$$
$$= \big|\boldsymbol{V}_{i,t}^{(k)}\big|^{-\frac{1}{2}}\big|\boldsymbol{V}_{t+1,j}^{(k)}\big|^{-\frac{1}{2}}\big|\boldsymbol{V}^{(k)}\big|^{\frac{1}{2}}\big|\boldsymbol{V}_{i,j}^{(k)}\big|^{\frac{1}{2}} f_{i,j}^{(k)}\exp\Big\{ - \tau_t\Big(- \big(\boldsymbol{z}_{i,j}^{(k)}\big)'\big(\boldsymbol{V}_{i,j}^{(k)}\big)^{-1}\boldsymbol{z}_{i,j}^{(k)} +$$

$$\big(\boldsymbol{z}_{i,t}^{(k)}\big)'\big(\boldsymbol{V}_{i,t}^{(k)}\big)^{-1}\boldsymbol{z}_{i,t}^{(k)} + \big(\boldsymbol{z}_{t+1,j}^{(k)}\big)'\big(\boldsymbol{V}_{t+1,j}^{(k)}\big)^{-1}\boldsymbol{z}_{t+1,j}^{(k)} - \big(\boldsymbol{z}^{(k)}\big)'\big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}^{(k)}\Big)\Big\}$$

where

$$\begin{aligned}
&\Big(\big(\boldsymbol{V}_{i,t}^{(k)}\big)^{-1} + \big(\boldsymbol{V}_{t+1,j}^{(k)}\big)^{-1} - \big(\boldsymbol{V}^{(k)}\big)^{-1}\Big)^{-1}\\
&= \Big(\sum_{r=i}^{t}\boldsymbol{x}_r\boldsymbol{x}_r' + \big(\boldsymbol{V}^{(k)}\big)^{-1} + \sum_{r=t+1}^{j}\boldsymbol{x}_r\boldsymbol{x}_r' + \big(\boldsymbol{V}^{(k)}\big)^{-1} - \big(\boldsymbol{V}^{(k)}\big)^{-1}\Big)^{-1}\\
&= \Big(\sum_{r=i}^{j}\boldsymbol{x}_r\boldsymbol{x}_r' + \big(\boldsymbol{V}^{(k)}\big)^{-1}\Big)^{-1} = \big(\boldsymbol{V}_{i,j}^{(k)}\big)^{-1}
\end{aligned}$$

and

$$\begin{aligned}
&\big(\boldsymbol{V}_{i,t}^{(k)}\big)^{-1}\boldsymbol{z}_{i,t}^{(k)} + \big(\boldsymbol{V}_{t+1,j}^{(k)}\big)^{-1}\boldsymbol{z}_{t+1,j}^{(k)} - \big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}^{(k)}\\
&= \sum_{r=i}^{t}\boldsymbol{x}_r y_r + \big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}^{(k)} + \sum_{r=t+1}^{j}\boldsymbol{x}_r y_r + \big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}^{(k)} - \big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}^{(k)}\\
&= \sum_{r=i}^{j}\boldsymbol{x}_r y_r + \big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}^{(k)}\\
&= \big(\boldsymbol{V}_{i,j}^{(k)}\big)^{-1}\boldsymbol{z}_{i,j}^{(k)}
\end{aligned}$$

Second

$$\begin{aligned}
\frac{h_{i,t}^{(k)}h_{t+1,j}^{(k)}}{h_{0,0}^{(k)}} = {}& \frac{\Gamma(g^{(k)})\big(\lambda^{(k)}\big)^{g^{(k)}}}{\Gamma(g_{i,t}^{(k)})(\lambda_{i,t}^{(k)})^{g_{i,t}^{(k)}}\Gamma(g_{t+1,j}^{(k)})(\lambda_{t+1,j}^{(k)})^{g_{t+1,j}^{(k)}}}\tau_t^{g_{i,t}^{(k)}-1}\\
&\exp\Big\{-\frac{\tau_t}{\lambda_{i,t}^{(k)}}\Big\}\tau_t^{g_{t+1,j}^{(k)}-1}\exp\Big\{-\frac{\tau_t}{\lambda_{t+1,j}^{(k)}}\Big\}\tau_t^{-g^{(k)}+1}\exp\Big\{\frac{\tau_t}{\lambda^{(k)}}\Big\}
\end{aligned}$$

where

$$\begin{aligned}
& g^{(k)}_{i,t} + g^{(k)}_{t+1,j} - g^{(k)} \\
&= \frac{t-i+1}{2} + g^{(k)} + \frac{j-t}{2} + g^{(k)} - g^{(k)} \\
&= \frac{j-i+1}{2} + g^{(k)} = g^{(k)}_{i,j}
\end{aligned}$$

and

$$\begin{aligned}
& \frac{1}{\lambda^{(k)}_{i,t}} + \frac{1}{\lambda^{(k)}_{t+1,j}} - \frac{1}{\lambda^{(k)}} - \big(\boldsymbol{z}^{(k)}_{i,j}\big)'\big(\boldsymbol{V}^{(k)}_{i,j}\big)^{-1}\boldsymbol{z}^{(k)}_{i,j} + \big(\boldsymbol{z}^{(k)}_{i,t}\big)'\big(\boldsymbol{V}^{(k)}_{i,t}\big)^{-1}\boldsymbol{z}^{(k)}_{i,t} \\
& + \big(\boldsymbol{z}^{(k)}_{t+1,j}\big)'\big(\boldsymbol{V}^{(k)}_{t+1,j}\big)^{-1}\boldsymbol{z}^{(k)}_{t+1,j} - \big(\boldsymbol{z}^{(k)}\big)'\big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}^{(k)} \\
={}& \frac{1}{\lambda^{(k)}} - \big(\boldsymbol{z}^{(k)}_{i,t}\big)'\big(\boldsymbol{V}^{(k)}_{i,t}\big)^{-1}\boldsymbol{z}^{(k)}_{i,t} + \sum_{r=i}^{t} y_r^2 + \big(\boldsymbol{z}^{(k)}\big)'\big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}^{(k)} + \frac{1}{\lambda^{(k)}} \\
& - \big(\boldsymbol{z}^{(k)}_{t+1,j}\big)'\big(\boldsymbol{V}^{(k)}_{t+1,j}\big)^{-1}\boldsymbol{z}^{(k)}_{t+1,j} + \sum_{r=t+1}^{j} y_r^2 + \big(\boldsymbol{z}^{(k)}\big)'\big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}^{(k)} - \frac{1}{\lambda^{(k)}} \\
& - \big(\boldsymbol{z}^{(k)}_{i,j}\big)'\big(\boldsymbol{V}^{(k)}_{i,j}\big)^{-1}\boldsymbol{z}^{(k)}_{i,j} + \big(\boldsymbol{z}^{(k)}_{i,t}\big)'\big(\boldsymbol{V}^{(k)}_{i,t}\big)^{-1}\boldsymbol{z}^{(k)}_{i,t} + \big(\boldsymbol{z}^{(k)}_{t+1,j}\big)'\big(\boldsymbol{V}^{(k)}_{t+1,j}\big)^{-1}\boldsymbol{z}^{(k)}_{t+1,j} \\
& - \big(\boldsymbol{z}^{(k)}\big)'\big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}^{(k)} \\
={}& \frac{1}{\lambda^{(k)}} - \big(\boldsymbol{z}^{(k)}_{i,j}\big)'\big(\boldsymbol{V}^{(k)}_{i,j}\big)^{-1}\boldsymbol{z}^{(k)}_{i,j} + \sum_{r=i}^{j} y_r^2 + \big(\boldsymbol{z}^{(k)}\big)'\big(\boldsymbol{V}^{(k)}\big)^{-1}\boldsymbol{z}^{(k)}
\end{aligned}$$

Therefore

$$\frac{f^{(k)}_{i,t} f^{(k)}_{t+1,j}}{f^{(k)}_{0,0}} \cdot \frac{h^{(k)}_{i,t} h^{(k)}_{t+1,j}}{h^{(k)}_{0,0}}$$

$$= \left|\boldsymbol{V}_{i,t}^{(k)}\right|^{-\frac{1}{2}} \left|\boldsymbol{V}_{t+1,j}^{(k)}\right|^{-\frac{1}{2}} \left|\boldsymbol{V}^{(k)}\right|^{\frac{1}{2}} \left|\boldsymbol{V}_{i,j}^{(k)}\right|^{\frac{1}{2}} \frac{\Gamma(g^{(k)})\left(\lambda^{(k)}\right)^{g^{(k)}}\Gamma(g_{i,j}^{(k)})\left(\lambda_{i,j}^{(k)}\right)^{g_{i,j}^{(k)}}}{\Gamma(g_{i,t}^{(k)})(\lambda_{i,t}^{(k)})^{g_{i,t}^{(k)}}\Gamma(g_{t+1,j}^{(k)})(\lambda_{t+1,j}^{(k)})^{g_{t+1,j}^{(k)}}} f_{i,j}^{(k)} h_{i,j}^{(k)}$$

$$= \frac{\phi_{0,0}^{(k)}\phi_{i,j}^{(k)}}{\phi_{i,t}^{(k)}\phi_{t+1,j}^{(k)}} f_{i,j}^{(k)} h_{i,j}^{(k)}$$

□

Theorem 5. *Derive $P(S_t = k, S_{t-1} = l|\mathcal{F}_T, \boldsymbol{\theta}_{old})$.*

Proof. Without loss of generality, $\boldsymbol{\theta}_{\text{old}}$ is omitted in the proof.

$$P(S_t = k, S_{t-1} = l|\mathcal{F}_T)$$

$$= \frac{P(S_t = k, S_{t-1} = l, \mathcal{F}_T)}{f(\mathcal{F}_T)}$$

$$= \frac{f(\mathcal{F}_t|S_t = k, S_{t-1} = l)f(\mathcal{F}_{t+1,T}|S_t = k, S_{t-1} = l)P(S_t = k, S_{t-1} = l)}{f(\mathcal{F}_T)}$$

$$= \frac{P(\mathcal{F}_t, S_t = k, S_{t-1} = l)}{P(S_t = k, S_{t-1} = l)} \frac{P(\mathcal{F}_{t+1,T}, S_t = k, S_{t-1} = l)}{\cancel{P(S_t = k, S_{t-1} = l)}} \frac{\cancel{P(S_t = k, S_{t-1} = l)}}{f(\mathcal{F}_T)}$$

$$= \frac{P(\mathcal{F}_t, S_t = k, S_{t-1} = l)P(\mathcal{F}_{t+1,T}, S_t = k, S_{t-1} = l)}{f(\mathcal{F}_T)P(S_t = k, S_{t-1} = l)}$$

$$= \frac{f(y_t|S_t = k, S_{t-1} = l, \mathcal{F}_{t-1})P(S_t = k, S_{t-1} = l, \mathcal{F}_{t-1})}{f(\mathcal{F}_T)P(S_t = k, S_{t-1} = l)}$$

$$\cdot P(S_{t-1} = l|S_t = k, \mathcal{F}_{t+1,T})P(S_t = k, \mathcal{F}_{t+1,T})$$

$$= \frac{f(y_t|S_t = k, S_{t-1} = l, \mathcal{F}_{t-1})P(S_t = k|S_{t-1} = l, \mathcal{F}_{t-1})P(S_{t-1} = l|\mathcal{F}_{t-1})f(\mathcal{F}_{t-1})}{f(\mathcal{F}_T)\cancel{P(S_{t-1} = l|S_t = k)}P(S_t = k)}$$

$$\cdot \cancel{P(S_{t-1} = l|S_t = k)}P(S_t = k|\mathcal{F}_{t+1,T})f(\mathcal{F}_{t+1,T})$$

$$= \frac{f(y_t|S_t = k, S_{t-1} = l, \mathcal{F}_{t-1})p_{lk}\xi_{t-1}^{(l)}P(S_t = k|\mathcal{F}_{t+1,T})}{\pi_k} \frac{f(\mathcal{F}_{t-1})f(\mathcal{F}_{t+1,T})}{f(\mathcal{F}_T)}$$

$$= \frac{f(y_t|S_t = k, S_{t-1} = l, \mathcal{F}_{t-1})p_{lk}\xi_{t-1}^{(l)}\sum_{i=1}^{K} P(S_t = k|S_{t+1} = i, \mathcal{F}_{t+1,T})P(S_{t+1} = i|\mathcal{F}_{t+1,T})}{\pi_k}$$

$$
\begin{aligned}
&\cdot \frac{f(\mathcal{F}_{t-1})f(\mathcal{F}_{t+1,T})}{f(\mathcal{F}_T)} \\
&= \frac{f(y_t|S_t = k, S_{t-1} = l, \mathcal{F}_{t-1})p_{lk}\xi_{t-1}^{(l)} \sum_{i=1}^K q_{ik}\eta_{t+1}^{(i)}}{\pi_k} \frac{f(\mathcal{F}_{t-1})f(\mathcal{F}_{t+1,T})}{f(\mathcal{F}_T)} \\
&\propto f(y_t|S_t = k, S_{t-1} = l, \mathcal{F}_{t-1}) \cdot \frac{p_{lk}\xi_{t-1}^{(l)} \sum_{i=1}^K q_{ik}\eta_{t+1}^{(i)}}{\pi_k}
\end{aligned}
$$

In the derivation above, we use the fact that S_{t-1} given S_t does not depend on $\mathcal{F}_{t+1,T}$ by property of reversible Markov Chain; Bayes' theorem; $P(\mathcal{F}_{t-1}|\boldsymbol{\theta}_{\text{old}})$, $f(\mathcal{F}_{t+1,T}|\boldsymbol{\theta}_{\text{old}})$ and $f(\mathcal{F}_T|\boldsymbol{\theta}_{\text{old}})$ are constants and that Given $S_t = k$, $\mathcal{F}_t$ and $\mathcal{F}_{t+1,T}$ are conditionally independent.

Next calculate $f(y_t|S_t = k, S_{t-1} = l, \mathcal{F}_{t-1})$

$$
\begin{aligned}
&P(y_t|S_t = k, S_{t-1} = l, \mathcal{F}_{t-1}) \\
&= \sum_{i=1}^{t} P(y_t, J_t^{(S_t)} = i|S_t = k, S_{t-1} = l, \mathcal{F}_{t-1}) \\
&= \sum_{i=1}^{t} P(y_t|J_t^{(S_t)} = i, S_t = k, S_{t-1} = l, \mathcal{F}_{t-1})P(J_t^{(S_t)} = i|S_t = k, S_{t-1} = l, \mathcal{F}_{t-1})
\end{aligned}
$$

When $k \neq l$, use the facts in (2.16) and (2.20)

$$
\begin{aligned}
P(y_t|S_t = k, S_{t-1} = l, \mathcal{F}_{t-1}) &= P(y_t|J_t^{(S_t)} = t, S_t = k, \mathcal{F}_{t-1}) \cdot 1 \\
&= \pi^{-\frac{1}{2}} \frac{\phi_{t,t}^{(k)}}{\phi_{0,0}^{(k)}}
\end{aligned}
$$

When $k = l$, by the facts in (2.18) and (2.21)

$$\begin{aligned}
&P(y_t|S_t = k, S_{t-1} = l, \mathcal{F}_{t-1}) \\
&= \sum_{i=1}^{t-1} P(y_t|J_{t-1}^{(S_{t-1})} = i, S_{t-1} = k, \mathcal{F}_{t-1}) P(J_{t-1}^{(S_{t-1})} = i|S_{t-1} = k, \mathcal{F}_{t-1}) \\
&= \sum_{i=1}^{t-1} P(y_t|J_{t-1}^{(S_{t-1})} = i, S_{t-1} = k, \mathcal{F}_{t-1}) \frac{P(J_{t-1}^{(S_{t-1})} = i, S_{t-1} = k|\mathcal{F}_{t-1})}{P(S_{t-1} = k|\mathcal{F}_{t-1})} \\
&= \sum_{i=1}^{t-1} \pi^{-\frac{1}{2}} \frac{\phi_{i,t}^{(k)}}{\phi_{i,t-1}^{(k)}} \cdot \frac{\xi_{i,t-1}^{(k)}}{\xi_{t-1}^{(k)}}
\end{aligned}$$

In summary

$$P(S_t = k, S_{t-1} = l|\mathcal{F}_T, \boldsymbol{\theta}_{\text{old}}) = \begin{cases} \frac{\phi_{t,t}^{(k)} \xi_{t-1}^{(l)}}{\phi_{0,0}^{(k)}} \frac{p_{lk}}{\pi_k} \sum_{r=1}^{K} q_{rk} \eta_{t+1}^{(r)} \Big/ A_t & k \neq l \\ \sum_{i=1}^{t-1} \frac{\phi_{i,t}^{(k)} \xi_{i,t-1}^{(k)}}{\phi_{i,t-1}^{(k)}} \frac{p_{kk}}{\pi_k} \sum_{r=1}^{K} q_{rk} \eta_{t+1}^{(r)} \Big/ A_t & k = l \end{cases}$$

where $A_t = \sum_{l=1}^{K} \left(\sum_{k \neq l}^{K} \frac{\phi_{t,t}^{(k)} \xi_{t-1}^{(l)}}{\phi_{0,0}^{(k)}} \frac{p_{lk}}{\pi_k} \sum_{r=1}^{K} q_{rk} \eta_{t+1}^{(r)} + \sum_{i=1}^{t-1} \frac{\phi_{i,t}^{(k)} \xi_{i,t-1}^{(k)}}{\phi_{i,t-1}^{(k)}} \frac{p_{kk}}{\pi_k} \sum_{r=1}^{K} q_{rk} \eta_{t+1}^{(r)} \right)$

□

Theorem 6. *Assume that*

(1) The Markov Chain $\{S_t \in \{1, 2, \ldots, K\}|t \geq 0\}$ is irreducible, and follows the transition probability matrix $A = (a_{ij})_{1 \leq i,j \leq K}$, i.e. $a_{ij} = P(S_t = j|S_{t-1} = i)$;

(2) Given S_t which defines various regimes, $(\beta_t,\ \nu_t) = (\beta^{(S_t)},\ \nu^{(S_t)})$;

(3) The Markov chain $\{S_t,\ t \geq 0\}$ has a stationary distribution $\pi =$

$(\pi_1, \ldots, \pi_K)^T$. With this assumption, a reversible Markov chain for $\{S_t\}$ can be defined as a chain with transition probability matrix $\widetilde{A} = (\tilde{a}_{ij})_{1\leq i,j,\leq K}$, where $\tilde{a}_{ij} = P(S_t = j|S_{t+1} = i) = a_{ji}\pi_j/\pi_i$

Denote that $p_t^{(k)} = P(S_t = k|\mathcal{F}_{1t})$ $(1 \leq k \leq K)$, $\widetilde{p}_t^{(k)} = P(S_t = k|\mathcal{F}_{t,T})$, $\phi_{t,k} = P(y_t|S_t = k)$, $p_t = (p_t^{(1)}, \ldots, p_t^{(K)})'$, $\phi_t = (\phi_1(y_t), \ldots, \phi_K(y_t))'$ and A_k is the k^{th} column of the transition matrix A and similarly $\widetilde{A}_k$ is the k^{th} column of the transition matrix $\widetilde{A}$.The smoothing estimate of $\{S_t\}$ given $\mathcal{F}_n$ can be written as

$$P(S_t = k|\mathcal{F}_n) = \frac{p_t^{(k)}\widetilde{A}'_k\tilde{p}_{t+1}/\pi_k}{\sum_{i=1}^K p_t^{(i)}\widetilde{A}'_i\tilde{p}_{t+1}/\pi_i} \tag{A.1}$$

and the smoothing estimate of transition probability is

$$P(S_t = j|S_{t-1} = i, \mathcal{F}_n) = \frac{\phi_{t,j}a_{ij}\widetilde{A}'_j\tilde{p}_{t+1}}{\pi_j}/C_1 \tag{A.2}$$

$$P(S_t = j, S_{t-1} = i|\mathcal{F}_n) = \frac{\phi_{t,j}a_{ij}p_{t-1}^{(i)}\widetilde{A}'_j\tilde{p}_{t+1}}{\pi_j}/C_2 \tag{A.3}$$

where C_1 and C_2 are the normalizing constants.

$$C_1 = \frac{P(S_{t-1} = i|\mathcal{F}_n)f(\mathcal{F}_n)}{p_{t-1}^{(i)}f(\mathcal{F}_{t-1})f(\mathcal{F}_{t+1,n})} \quad or \quad C_1 = \sum_{k=1}^K \frac{\phi_{t,k}a_{ik}\widetilde{A}'_k\tilde{p}_{t+1}}{\pi_k}$$

and

$$C_2 = \frac{f(\mathcal{F}_n)}{f(\mathcal{F}_{t-1})f(\mathcal{F}_{t+1,n})} \quad or \quad C_2 = \sum_{k=1}^K \sum_{l=1}^K \frac{\phi_{t,k}a_{lk}p_{t-1}^{(l)}\widetilde{A}'_k\tilde{p}_{t+1}}{\pi_k}$$

Proof.

$$
\begin{aligned}
&P(S_t = j|S_{t-1} = i, \mathcal{F}_n)\\
&= \frac{P(S_t = j, S_{t-1} = i, \mathcal{F}_t, \mathcal{F}_{t+1,n})}{f(\mathcal{F}_n)P(S_{t-1} = i|\mathcal{F}_n)}\\
&= \frac{f(\mathcal{F}_t|S_t = j, S_{t-1} = i)f(\mathcal{F}_{t+1,n}|S_t = j, S_{t-1} = i)P(S_t = j, S_{t-1} = i)}{f(\mathcal{F}_n)P(S_{t-1} = i|\mathcal{F}_n)}\\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{(by indepdence of } \mathcal{F}_t \text{ and } \mathcal{F}_{t+1,n} \text{ given } S_t = j)\\
&= \frac{P(\mathcal{F}_t, S_t = j, S_{t-1} = i)P(\mathcal{F}_{t+1,n}, S_t = j, S_{t-1} = i)P(S_t = j, S_{t-1} = i)}{P(\mathcal{F}_n)P(S_{t-1} = i|\mathcal{F}_n)P(S_t = j, S_{t-1} = i)P(S_t = j, S_{t-1} = i)}\\
&= \frac{P(\mathcal{F}_t, S_t = j, S_{t-1} = i)P(\mathcal{F}_{t+1,n}, S_t = j, S_{t-1} = i)}{P(\mathcal{F}_n)P(S_{t-1} = i|\mathcal{F}_n)P(S_t = j, S_{t-1} = i)}\\
&= \frac{f(y_t|S_t = j, S_{t-1} = i, \mathcal{F}_{t-1})P(S_t = j, S_{t-1} = i, \mathcal{F}_{t-1})}{f(\mathcal{F}_n)P(S_{t-1} = i|\mathcal{F}_n)}\\
&\quad \cdot \frac{P(S_{t-1} = i|S_t = j, \mathcal{F}_{t+1,n})P(S_t = j, \mathcal{F}_{t+1,n})}{P(S_{t-1} = i|S_t = j)P(S_t = j)}\\
&= \frac{f(y_t|S_t = j)P(S_t = j|S_{t-1} = i, \mathcal{F}_{t-1})P(S_{t-1} = i|\mathcal{F}_{t-1})f(\mathcal{F}_{t-1})}{f(\mathcal{F}_n)P(S_{t-1} = i|\mathcal{F}_n)}\\
&\quad \frac{P(S_{t-1} = i|S_t = j)P(S_t = j, \mathcal{F}_{t+1,n})}{P(S_{t-1} = i|S_t = j)P(S_t = j)}\\
&= \frac{f(y_t|S_t = j)P(S_t = j|S_{t-1} = i)P(S_{t-1} = i|\mathcal{F}_{t-1})f(\mathcal{F}_{t-1})}{f(\mathcal{F}_n)P(S_{t-1} = i|\mathcal{F}_n)}\\
&\quad \frac{P(S_{t-1} = i|S_t = j)P(S_t = j|\mathcal{F}_{t+1,n})f(\mathcal{F}_{t+1,n})}{P(S_{t-1} = i|S_t = j)P(S_t = j)}\\
&= \frac{\phi_{t,j}a_{ij}p^{(i)}_{t-1}P(S_t = j|\mathcal{F}_{t+1,n})}{f(S_{t-1} = i|\mathcal{F}_n)\pi_j}\frac{f(\mathcal{F}_{t-1})f(\mathcal{F}_{t+1,n})}{f(\mathcal{F}_n)}\\
&= \frac{\phi_{t,j}a_{ij}p^{(i)}_{t-1}\sum_{l=1}^{K} P(S_t = j, S_{t+1} = l|\mathcal{F}_{t+1,n})}{P(S_{t-1} = i|\mathcal{F}_n)\pi_j}\frac{f(\mathcal{F}_{t-1})f(\mathcal{F}_{t+1,n})}{f(\mathcal{F}_n)}\\
&= \frac{\phi_{t,j}a_{ij}p^{(i)}_{t-1}\sum_{l=1}^{K} P(S_t = j|S_{t+1} = l, \mathcal{F}_{t+1,n})P(S_{t+1} = l|\mathcal{F}_{t+1,n})}{f(S_{t-1} = i|\mathcal{F}_n)\pi_j}\frac{f(\mathcal{F}_{t-1})f(\mathcal{F}_{t+1,n})}{f(\mathcal{F}_n)}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{\phi_{t,j} a_{ij} p_{t-1}^{(i)} \sum_{l=1}^{K} \tilde{a}_{lj} \tilde{p}_{t+1}^{l}}{P(S_{t-1} = i|\mathcal{F}_n)\pi_j} \frac{f(\mathcal{F}_{t-1}) f(\mathcal{F}_{t+1,n})}{f(\mathcal{F}_n)} \\
&= \frac{\phi_{t,j} a_{ij} \widetilde{A}'_j \tilde{p}_{t+1}}{\pi_j} \frac{p_{t-1}^{(i)} f(\mathcal{F}_{t-1}) f(\mathcal{F}_{t+1,n})}{P(S_{t-1} = i|\mathcal{F}_n) f(\mathcal{F}_n)}
\end{aligned}
$$

Since $\frac{f(\mathcal{F}_{t-1})f(\mathcal{F}_{t+1,n})}{f(\mathcal{F}_n)}$ is a constant and $p_{t-1}^{(i)},\ P(S_{t-1} = i|\mathcal{F}_n)$ does not depend on j, thus also a constant. Then

$$
P(S_t = j|S_{t-1} = i, \mathcal{F}_n) \propto \frac{\phi_{t,j} a_{ij} \widetilde{A}'_j \tilde{p}_{t+1}}{\pi_j}
$$

and

$$
\begin{aligned}
P(S_t = j, S_{t-1} = i|\mathcal{F}_n) &= P(S_t = j|S_{t-1} = i\mathcal{F}_n) P(S_{t-1} = i|\mathcal{F}_n) \\
&= \frac{\phi_{t,j} a_{ij} \widetilde{A}'_j \tilde{p}_{t+1}}{\pi_j} \frac{p_{t-1}^{(i)} f(\mathcal{F}_{t-1}) f(\mathcal{F}_{t+1,n})}{\cancel{P(S_{t-1} = i|\mathcal{F}_n)} f(\mathcal{F}_n)} \cancel{P(S_{t-1} = i|\mathcal{F}_n)} \\
&\propto \frac{\phi_{t,j} a_{ij} p_{t-1}^{(i)} \widetilde{A}'_j \tilde{p}_{t+1}}{\pi_j}
\end{aligned}
$$

This formula is correct because we can show the following relation. Let $c_0 = \frac{f(\mathcal{F}_{t-1})P(\mathcal{F}_{t+1,n})}{f(\mathcal{F}_n)}$

$$
\begin{aligned}
P(S_t = k|\mathcal{F}_n) &= \sum_{l=1}^{K} P(S_t = k, S_{t-1} = l|\mathcal{F}_n) \\
&= \sum_{l=1}^{K} P(S_t = k|S_{t-1} = l, \mathcal{F}_n) P(S_{t-1} = l|\mathcal{F}_n)
\end{aligned}
$$

$$= \sum_{l=1}^{K} \frac{\phi_{t,k} a_{lk} p_{t-1}^{(l)} \widetilde{A}'_k \tilde{p}_{t+1}}{\pi_k P(S_{t-1} = l|\mathcal{F}_n)} \cdot c_0 P(S_{t-1} = l|\mathcal{F}_n)$$

(from formular (2) with constant C_1)

$$= \sum_{l=1}^{K} \frac{\phi_{t,k} a_{lk} p_{t-1}^{(l)} \widetilde{A}'_k \tilde{p}_{t+1}}{\pi_k} \cdot c_0$$

$$= \frac{\phi_{t,k} \sum_{l=1}^{K} a_{lk} p_{t-1}^{(l)} \widetilde{A}'_k \tilde{p}_{t+1}}{\pi_k} \cdot c_0$$

$$= \frac{P(y_t|S_t = k) P(S_t = k|\mathcal{F}_{t-1}) \widetilde{A}'_k \tilde{p}_{t+1}}{\pi_k} \cdot c_0$$

$$= \frac{P(y_t|S_t = k, \mathcal{F}_{t-1}) P(S_t = k|\mathcal{F}_{t-1}) \widetilde{A}'_k \tilde{p}_{t+1}}{\pi_k} \cdot c_0$$

$$= \frac{P(y_t, S_t = k|\mathcal{F}_{t-1}) \widetilde{A}'_k \tilde{p}_{t+1}}{\pi_k} \cdot c_0$$

$$= \frac{P(S_t = k|\mathcal{F}_t) f(y_t|\mathcal{F}_{t-1}) \widetilde{A}'_k \tilde{p}_{t+1}}{\pi_k} \cdot c_0$$

$$= \frac{p_t^{(k)} \tilde{A}'_k \tilde{p}_{t+1}}{\pi_k} \cdot f(y_t|\mathcal{F}_{t-1}) c_0$$

$$\propto \frac{p_t^{(k)} \widetilde{A}'_k \tilde{p}_{t+1}}{\pi_k} \qquad \square$$

Theorem 7. *If Let* $X = \frac{1}{2Y^2}$ *and* $X \sim$ *Gamma* $(g,\ \lambda)$*, then*

$$E(Y^2) = \frac{1}{2\lambda(g-1)}, \qquad E(Y) = (2\lambda)^{-\frac{1}{2}} \frac{\Gamma(g - \frac{1}{2})}{\Gamma(g)}$$

and

$$var(Y) = \frac{1}{2\lambda(g-1)} - \frac{\Gamma^2(g - \frac{1}{2})}{2\lambda\Gamma^2(g)}$$

Proof. Let pdf of X be $f_X(x) = \frac{1}{\Gamma(g)\lambda^g} x^{g-1} \exp\left\{\frac{x}{\lambda}\right\}$, and

$$\frac{dx}{dy} = -\frac{1}{Y^3}$$

, then pdf of Y is

$$f_Y(y) = \frac{1}{\Gamma(g)\lambda^g} 2^{-g+1} Y^{-2g-1} \exp\left\{\frac{1}{2\lambda Y^2}\right\} \quad \forall y > 0$$

$$\begin{aligned} EY &= \int_0^\infty y f_Y(y) dy = \int_0^\infty \frac{1}{\Gamma(g)\lambda^g} 2^{-g+1} Y^{-2(g-\frac{1}{2})-1} \exp\left\{\frac{1}{2\lambda Y^2}\right\} dy \\ &= \frac{\Gamma(g-\frac{1}{2})\lambda^{g-\frac{1}{2}}}{\Gamma(g)\lambda^g} 2^{-g+1} 2^{g-\frac{1}{2}-1} = \frac{\Gamma(g-\frac{1}{2})}{\Gamma(g)} (2\lambda)^{-\frac{1}{2}} \end{aligned}$$

$$\begin{aligned} EY^2 &= \int_0^\infty y^2 f_Y(y) dy = \int_0^\infty \frac{1}{\Gamma(g)\lambda^g} 2^{-g+1} Y^{-2(g-1)-1} \exp\left\{\frac{1}{2\lambda Y^2}\right\} dy \\ &= \frac{\Gamma(g-1)\lambda^{g-1}}{\Gamma(g)\lambda^g} 2^{-g+1} 2^{g-2} = \frac{1}{2\lambda(g-1)} \end{aligned}$$

$$var(Y) = EY^2 - (EY)^2 = \frac{1}{2\lambda(g-1)} - \frac{\Gamma^2(g-\frac{1}{2})}{2\lambda\Gamma^2(g)}$$

□

Theorem 8. *Suppose that X follows Gamma distribution with parameters*

α and β with density function

$$f(x;\alpha,\beta)=\frac{1}{\Gamma(\alpha)\beta^{\alpha}}x^{\alpha-1}e^{-\frac{x}{\beta}}$$

Show that

$$E\big[logX\big]=\frac{d}{d\alpha}log\ \Gamma(\alpha)+\ log\beta$$

Proof. First consider the definition of Gamma function and its derivative, if Gamma function is defined as

$$\Gamma(\alpha)=\int_0^{\infty}y^{\alpha-1}\mathrm{e}^{-y}\mathrm{d}y$$

whose derivative is

$$\frac{\mathrm{d}\Gamma(\alpha)}{\mathrm{d}\alpha}=\int_0^{\infty}(\mathrm{log}y)y^{\alpha-1}\mathrm{e}^{-y}dy$$

Then

$$E[\mathrm{log}X]=\int_0^{\infty}(\ \mathrm{log}\ x)\frac{1}{\Gamma(\alpha)\beta^{\alpha}}x^{\alpha-1}\mathrm{e}^{-\frac{x}{\beta}}\mathrm{d}x$$

Let $y=\frac{x}{\beta}$ then $\mathrm{d}x=\beta\mathrm{d}y$

$$\begin{aligned}&E[\mathrm{log}X]\\&=\frac{1}{\Gamma(\alpha)\beta^{\alpha}}\int_0^{\infty}(\ \mathrm{log}\ y+\ \mathrm{log}\ \beta)y^{\alpha-1}\beta^{\alpha-1}\mathrm{e}^{-y}\beta\mathrm{d}y\\&=\frac{1}{\Gamma(\alpha)}\bigg(\int_0^{\infty}(\ \mathrm{log}\ y)y^{\alpha-1}e^{-y}dy+\int_0^{\infty}(\ \mathrm{log}\ \beta)y^{\alpha-1}\mathrm{e}^{-y}\mathrm{d}y\bigg)\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\Gamma(\alpha)} \frac{\mathrm{d}\Gamma(\alpha)}{\mathrm{d}\alpha} + \log\, \beta \\
&= \frac{\mathrm{d}}{\mathrm{d}\alpha} \log\, \Gamma(\alpha) + \ \log\, \beta \qquad \square
\end{aligned}
$$

Theorem 9. *Theorems about matrix expectations*

$$E\big[\boldsymbol{X}\boldsymbol{X}'\big] = \Sigma_{\boldsymbol{X}} + E(\boldsymbol{X})(E\boldsymbol{X})'$$

$$E(\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}) = tr\big(\boldsymbol{A}\Sigma_{\boldsymbol{X}}\big) + (E\boldsymbol{X})'\boldsymbol{A}(E\boldsymbol{X})$$

Theorem 10. *If A is a symmetric matrix and $\mathcal{A}_{ij}$ is the i, j^{th} cofactor of A, then*

$$\frac{\partial\ log\ (|A|)}{\partial A} = 2A^{-1} - \ \ diag\ (A^{-1})$$

If A is a symmetric matrix and B is an arbitrary matrix and AB meaningful, then

$$\frac{\partial tr(AB)}{\partial A} = B + B^T - diag(B)$$

extracted from Bilmes (1998) [1]*.*

Theorem 11. *Suppose under true model $y_t \sim\ N\,(\boldsymbol{x}_t'\boldsymbol{\beta}_t, \sigma_t^2)$ and under the estimated model, $y_t \sim\ N\,(\boldsymbol{x}_t'\widehat{\boldsymbol{\beta}}_t, \widehat{\sigma}_t^2)$. Let $\boldsymbol{\theta}_t = (\boldsymbol{\beta}_t,\ \sigma_t)$, Kullback-Leibler*

[1] A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models, *International Computer Science Institute*

(KL) divergence is

$$KL(\boldsymbol{\theta}_t, \widehat{\boldsymbol{\theta}}_t) = \frac{1}{2}\left(\frac{\big(\boldsymbol{x}_t'(\boldsymbol{\beta}_t - \widehat{\boldsymbol{\beta}}_t)\big)^2}{\hat{\sigma}_t^2} + \frac{\sigma_t^2}{\hat{\sigma}_t^2} - 1 - log\left(\frac{\sigma_t^2}{\hat{\sigma}_t^2}\right)\right)$$

Proof. By definition of KL divergence, i.e. $E_{\boldsymbol{\theta}_t}\left[\log \frac{f_{\boldsymbol{\theta}_t}(y_t)}{f_{\widehat{\boldsymbol{\theta}}_t}(y_t)}\right]$, and let $f(y_t)$ be $f_{\boldsymbol{\theta}_t}(y_t)$ for short

$$\begin{aligned}
KL &= \int_{-\infty}^{\infty} f_{\boldsymbol{\theta}_t}(y_t)\Bigg[\ \log\frac{\widehat{\sigma}_t}{\sigma_t} - \frac{(y_t - \boldsymbol{x}_t'\boldsymbol{\beta}_t)^2}{2\sigma_t^2} + \frac{(y_t - \boldsymbol{x}_t'\widehat{\boldsymbol{\beta}}_t)^2}{2\widehat{\sigma}_t}\Bigg]dy_t \\
&= \log\frac{\widehat{\sigma}_t}{\sigma_t} - \frac{1}{2\sigma_t^2}\Bigg[\int y_t^2 f(y_t)dy_t - 2\boldsymbol{x}_t'\boldsymbol{\beta}_t\int y_t f(y_t)dy_t + (\boldsymbol{x}_t'\boldsymbol{\beta}_t)^2\int f(y_t)dy_t\Bigg] \\
&\quad + \frac{1}{2\widehat{\sigma}_t^2}\Bigg[\int y_t^2 f(y_t)dy_t - 2\boldsymbol{x}_t'\widehat{\boldsymbol{\beta}}_t\int y_t f(y_t)dy_t + (\boldsymbol{x}_t'\widehat{\boldsymbol{\beta}}_t)^2\int f(y_t)dy_t\Bigg] \\
&= \log\frac{\widehat{\sigma}_t}{\sigma_t} - \frac{1}{2\sigma_t^2}\big((\boldsymbol{x}_t'\boldsymbol{\beta}_t)^2 + \sigma_t^2 - 2(\boldsymbol{x}_t'\boldsymbol{\beta}_t)^2 + (\boldsymbol{x}_t'\boldsymbol{\beta}_t)^2\big) \\
&\quad + \frac{1}{2\widehat{\sigma}_t^2}\big(\sigma_t^2 + (\boldsymbol{x}_t'\boldsymbol{\beta}_t)^2 - 2\boldsymbol{x}_t'\widehat{\boldsymbol{\beta}}_t\boldsymbol{x}_t'\boldsymbol{\beta}_t + (\boldsymbol{x}_t'\widehat{\boldsymbol{\beta}}_t)^2\big) \\
&= \frac{1}{2}\left(\frac{\big(\boldsymbol{x}_t'(\boldsymbol{\beta}_t - \widehat{\boldsymbol{\beta}}_t)\big)^2}{\hat{\sigma}_t^2} + \frac{\sigma_t^2}{\hat{\sigma}_t^2} - 1 - \log\left(\frac{\sigma_t^2}{\hat{\sigma}_t^2}\right)\right)
\end{aligned}$$

□

References

Acemoglu, D., & Scott, A. (1997). Asymmetric business cycles: theory and time-series evidence. *Journal of Monetary Economics*, *40*, 501–533.

Albert, J. H., & Chib, S. (1993). Bayes inference via gibbs sampling of autoregressive time series subject to markov mean and variance shifts. *Journal of Business & Economic Statistics*, *11*.

Albert, P. (1991). A two-stage markov mixture model for a time series of epileptic seizure counts. *Biometrics*, *47*, 1371–1381.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*, 307–327.

Cai, J. (1994). A markov model of switching-regime arch. *Journal of Business & Economic Statistics*, *12*, 309–316.

Cappé, O., Moulines, E., & Rydén, T. (2005). *Inference in hidden markov models*. Springer.

Casella, G., & Berger, R. L. (2001). *Statistical infference* (2, Ed.). Duxbury Press.

Chauvet, M., Juhn, C., & Potter, S. (2002). Markov switching in disaggregate unemployment rates. *Empirical Economics*, *27*, 205–232.

Chib, S. (1996). Calculating posterior distributions and modal estimates in markov mixture models. *Journal of Econometrics*, *75*, 79–97.

Cho, J., & White, H. (2007). Testing for regime switching. *Econometrica*, *75*, 1671–1720.

Dempster, A., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood for incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Ser. B*, *39*, 1–37.

Diebold, F., Lee, J.-H., & Weinbach, G. (1994). Regime switching with time-varying transition probabilities. In C. Hargreaves (Ed.), *Nonstationary time series analysis and cointegration.* Oxford: Oxford University Press.

Dueker, M. (1997). Markov switching in garch processes and mean-reverting stock-market volatility. *Journal of Business & Economic Statistics*, *15*, 26–34.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, *50*, 987–1007.

Everitt, B. S., & Hand, D. J. (1981). *Finite mixture distributions.* London: Chapman & Hall.

Filardo, A. (1998). Business cycle phases and their transitional dynamics. *Journal of Business & Economic Statistics*, *12*, 299–308.

Francq, C., Roussignol, M., & Zakoian, J. (2001). Conditional heteroscedasticity driven by hidden markov chains. *Journal of Time Series Analysis*, *22*, 197–220.

Frühwirth-Schnatter, S. (2004). Estimating marginal likelihoods for mixture and markov switching models using bridge sampling techniques. *The Econometrics Journal*, *7*, 143–167.

Frühwirth-Schnatter, S. (2006). *Finite mixture and markov switching models.* Springer.

Garcia, R. (1998). Asymptotic null distribution of the likelihood ratio test in markov switching models. *International Economic Review*, *39*, 763–788.

Ghahramani, Z. (2001). An introduction of hidden markov models and bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence*, *15*(1), 9–42.

Goldfeld, S. M., & Quandt, R. (1973). A markov model for switching regressions. *Journal of Econometrics*, *1*, 3–16.

Gray, S. F. (1996). Modelling the conditional distribution of interest rates as a regime switching process. *Journal of Financial Economics*, *42*, 27–62.

Hamilton, J. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, *57*, 357–384.

Hamilton, J. (1990). Analysis of time series subject to changes in regime. *Journal of Econometrics*, *45*, 39–70.

Hamilton, J. (1996). Specification testing in markov-switching time-series models. *Journal of Econometrics*, *70*, 127–157.

Hamilton, J., & Raj, B. (2002). New directions in business cycle research and financial analysis. *Empirical Economics*, *27*, 149–162.

Hamilton, J., & Susmel, R. (1994). Autoregressive conditional heteroskedasticity and changes in regime. *Journal of Econometrics*, *64*, 307–333.

Hamilton, J. D. (1994). *Time series analysis.* Princeton University Press.

Hansen, B. (1992). The likelihood ratio test under nonstandard conditions: testing the markov switching model of gnp. *Journal of Applied Econometrics*, *7*, 61–82.

Holst, U., Lindgren, G., Holst, J., & Thuvesholmen, M. (1994). Recursive estimation in switching autoregressions with a markov regime. *Journal of Time Series Analysis*, *15*, 489–506.

Kaufmann, S. (2000). Measuring business cycles with a dynamic markov switching factor model: an assessment using bayesian simulation methods. *Econometrics Journal*, *3*, 39–65.

Kaufmann, S., & Frühwirth-Schnatter, S. (2002). Bayesian analysis of switching arch models. *Journal of Time Series Analysis*, *23*, 425–458.

Kim, C.-J. (1994). Dynamic linear models with markov-switching. *Journal of Econometrics*, *60*, 1–22.

Kim, C.-J., & Nelson, C. R. (1998). Business cycle turning points, a new coincident index, and tests of duration dependence based on a dynamic factor model with regime switching. *The Review of Economics and Statistics*, *80*, 188–201.

Kim, C.-J., & Nelson, C. R. (1999). *State-space models with regime switching: Classical and gibbs-sampling approaches with applications.* The MIT Press.

Klaasen, F. (2002). Improving garch volatility forecasts with regime-switching garch. *Empirical Economics*, *27*, 363–394.

Krolzig, H.-M. (1997). Markov-switching vector autoregressions: modelling, statistical inference, and application to business cycle analysis. In *Lecture notes in economics and mathematical systems.* New York/Berlin/Heidelberg: Springer.

Lai, T., Liu, H., & Xing, H. (2005). Autoregressive models with piecewise constant volatility and regression parameters. *Statistical Sinica*, *15*, 279–301.

Lai, T., & Xing, H. (2011). A simple bayesian approach to multiple change-points. *Statistical Sinica*, *21*, 539–569.

Lai, T., Xing, H., & Zhang, N. (2008). Stochastic segmentation models for array-based comparative genomic hybridization data analysis. *Biostatistics*, *9*, 290–307.

MacDonald, I., & Zucchini, W. (1997). Hidden markov and other models for discrete-valued time series. In *Monographs on statistics and applied probability.* London: Chapman & Hall.

McCulloch, R. E., & Tsay, R. S. (1994). Statistical analysis of economic time series via markov switching models. *Journal of Time Series Analysis*, *15*, 523–539.

Neftçi, S. (1984). New directions in business cycle research and financial analysis. *Empirical Economics*, *27*, 149–162.

Peria, M. (2002). A regime-switching approach to the study of speculative attacks: A focus on the ems crisis. *Empirical Economics*, *27*, 299–334.

Perlin, M. (n.d.-a). *Ms_ regress - a package for markov regime switching models in matlab.* `https://sites.google.com/site/marceloperlin/matlab-code/ms_regress---a-package-for-markov-regime-switching-models-in-matlab`. (Accessed Feb 4, 2016)

Perlin, M. (n.d.-b). *Ms-regress r package.* `https://sites.google.com/site/marceloperlin/r-code`. (Accessed Feb 4, 2016)

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

Quandt, R. (1972). A new approach to estimating switching regressions. *Journal of the American Statistical Association*, *67*, 306–310.

Rabiner, L., & Juang, B. (1986). An introduction to hidden markov models. In *Acoustic, speech, signal processing magazine* (Vol. 3, p. 4-16). IEEE.

Scott, S. (2002). Bayesian methods for hidden markov models: Recursive computing in the 21st century. *Journal of the American Statistical Association*, *97*, 337–351.

Timmermann, A. (2000). Moments of markov switching models. *Journal of Econometrics*, *96*, 75–111.

Titterington, D., Smith, A. F. M., & Markov, U. E. (1985). *Statistical* *analysis of finite mixture distributions*. New York: Wiley Series in Probability and Statistics.

Wang, P., & Puterman, M. L. (1999). Markov poisson regression models for discrete time series. part i: Methodology. *Journal of Applied Statistics*, *26*, 855–869.

Wong, C., & Li, W. (2001). On a mixture autoregressive conditional heteroscedastic model. *Journal of American Statistical Association*, *96*, 982–995.

Yao, Y. C. (1984). Estimation of a noisy discrete-time step function: Bayes and empirical bayes approaches. *Annals of Statistics*, *12*, 1434–1447.