NEPAL RESPIRATORY INFECTION ANALYSIS WITH LATENT VARIABLE MODELS

by

Wenyi Lin

A thesis submitted to Johns Hopkins University in conformity with the requirements for the degree of Master of Science

Baltimore, Maryland
April, 2017
©2017 Wenyi Lin
All Rights Reserved
Abstract

In this thesis, we implement latent variable models, including latent class analysis (LCA) and hidden Markov models (HMMs), to estimate whether or not an infant has acute respiratory infection (ARI) each day, using 4 common ARI symptoms. Many studies define ARI as a simple function of the observed symptoms ignoring the measurement error and time dependence across successive days. LCA accounts for the measurement error, but not the time dependence; HMMs account for both. We implement three computational approaches to estimate HMMs: EM algorithm, MCMC-based Bayes inference, and the method of moments spectral algorithm. Using pilot simulation experiments, we compare these three methods and conclude that EM algorithm and MCMC inference produce more accurate estimates of model parameters but are much less computationally efficient compared with the spectral algorithm. In an analysis of respiratory infection data for Nepali infants, we compare rule-based, LCA and spectral HMM approaches. The LCA results show that all four symptoms have high specificities but much lower and different sensitivities. The HMMs provide estimates of incidence and transition rates but its estimates of the marginal rates, directly observable from the data, are biased downward. Further investigation of its utility for ARI monitoring is needed to address questions remaining about the bias in the spectral implementation of HMMs with low prevalence outcomes.

Thesis Advisor: Dr. Scott L. Zeger, Professor, Department of Biostatistics
Committee Member: Dr. Joanne Katz, Professor, Department of International Health
Acknowledgments

I would like to express my deep gratitude to my thesis advisor, Dr. Scott L. Zeger, for his consistent patience, detailed instruction and insightful comments at every stage of this thesis. His attitude and carefulness toward research has a great impact on my thesis work and will guide me through my future academic studies.

I also would like to thank Dr. Joanne Katz for being the committee member for my thesis. She helped me understand the background of the problem and provided valuable advices for my thesis writing.

Furthermore, many of my fellows offered great help during my study and research in this department. I would like to thank Dr. Zhenke Wu for helping me understand the models and methods at the very beginning; Detian Deng for sharing his experiences with hidden Markov models; Junrui Di for editing the writing of my thesis; other friends for their kindness and encouragement.

Finally, I would like to thank my family for supporting me all the time and words can’t express my gratitude for them.
# Table of Contents

1 Introduction .................................................. 1
   1.1 Motivation .............................................. 1
   1.2 Statement of problems .................................. 1
   1.3 Previous work .......................................... 2
   1.4 Our approach ........................................... 4
   1.5 Organization of the thesis ............................. 5

2 Methods ....................................................... 6
   2.1 Hidden Markov model .................................... 6
      2.1.1 Model Definition .................................. 6
      2.1.2 Properties ........................................ 8
   2.2 EM algorithm ............................................ 9
      2.2.1 The forward-backward procedure .................. 10
      2.2.2 Baum-Welch algorithm ............................ 11
   2.3 Bayesian inference ...................................... 12
      2.3.1 MCMC and Gibbs sampler .......................... 12
      2.3.2 Posterior sampling of parameters ................ 13
      2.3.3 Label switching .................................... 15
   2.4 Spectral HMM ........................................... 15
      2.4.1 Methods of Moments ............................... 16
      2.4.2 Spectral algorithm ............................... 16
      2.4.3 Observable moments and operators ............... 17
      2.4.4 Estimation procedure ............................. 19
      2.4.5 Complexity bound ................................ 20
   2.5 Viterbi algorithm ...................................... 20
   2.6 Latent class analysis ................................... 21
   2.7 Summary ............................................... 23

3 Simulation .................................................... 24
   3.1 Simulation data description ............................ 24
   3.2 Hidden Markov approaches comparison ................. 25
List of Tables

1. Prediction results and errors of Scenarios 1, 2, 3 using rule-based approach, LCA and HMM-VA, compared with simulated true values. ................................................................. 28
2. Frequencies of the 16 possible combinations of symptoms in the Nepal respiratory infection data ............................................................ 31
3. Empirical symptoms combination transition ($\hat{P}(y_t|y_{t-1})$) matrix .................................................. 32
4. LCA and HMM estimated percentage of observed phenotypes compared with observed values. ......................................................... 34
5. Comparison of estimated prevalence, transition probability from health to sick, transition probability from sick to sick, false positive rate and true positive rate for rule-based approach, LCA and HMM. ................................................................................ 36
List of Figures

1 Comparison of the accuracy of estimating $P(\eta_t = 1|\eta_{t-1} = 0)$ (top row) and $P(\eta_t = 1|\eta_{t-1} = 1)$ (bottom row) of EM algorithm, MCMC and Method of Moments for each of the three measurement error scenarios (columns 1-3). .......................... 26

2 Comparison of the accuracy of estimating $P(\eta_t = 1|\eta_{t-1} = 0)$ and $P(\eta_t = 1|\eta_{t-1} = 1)$ using the Method of Moments with increasing sample size $N = 100, 300, 500$ .................................................. 27

3 Observed ARI prevalence in a calendar time order from August 2012 to April 2014. ................................................................. 32

4 Estimated sensitivity ($P(y = 1|\eta = 1)$) and specificity ($P(y = 0|\eta = 0)$) from LCA for each symptom respectively. ................. 34

5 Prevalence in calendar time of HMM predicted trend and observed trend. ................................................................. 36
1 Introduction

1.1 Motivation

An estimated 1.9 million children under 5 died from ARIs in 2000, 70% of them in Africa and Southeast Asia (Williams, et al, 2002)[1]. Respiratory Infections (ARI) can be present in the upper or lower airway, with lower ARI being more severe and often leading to pneumonia. Pneumonia is the primary cause of death among children 1 month through 5 years of age worldwide, with the greatest burden in developing countries (Liu, et al, 2015)[2]. The World Health Organization defines pneumonia as an elevated respiratory rate or chest indrawing, with or without fever or cough [3]. Other symptoms of ARI in children, in addition to cough and fever are otitis media and runny nose. The broad objective of this thesis is to compare rule-based definitions of ARI to an alternate approach that treats the actual ARI state as a latent (unobserved) variable and then infers its value from reports on multiple symptoms using a statistical model.

1.2 Statement of problems

Johns Hopkins researchers conducted two consecutive, population-based, randomized, placebo-controlled trials of maternal influenza immunization in rural southern Nepal from April 2011 to May 2014 (Tielsch, et al.)[4]. Households were visited every week and mothers were asked to recall any daily respiratory symptoms in the past week for their infants under the age of 6 months. If the infant met an ARI symptoms’ criterion in the past week, health workers also collected infants’ nasal swabs to attempt to identify the pathogen causing the infection. In this thesis, we focus on the daily symptom reports because the lab testing is only available on a small subset of days. Prior statistical analyses[4] in these studies used rule-based definitions of ARI applied to multiple symptoms each day.

There are strengths and weaknesses of a rule-based approach. A major strength
is that the ARI measure is observed so that standard regression models can be used to address the study hypotheses. A limitation is that the definition and estimated rates of ARI vary among studies with different symptom rules. For example, Taylor and Nakai (2012)[5] proposed that ARI was defined as cough followed by rapid breathing. However, Kumar, et al. (2015)[6] included other recognizable symptoms including runny nose, sore throat, chest retraction into the definition of ARI. The variation in ARI definition can complicate inference and interpretation of estimates of ARI prevalence, incidence and episode duration. A second potential limitation is that all respiratory symptoms may not have the same degree of information about the actual ARI state. The sensitivity and specificity of these symptom reports likely vary among the symptoms.

The primary question addressed in this thesis is whether we can reduce the reliance on specific definitions and better estimate the underlying respiratory infection state by scientifically pooling the available symptom reports using a latent variable statistical modeling approach.

1.3 Previous work

Previous work (Spycher, et al., 2008)[7] on asthma has relied upon latent class analysis (LCA) to infer a child’s asthma state, given reports on multiple symptoms. Unlike an ARI, asthma is a chronic respiratory disease, characterized by the respiratory symptoms of wheeze and coughing that can persist for long periods of time. It is widely accepted that wheeze and coughing are symptoms of several distinct diseases for which the appropriate therapy differs. Latent class analysis (LCA) was introduced in (Spycher, et al. 2008)[7] to identify wheeze and coughing phenotypes and their research provided support for the distinction between transient and persistent wheeze and the existence of a third form of wheezing. The model also identified a group that exhibits features of a condition called “cough-variant asthma”.

However, LCA assumes that the children’s state on one day is independent
of the state on days before and after. While this limitation may not be serious with observations separated by weeks or months, it is an issue for the 6 months of daily ARI data in our motivating data because serial correlation of the true state across successive days is expected. This time series feature is not acknowledged in LCA methods.

The natural extension of LCA to allow for serial correlation is the Hidden Markov model (HMM). Here, the true infection state is assumed to follow a Markov process. Conditional on the true state, the symptoms are assumed to be independent of one another at the same and different times. Strat and Carrat (1999)[8] proposed a precursor to an HMM with a continuous state variable. They characterized the incidence rate of influenza-like illness (ILI) based on surveillance data using a mixture of Gaussian distributions. They computed weekly ILI incidence rates taking into account the size of the underlying population and the participating physicians. Rath, et al. (2003)[9] then modified their model into a two-state HMM, assigning different distributions to epidemic and non-epidemic rates. In these studies, continuous probability distributions were applied to observed states given hidden variables using surveillance data. These methods were not directly appropriate for our problem where the underlying disease state is more reasonably assumed to be binary (infected or not) or at a minimum discrete.

Scott, et al. (2005)[10] proposed an HMM to analyze multivariate longitudinal data, comparing the side effects of patients with schizophrenia in two different medical groups. They inferred that there are seven states of health ranging from no side effects to severe side effects. Similarly, Mass, et al. (2006)[11] used a hidden Markov model to model the process underlying migraine symptoms. They used the HMM to quantify drug effects on mitigating the severity of migraine attacks. These prior studies demonstrate the utility of hidden Markov models in longitudinal studies.
1.4 Our approach

Based on this previous work, we propose a hidden Markov model to model the binary latent acute respiratory infection (ARI) state. The latent states are denoted ARI absence (0) or presence (1). We also assume that the observed respiratory symptoms are independent of one another at different times given the hidden states. Two models are used to predict the distribution of hidden states. The first approach is to assume the symptoms are also conditionally independent at the same time. In this case, we allow each symptom to have its own sensitivity and specificity. In this way, we allow each respiratory symptom to provide different information about the underlying infectious state. The second approach is to assume that symptoms may be correlated—for instance, cough is supposed to arise along with fever. Here, we model the most common combinations of symptoms to infer the hidden state.

A natural approach to parameter estimation in HMMs is to treat hidden states as missing data and to maximize the observed data likelihood using the EM algorithm (Dempster, Laird and Rubin, 1977)[12]. The EM algorithm for the HMM uses a forward−backward algorithm (Rabiner and Juang, 1986)[13]. However, the HMM EM algorithm has been shown to be susceptible to converging to a local rather than global maximum and has the drawback of slow convergence (Redner and Walker, 1984)[14]. Two alternative approaches we explore are a full Bayesian inference using Markov Chain Monte Carlo (MCMC) and a method of moments approach called spectral−computation. MCMC output can be efficiently used to estimate the hidden chain (Scott 2002)[15] and gives credible intervals in addition to a point estimate (Ryden 2008)[16]. The spectral approach, on the other hand, is computationally efficient. Hsu, et al. (2012)[17] proposed a related algorithm, giving the idea of observable representation of probabilistic quantities, makes it particularly applicable to settings with a large number of observations. Anandkumar, et al. (2012)[18] followed their basic ideas and extended it to multi-view mixture models, discussing both multi-view mixtures of Gaussians and HMMs. Both the Bayesian
approach and spectral methods have their own advantages that we explore in
the coming chapters.

1.5 Organization of the thesis

The outline of the remainder of this thesis is as follows. In Chapter 2, we
review the literature of methods dealing with respiratory infection data incor-
porating latent class analysis, in particular, the hidden Markov model. For the
hidden Markov model, a detailed derivation done by others will be presented
for three estimation approaches: the EM algorithm, Bayesian inference, and
the method of moments method called spectral HMM. The Viterbi algorithm
(Forney 1973)[19] will be introduced as a method for estimating the latent ARI
state, what computer scientists call the decoding approach for HMMs.

The third chapter presents and discusses the results from a small (pilot) simu-
lation study designed to compare the competing methods. First, we introduce
the parameter cases for which data are simulated. Then, we apply the compet-
ing methods, including rule-based, LCA, and the three HMM computational
approaches to compare their results.

In the fourth chapter, we apply the competing methods to the Nepal infants’
respiratory infection data. In the first sub-section, the source and structure
of the data are discussed. Then, we apply hidden Markov models and LCA
to estimate the latent states of respiratory symptoms. In the last sub-section,
we will predict and interpret the results based on our algorithms. Chapter
five is a discussion of the main contributions and limitations of this thesis and
opportunities for future work.
2 Methods

In this chapter, a thorough review of the methodology used in this thesis is presented. Section 2.1 covers the background and properties of hidden Markov models. In Sections 2.2, 2.3, and 2.4, EM algorithm, Bayesian inference, and the spectral approach of estimating HMMs are presented respectively. Section 2.5 and 2.6 provides an overview to the Viterbi algorithm and the latent class model.

2.1 Hidden Markov model

An early application of Hidden Markov models was introduced by Leonard E. Baum and his colleagues (1966)[20] in speech processing. Since then, they have been used for at least three decades in speech and other signal-processing applications (Lee and Hon 1989[21], Rabiner and Juang 1993[22]). More recently, HMMs has gained more popularity in other fields including bioinformatics (Churchill 1999[23]), image analysis (Romberg, Choi, and Baraniuk 2000[24]) and finance (Hamilton 1989[25]).

2.1.1 Model Definition

i) Markov process
Let \( p(\cdot) \) denote a probability distribution function. A sequence of discrete random variables \( \{\eta_t : t \in \mathbb{N}\} \) is considered to be a first-order Markov chain if, for all \( t \in \mathbb{N} \), it satisfies the Markov property

\[
p(\eta_{t+1} | \eta_t, \ldots, \eta_1) = p(\eta_{t+1} | \eta_t).
\]

That is, conditioning on all previous observations up to time \( t \) is equivalent to conditioning only on the most recent value \( \eta_t \). Said another way, for a first order Markov chain, the future state is independent of the past given the present. The conditional probability of the observation at time \( t + 1 \) given the previous observation at time \( t \) is defined as a first-order transition probability and denoted by:

\[
\gamma_{ij} = p(\eta_{t+1} = j | \eta_t = i)
\]
The matrix $\Gamma$ is defined as the transition matrix of this Markov chain whose $(i, j)$ element is $\gamma_{ij}$ and with row sums equal to 1.

If a Markov chain has a stationary distribution $\delta$, then it satisfies $\delta\Gamma = \delta$ and $\delta 1' = 1$. (Zucchini and MacDonald 2009)[26]

ii) Hidden states as latent variables

A hidden Markov model or HMM is a stochastic process with an underlying Markov process that is not directly observable. Rather, the process state is manifest through a series of observable symbols or variables (Rabiner and Juang 1986)[13]. With $\eta_t$ and $y_t$ representing the hidden states and observed values from one specific sequence, following the notation in Zucchini and MacDonald’s book [26], we let:

- $|T| = 1, 2, ..., T$ time points of the observation sequence
- $|K| = 1, 2, ..., K$ discrete possible hidden states
- $|D| = 1, 2, ..., D$ discrete observations (called symbols) in the literature [26]
- $\Gamma = \{\gamma_{ij}\} = p(\eta_{t+1} = j | \eta_t = i), \ i, j \in [K]$, state transition probability distribution
- $\pi = \{\pi_i\} = p(\eta_1 = i), \ i \in [K]$, initial state probability distribution
- $O = \{o_{ij}\} = \{o_i(j)\} = p(y_t = j | \eta_t = i), \ i \in [K], j \in [D]$, observation probability distribution

iii) Transition matrix and observation matrix

We use the compact notation $\lambda = (\pi, \Gamma, O)$ to represent an HMM. If we define $\delta$ to be the stationary distribution of the Markov chain, we would have the relationship between initial probability and transition probability given by $\delta\Gamma = \delta$.

The observation matrix can be expressed in other ways depending on the specific model structures. For example, it can be rewritten in a vector-valued form. An observation at time point $t$ is represented by a random vector $y_t$ and in this
case, the observation matrix would be introduced as [18]

\[ O = [o_1 | o_2 | ... | o_k] \in \mathbb{R}^{d \times k} \]

The columns of the matrix denote the conditional means of the observations \( y_t \), given the corresponding discrete hidden state \( \eta_t \)

\[ E(y_t | \eta_t = i) = O e_j = o_i, i \in [k] \]

where \( e_i \) is the \( i \)-th vector of the \( d \)-dimensional standard basis.

In some cases, we have \( n \) possible conditionally independent measurements given a hidden state. If each measurement has \( m \) possible levels, we would have \( m^n = d \) observations in all. Consider a simple case involving binary hidden states (\( k = 2; \eta_t = 0, 1, t \in [T] \)) with one observable measurement that also has binary outcomes: 0, 1. In order to calculate the probability \( p(y_t = i | \eta_t = j) \), where \( i \in \{1, 2\}, j \in \{1, 2\} \), it is essential to estimate the true positive rate \( \theta = P(y_t = 1 | \eta_t = 1) \) and false positive rate \( \psi = P(y_t = 1 | \eta_t = 0) \) of each variable. We assume TPR and FPR are constants over the time sequence. A helpful 2 × 2 table is constructed as below:

<table>
<thead>
<tr>
<th></th>
<th>( \eta_t = 0 )</th>
<th>( \eta_t = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \eta_t = 0 )</td>
<td>1-( \psi )</td>
<td>( \psi )</td>
</tr>
<tr>
<td>( \eta_t = 1 )</td>
<td>1-( \theta )</td>
<td>( \theta )</td>
</tr>
</tbody>
</table>

This setting is useful when we consider a disease transmission problem, because the sensitivity and specificity of disease measurement are parameters of interest in their own right and because there is often prior knowledge about a plausible range of values for them.

2.1.2 Properties

Rabiner and Juang (1986)[13] proposed three basic questions, the answers to which make a hidden Markov model useful in real world applications.
• Given the observation sequence $\mathbf{y} = y_1, y_2, ..., y_T$ and the model $\lambda = (\Gamma, \mathcal{O}, \pi)$, how do we compute $p(\mathbf{y}|\lambda)$, the probability of the observation sequence?

• Given the observation sequence $\mathbf{y} = y_1, y_2, ..., y_T$, what can we optimally infer about the state sequence $\mathbf{\eta} = \eta_1, \eta_2, ..., \eta_T$?

• How can we efficiently estimate the model parameters $\lambda = (\Gamma, \mathcal{O}, \pi)$ to maximize the likelihood of the observations $p(\mathbf{y}|\lambda)$.

The methods discussed below have been developed by others over the last 30 years (Rabiner and Juang 1986[13], Scott 2002[15], Hsu et al. 2012[17]) to answer their questions. For simplicity, but without loss of generality, methods and algorithms are described with notation for a single individual. Models for multiple objects follow directly. The EM algorithm and spectral HMM are introduced as frequentist solutions to estimating the HMM parameters. The Viterbi algorithm[19] is a frequentist method for predicting a realization of the hidden states. For contrast, we also consider Bayesian inference that treats the model parameters $\lambda$ and latent variables $\mathbf{\eta}$ as random variables producing a high-dimensional parameter structure.

2.2 EM algorithm

A sequence of hidden states is not observed in HMM and a natural approach for parameter estimation would be EM algorithm (Dempster, Laird and Rubin 1977)[12], treating hidden states as missing data. EM algorithm is used to find the value of parameters that maximize the likelihood given the observed data. Each iteration of EM algorithm involves two steps: expectation step (E-step) and maximization step (M-step). The algorithm is described informally below. Choose the initial value for parameters and then iterate these steps (Zucchini and MacDonald 2009 Chap 4)[26].

• **E step** Compute the conditional expectations of the complete data log likelihood (or score function) that involves missing data given the observations and current value of parameters.
• **M step** With respect to parameters, maximize the complete-data log-likelihood.

The procedure repeats until convergence when the change in parameters is less than some threshold. The estimated parameters converge to a stationary point; in some cases, the stationary point could be a local maximum or a saddle point. The specific algorithm by Rabiner and Juang (1986)[13] is shown below.

### 2.2.1 The forward-backward procedure

For every fixed state sequence \( \eta = \eta_1, ..., \eta_T \), the probability of the observation sequence \( y \) is,

\[
p(y|\eta, \lambda) = o_{\eta_1}(y_1)o_{\eta_2}(y_2)...o_{\eta_T}(y_T)
\]

We also know the probability of such a state sequence is,

\[
p(\eta|\lambda) = \pi_{\eta_1}\gamma_{\eta_1,\eta_2}\gamma_{\eta_2,\eta_3}...\gamma_{\eta_{T-1},\eta_T}
\]

The joint probability of \( \eta \) and \( y \) is simply the product of the above two terms,

\[
p(\eta, y|\lambda) = p(y|\eta, \lambda)p(\eta|\lambda).
\]

With this basic setting, we introduce the forward-backward procedure to perform the following calculation.

For \( t = 1, 2, ..., T \), the forward variable, \( \alpha_t(i) \), is defined as,

\[
\alpha_t(i) = P(y_1, y_2, ..., y_t, \eta_t = i|\lambda)
\]

that is the joint probability of partial observation sequence up to time \( t \) and hidden state \( \eta_t \), given the parameters \( \lambda \). As stated in the reference, We can solve for \( \alpha_t(i) \) inductively, as follows:

**Proposition 2.1.** [13]

1. \( \alpha_1(i) = \pi_i o_i(y_1), \quad 1 \leq i \leq k \);

2. \( \alpha_{t+1}(j) = \left[ \sum_{i=1}^k \alpha_t(i)\gamma_{ij} \right] o_j(y_{t+1}), \quad 1 \leq j \leq k, \quad t = 1, 2, ..., T - 1 \)

Similarly, we can consider a backward variable, \( \beta_t(i) \), is defined as,

\[
\beta_t(i) = P(y_{t+1}, y_{t+2}, ..., y_T|\eta_t = i, \lambda)
\]

the joint probability of partial observation sequence from \( t + 1 \) to the end, given the hidden state \( \eta_t \) and parameters \( \lambda \). Again we can solve it inductively,

**Proposition 2.2.** [13]
1. $\beta_T(i) = 1, \quad 1 \leq i \leq k$;
2. $\beta_t(i) = \sum_{j=1}^{d} \gamma_{ij} o_j(y_{t+1}) \beta_{t+1}(j), \quad 1 \leq i \leq k, \quad t = T - 1, T - 2, \ldots, 1$

With forward and backward variables, we now establish the probability

$$P(y_1, y_2, \ldots, y_T, \eta_t = i|\lambda) = \alpha_t(i) \beta_t(i)$$

Consequently, the full likelihood probability for each such $t$ is, $\sum_{i=1}^{k} \alpha_t(i) \beta_t(i)$.

### 2.2.2 Baum-Welch algorithm

For maximum likelihood estimation, the goal is to find the parameter values in the HMM given the observations that maximize the probability of the observation sequence. Here we state the iterative procedure, Baum-Welch algorithm, to estimate the model parameters.

Given the observation sequence $\mathbf{y}$ and parameters $\lambda$, the probability of hidden state $\eta_t$ is,

$$\mu_t(i) = P(\eta_t = i|\mathbf{y}, \lambda) = \frac{\alpha_t(i) \beta_t(i)}{p(\mathbf{y}|\lambda)}$$

Meanwhile, the second-order probability of $\eta_t$ and $\eta_{t+1}$ given the model is defined as,

$$\nu_t(i, j) = P(\eta_t = i, \eta_{t+1} = j|\mathbf{y}, \lambda) = \frac{\alpha_t(i) \gamma_{ij} o_j(y_{t+1}) \beta_{t+1}(j)}{p(\mathbf{y}|\lambda)}$$

hence we relate $\mu_t(i)$ to $\nu_t(i, j)$ by summing $\nu_t(i, j)$ over $j$, giving, $\mu_t(i) = \sum_{j=1}^{d} \nu_t(i, j)$.

By defining these two probabilities, both the E-step and M-step can be generated as follows:[13]

**E step** Sum $\mu_t(i)$ over the time index $t$ to get the expected number of times that state $i$ is visited and regard the sum $\nu_t(i, j)$ over $t$ as the expected value of number of transitions form state $i$ to $j$, that is,

$$\sum_{t=1}^{T-1} \mu_t(i) = \text{Expected number of transitions made from } i$$

$$\sum_{t=1}^{T-1} \nu_t(i, j) = \text{Expected number of transitions made from } i \text{ to } j$$

**M step** The reestimation formulas for the parameters are:

$$\hat{\pi}_i = \mu_1(i), \quad 1 \leq i \leq k$$
\[
\hat{\gamma}_{ij} = \frac{\sum_{t=1}^{T-1} \nu_t(i, j)}{\sum_{t=1}^{T-1} \mu_t(i)}
\]

\[
\hat{o}_i(j) = \frac{\sum_{t=1, y_t=j}^{T} \mu_t(i)}{\sum_{t=1}^{T} \mu_t(i)}
\]

The newly estimated model parameters \( \hat{\lambda} = (\hat{\pi}, \hat{\Gamma}, \hat{O}) \) are used to iteratively replace the initial setting of \( \lambda \) and the above procedure is repeated. Then, the probability of observation sequence \( y \) is improved until some limiting point is reached.

### 2.3 Bayesian inference

An alternative approach to the EM algorithm is Bayesian estimation. Scott (2002)[15] used a Markov chain Monte Carlo (MCMC) sampling strategy to simulate HMM parameters and latent variables from their posterior distribution given the observed data.

#### 2.3.1 MCMC and Gibbs sampler

Geman and Geman were early adapters of The Gibbs sampler among statistical scientists (1984)[27]. MCMC enjoyed a surge of popularity due to the work of Gelfand and Smith (1990)[28] who applied the approach to solve the generic Bayes estimation problem and stimulated its application to a wide variety of substantive problems. The Gibbs sampler is an MCMC technique for generating random variables from the marginal distributions of the joint posterior distribution indirectly (Casella and George 1992)[29]. An example of sampling a pair of random variables in Casella and George (1992)[29] illustrates the basic Gibbs sampler process.

Suppose a joint density \( f(x, y) \) was given and we are interested in the properties of the marginal density

\[
f(x) = \int f(x, y) dy
\]
such as its mean or variance. Rather than directly compute the $f(x)$, the Gibbs sampler approach is to sample $X_1, \ldots, X_m \sim f(x)$ without requiring $f(x)$. With a randomly initial pair of $(X, Y)$, a “Gibbs sequence” $(Y'_0, X'_0, Y'_1, X'_1, \ldots, Y'_m, X'_m)$ is generated from

$$X'_j \sim f(x|Y'_j = y'_j) \quad Y'_{j+1} \sim f(y|X'_j = x'_j)$$

which is referred as Gibbs sampling. Under reasonable conditions, the distribution of $X'_m$ converges to $f(x)$ as $m \to \infty$.

In this way, to calculate the mean of $f(x)$, we could use the fact that

$$\lim_{m \to \infty} \frac{1}{m} \sum_{i=1}^{m} X_i = \int_{-\infty}^{\infty} x f(x) dx = EX$$

The forward-backward algorithm in HMM is applied to many MCMC methods. By mixing it with the MCMC algorithms, the hidden chain can be easily estimated while avoiding the “black box” problem in traditional recursive algorithms (Scott 2002)[15].

Comparing with the other MCMC procedure, i.e. the Metropolis-Hastings procedures (M-H), Gibbs sampler (GS) has its own advantages when dealing with HMM. First of all, M-H algorithms tend to have undesirable performance to solve for high dimensional parameters. Secondly, given the hidden states, the parameters are often nearly independent, which offer a preference of sampling hidden states from posterior distribution by GS (Scott 2002)[15].

2.3.2 Posterior sampling of parameters

By including the hidden states $\eta_1, \ldots, \eta_T$ in the MCMC state space, the Gibbs sampler produces simulated values for both the model parameters and latent states from their respective full conditional distributions given the observations. The reason why this algorithm is effective is that the parameters are conditionally independent given hidden states and observations; moreover, given the model parameters and observed data, the hidden states form a non-homogeneous Markov chain and can be sampled easily.
We consider a hidden Markov model with \( k = 2 \) hidden states \( (\eta_t = 0 \text{ or } 1) \) and observed data that are conditionally independent given the hidden state. For each \( y_t \), there are two different measurements \( y_{t1}, y_{t2} \) and each measurement has 2 possible values \( (y_{ti} = 0 \text{ or } 1) \). The essential assumptions for the two measurements are that: (1) given the hidden state at time point \( t \),

\[
y_{ti}|\eta_t, \theta, \psi \sim \text{Bernoulli}\{I(\eta_t = 1)\theta_i + (1 - I(\eta_t = 0))\psi_i)\}, \quad i = 1, 2
\]

and; (2) the two observations are conditionally independent given the underlying state space. The measurement parameters are the true positive rate (sensitivity) \( \theta \) and false positive rate (1-specificity) \( \psi \) defined as

\[
\theta_i = P(y_i = 1|\eta = 1) \quad \psi_i = P(y_i = 1|\eta = 0)
\]

where \( I(\cdot) \) denotes the indicator function.

With the above specification, the HMM model parameters are \( \lambda = (\pi, \Gamma, \theta, \psi) \). The prior distributions of the parameters are taken to be the conjugate priors. Each row of the transition matrix as well as the initial distribution were given are respectively independent Dirichlet distribution prior \( \text{Dir}(a_1, a_2) \) and \( \text{Dir}(b_1, b_2) \). Each \( \theta_i \) and \( \psi_i \) was given a independent Beta prior \( \text{Beta}(c_1, c_2) \) and \( \text{Beta}(d_1, d_2) \).

Under the prior specification, the full conditional distribution of one sequence of HMM is given by

\[
(\pi_0, \pi_1)|\text{others} \sim \text{Dir}(a_1 + I(\eta_1 = 0), a_2 + I(\eta_1 = 1))
\]

\[
(\gamma_{i0}, \gamma_{i1})|\text{others} \sim \text{Dir}(b_1 + n_{i0}, b_2 + n_{i1})
\]

where \( n_{ij} = \sum_{1 \leq t \leq T-1} I(\eta_t = i, \eta_{t+1} = j) \) is the number of transitions for state \( i \) to \( j \) in the latent sequence and with conditional independence across rows \( i = 1, 2 \).

\[
\theta_k|\text{others} \sim \text{Beta}(c_1 + \sum_{t: \eta_t = k} I(y_{tk} = 1), c_2 + \sum_{t: \eta_t = k} I(y_{tk} = 0))
\]
\[
\psi_k|\text{others} \sim \text{Beta}(d_1 + \sum_{t: y_{tk} = 1} I(y_{tk} = 1), d_2 + \sum_{t: y_{tk} = 0} I(y_{tk} = 0))
\] (4)

where \( k = 1, 2 \). Moreover, the latent chain forms a non-homogeneous Markov chain given model parameters and observed data. The process can be generated by the following distribution:

\[
P(\eta_1 = i|\text{others}) \propto \pi_1 \phi(y_1; \theta_1, \theta_2, \psi_1, \psi_2) p(y_2, ..., y_T|\eta_1 = i, \lambda)
\] (5)

\[
P(\eta_t = j|\eta_{t-1} = i) \propto \gamma_{i,j} \phi(y_t; \theta_1, \theta_2, \psi_1, \psi_2) p(y_{t+1}, ..., y_T|\eta_t = j, \lambda)
\] (6)

where \( \phi \) is the density of observed data given a hidden state. \( p(\cdot) \) is the backward variable of hidden state at time point \( t \). One advantage of applying a Gibbs sampler is that it gives credible intervals immediately after repeated sampling process, while EM only provide a point estimate.

### 2.3.3 Label switching

Sampling parameters from its complete posterior could be affected by an identifiability issue known as label switching, which indicates that the order of estimated parameters could be changed during sampling process (Scott 2002)[15]. One way to solve this problem is by introducing identifiability constraints. Under the constraints’ setting, the posterior is then zero outside this region, which protects the ordering of parameters. The other way, given by Fruhwirth-Schnatter (2001)[30], was to use a permutation-invariant prior for the MCMC simulations.

### 2.4 Spectral HMM

The EM algorithm has a number of well-documented drawbacks, including high computational cost, slow convergence and convergence to local optima.[14] In the past five years, several techniques based on method of moments (Pearson 1894)[31] have been proposed as an alternative to maximize likelihood for learning HMM.
2.4.1 Methods of Moments

In statistics, the method of moments is a method of estimation of population parameters that was first introduced with Pearson’s solution for identifying the parameters of a mixture of two univariate Gaussian.[31] In this approach, model parameters are chosen to specify a distribution whose \( p \)-th order moments, for several values of \( p \), are equal to the corresponding empirical moments observed in the data. But the method often runs into trouble when applying to large mixtures of high dimensional models. Recent work develops a computationally efficient method of moments based on only low-order moments that can be used to estimate the parameters of a broad class of high-dimensional mixture models. The method is implemented with standard numerical linear algebra routines and the estimators tend to be efficient because they rely mainly on lower-order moment (Anandkumar et al. 2012)[18].

The algorithm by (Anandkumar et al. 2012)[18] will be presented in Section 2.4.2-2.4.5. Anandkumar’s notation, lemmas, and theorems are adapted from his paper to the case of the hidden Markov model not explicitly addressed in the original paper. We have also constructed a specific algorithm for HMMs.

2.4.2 Spectral algorithm

i) Notation

The standard inner product between vectors \( \mathbf{u} \) and \( \mathbf{v} \) is denoted by \( \langle \mathbf{u} \mathbf{v} \rangle = \mathbf{u}^T \mathbf{v} \). The \( p \)-norm of a vector \( \mathbf{u} \) is denoted by \( \| \mathbf{u} \|_p \). For a matrix \( A \in \mathbb{R}^{m \times n} \), let \( \| A \|_2 \) denote its spectral norm \( \| A \|_2 := \sup_{\mathbf{v} \neq 0} \| A \mathbf{v} \|_2 / \| \mathbf{v} \|_2 \). \( \sigma_i(A) \) denote the \( i \)-th largest singular value. Let \( \Delta^{n-1} := \{(p_1, p_2, ..., p_n) \in \mathbb{R}^n : p_i \leq 0 \ \forall i, \ \sum_{i=1}^n p_i = 1\} \)

ii) General Setting

Consider the following HMM model; \( k \) denotes the number of hidden states, \( d \) denotes the number of possible observations and \( T \leq 3 \) denotes the number of time points. Let \( \omega = (\omega_1, \omega_2, ..., \omega_k) \in \Delta^{k-1} \) be a vector of mixing weights that
satisfies $\omega = \pi \Gamma$ and $p(\eta = j) = \omega_j$, for all $j \in [k]$. Let $y_1, y_2, ..., y_T \in \mathbb{R}^d$ be $T$ random vectors that are conditionally independent given $\eta$, and was defined by setting

$$y_v = e_i \Leftrightarrow \text{the } v\text{-th observation is } i, \quad i \in [d]$$

Define the conditional mean vectors as

$$\mu_{v,j} := \mathbb{E}(y_v | \eta = j), \quad v \in [T], j \in [k]$$

and let $M_v \in \mathbb{R}^{d \times k}$ be the matrix whose $j$-th column is $\mu_{v,j}$.

We assume the following conditions on $\omega$ and the $M_v$.

**Condition 1** $\omega_j > 0$ for all $j \in [k]$, and $M_v$ has rank $k$ for all $v \in [T]$.

### 2.4.3 Observable moments and operators

Triples are defined as any three consecutive observation of the random vectors $\{y_i, y_{i+1}, y_{i+2}\}$, $i, i+1, i+2 \in [T]$. We focus on the moments concerning $\{y_1, y_2, y_3\}$.

Define $\text{Pairs} \in \mathbb{R}^{d \times d}$ to be the matrix of pair-wise probabilities whose $(i,j)$-th entry is

$$\text{Pairs}_{i,j} := p(y_i = e_i, y_j = e_j), \quad i, j \in [d]$$

Also define $\text{Triples} \in \mathbb{R}^{d \times d \times d}$ to be the third-order tensor of triple-wise probability whose $(i,j,k)$-th entry is

$$\text{Triples}_{i,j,k} := p(y_i = e_i, y_j = e_j, y_k = e_k), \quad i, j, k \in [d]$$

Pairs and Triples could also be viewed as expectations of tensor products of the random vectors $y_1, y_2$ and $y_3$:

$$\text{Pairs} = P_{1,2} = \mathbb{E}(y_1 \otimes y_2) \quad \text{and} \quad \text{Triples} = P_{1,2,3} = \mathbb{E}(y_1 \otimes y_2 \otimes y_3)$$

$P_{1,2,3}$ is regarded as the linear operator $P_{1,2,3}(\eta) := \mathbb{E}[(y_1 \otimes y_2)\langle \eta, y_2 \rangle]$.

**Lemma 1.** $P_{1,2} = M_1 \text{diag}(\omega) M_2^T$ and $P_{1,2,3}(\eta) = M_1 \text{diag}(M_2^T \eta) \text{diag}(\omega) M_2^T$
proof. By conditional independence,

\[
P_{1,2} = \mathbb{E}[y_1 \otimes y_2] = \mathbb{E}[\mathbb{E}[y_1 \otimes y_2|h]]
\]

\[
= \mathbb{E}[\mathbb{E}[y_1|h] \otimes \mathbb{E}[y_2|h]]
\]

\[
= \mathbb{E}[M_1 e_h \otimes M_2 e_h]
\]

\[
= M_1(\Sigma_{k=1}^{k} \omega_k e_t \otimes e_t)M_2^T
\]

\[
= M_1 \text{diag}(\omega)M_2^T
\]

Similarly,

\[
P_{1,2,3}(\eta) = \mathbb{E}[(y_1 \otimes y_2)\langle \eta, y_3 \rangle] = \mathbb{E}[\mathbb{E}[(y_1 \otimes y_2)\langle \eta, y_3 \rangle|h]]
\]

\[
= \mathbb{E}[\mathbb{E}[y_1|\eta] \otimes \mathbb{E}[y_2|\eta]\langle \eta, \mathbb{E}[y_3|\eta]]
\]

\[
= \mathbb{E}[M_1 e_\eta \otimes M_2 e_\eta]\langle \eta, M_3 e_\eta]\rangle
\]

\[
= M_1(\Sigma_{k=1}^{k} \omega_k e_t \otimes e_t\langle \eta, M_3 e_\eta\rangle)M_2^T
\]

\[
= M_1 \text{diag}(M_3^T \eta)\text{diag}(\omega)M_2^T
\]

Lemma 2. Assume condition 1. For \(v \in \{1, 2, 3\}\), let \(U_v \in \mathbb{R}^{d \times k}\) be a matrix such that \(U_v^T M_v\) is invertible. For all \(\eta \in \mathbb{R}^d\), the observable operator is defined by \(B_{1,2,3}(\eta) = (U_1^T P_{1,2,3}(\eta)U_2)\) and it satisfies

\[
B_{1,2,3}(\eta) = (U_1^T M_1) \text{diag}(M_3^T \eta)(U_1^T M_1)^{-1}.
\]

proof.

\[
B_{1,2,3}(\eta) = (U_1^T P_{1,2,3}(\eta)U_2)\]

\[
= (U_1^T M_1) \text{diag}(M_3^T \eta)\text{diag}(\omega)(M_3^T \eta)(U_1^T M_1)^{-1}\]

\[
= (U_1^T M_1) \text{diag}(M_3^T \eta)(U_1^T M_1)^{-1}(U_1^T M_1)\text{diag}(\omega)(M_3^T \eta)(U_1^T M_1)^{-1}\]

\[
= (U_1^T M_1) \text{diag}(M_3^T \eta)(U_1^T M_1)^{-1}(U_1^T P_{1,2,3}(\eta)U_2)(U_1^T M_1)^{-1}\]

\[
= (U_1^T M_1) \text{diag}(M_3^T \eta)(U_1^T M_1)^{-1}.
\]

Lemma 3. In HMM, we have \(M_1 := O \text{diag}(\pi)\Gamma \text{diag}(\pi\Gamma)^{-1}\), \(M_2 := O\) and \(M_3 := O\).

proof. First of all, we know that

\[
p(h_1 = i|\eta_2 = i) = \frac{p(h_2 = j|\eta_1 = i)}{p(\eta_2 = j)} = \frac{\Gamma_{i,j} \cdot \pi_i}{(\Gamma\pi)_j}
\]

\[
= e_i \text{diag}(\pi)\Gamma \text{diag}(\pi\Gamma)^{-1} e_j
\]
Thus,

\[ M_1 e_j = \mathbb{E}[y_1 | \eta_2 = j] = \mathbb{E}[\mathbb{E}[y_1 | h_2 = j, \eta]] \]

\[ = \mathbb{E}[O_{\eta} | \eta_2 = j] = O \mathbb{E}[e_{\eta} | \eta_2 = j] \]

\[ = O diag(\pi) \Gamma diag(\Gamma \pi)^{-1} e_j \]

and similarly,

\[ M_2 e_j = \mathbb{E}[y_2 | \eta_2 = j] = O e_j \]

\[ M_3 e_j = \mathbb{E}[y_3 | \eta_2 = j] = \mathbb{E}[O h_3 | \eta_2 = j] = O \Gamma e_j \]

From Lemma 3, it is easy to verify that

\[ B_{3,1,2}(\eta) = (U_3^T P_{3,1,2}(\eta) U_1)(U_3^T P_{3,1} U_1)^{-1} \]

\[ = (U_3^T M_3) diag(M_2^T \eta)(U_3^T M_3)^{-1} \]

\[ = (U_3^T O \Gamma) diag(O^T \eta)(U_3^T O \Gamma)^{-1} \]

In order to reconstruct the parameters from different observable operators, we propose the following lemma to solve it.

**Lemma 4.** Let \( \Theta \in \mathbb{R}^{k \times k} \) be an invertible matrix, and \( \theta_i^T \in \mathbb{R}^k \) be its \( i \)-th row. And for all \( i \in [k] \), let \( \lambda_{i,1}, \lambda_{i,2}, \ldots, \lambda_{i,k} \) denote the \( k \) eigenvalues of \( B_{3,1,2}(\eta) \), where \( \eta = U_2 \theta_1 \). Let \( L \in \mathbb{R}^{k \times k} \) be the matrix whose \((i, j)\)-th entry is \( \lambda_{i,j} \). Then

\[ \Theta U_2^T M_2 = L \]

### 2.4.4 Estimation procedure

Based on the lemmas we proposed in the last part, we construct an algorithm to estimate our parameters.

1. Compute the empirical averages from \( N \) independent copies of \( y_3 \otimes y_1 \) to form \( \hat{P}_{3,1} \in \mathbb{R}^{d \times d} \), and do the same for \( y_3 \otimes y_2 \) to form \( \hat{P}_3 \in \mathbb{R}^{d \times d} \) and for \( y_3 \otimes y_1 \otimes y_2 \) to form \( \hat{P}_{3,1,2} \in \mathbb{R}^{d \times d \times d} \).

2. Let \( \hat{U}_1 \in \mathbb{R}^{d \times k} \) and \( \hat{U}_3 \in \mathbb{R}^{d \times k} \) be, respectively, matrices of orthonormal left and right singular vectors of \( \hat{P}_{3,1} \), corresponding to its \( k \) singular values. Similarly, let \( \hat{U}_2 \in \mathbb{R}^{d \times k} \) be the matrix of orthonormal right singular vectors of \( \hat{P}_{3,2} \).
3. Take a randomly chosen \( \Theta \in \mathbb{R}^{k \times k} \) from the subspace range \((\hat{U}_2)\) with its \(i\)-th row equal to \(\theta_i^T\). Let \(\eta = \hat{U}_2 \theta_1\).

Form the matrix, \(\hat{B}_{3,1,2}(\eta) = (\hat{U}_3^T \hat{P}_{3,1,2}(\eta) \hat{U}_1)(\hat{U}_3^T \hat{P}_{3,1} \hat{U}_1)^{-1}\).

Compute \(\hat{R}_1 \in \mathbb{R}^{k \times k}\) that diagonalizes \(\hat{B}_{3,1,2}(\hat{U}_2 \theta_1)\), i.e., \(\hat{R}_1^{-1} \hat{B}_{3,1,2}(\hat{U}_2 \theta_1) \hat{R}_1 = \text{diag}(\lambda_{1,1}, \lambda_{1,2}, \ldots, \lambda_{1,k})\).

4. For each \(i \in [k]\), compute the diagonal entries \(\hat{\lambda}_{i,1}, \hat{\lambda}_{i,2}, \ldots, \hat{\lambda}_{i,k}\) of \(\hat{R}_1^{-1} \hat{B}_{3,1,2}(\hat{U}_2 \theta_1) \hat{R}_1\), and form the matrix \(L \in \mathbb{R}^{k \times k}\) whose \((i, j)\)-th entry is \(\lambda_{i,j}\).

5. Return \(\hat{O} = \hat{M}_2 = \hat{U}_2 \Theta^{-1} L\)

6. Return \(\hat{T} = (\hat{U}_3^T O)^{-1} \hat{R}_1\)

### 2.4.5 Complexity bound

The sample complexity of the algorithm depends on the specific concentration properties of \(\{y_1, y_2, y_3\}\) and the dependence is abstracted in the following:

**Condition 2.** There exists positive scalars \(N_0, C_{3,1}, C_{3,2}, C_{3,1,2}\) and a function \(f(N, \delta)\) (decreasing in \(N\) and \(\delta\)) such that for any \(N \geq N_0\) and \(\delta \in (0, 1)\),

1. \(P\left[\|\hat{P}_{a,b} - P_{a,b}\|_2 \leq C_{a,b} \cdot f(N, \delta)\right] \geq 1 - \delta\) for \(\{a, b\} \in \\{\{3,1\}, \{3,2\}\}\),

2. \(\forall \eta \in \mathbb{R}^d, P\left[\|\hat{P}_{3,1,2}(\eta) - P_{3,1,2}(\eta)\|_2 \leq C_{3,1,2} \cdot \|\eta\|_2 \cdot f(N, \delta)\right] \geq 1 - \delta\).

For discrete HMM, Condition 2 holds with \(N_0 = C_{3,1} = C_{3,2} = C_{3,1,2} = 1\), and \(f(N, \delta) = (1 + \sqrt{\ln(1/\delta)})/\sqrt{N}\).

**Theorem 1.** There exists a constant \(C > 0\) such that the following holds. Assume the three-view mixture model satisfies Condition 1 and Condition 2. Pick any \(\epsilon \in (0, 1)\) and \(\delta \in (0, \delta_0)\). Assume \(\Theta \in \mathbb{R}^{k \times k}\) is an independent random rotation matrix distributed uniformly over the Stiefel manifold \(\{Q \in \mathbb{R}^{k \times k} : Q^T Q = I\}\). If the number of samples \(N\) satisfies \(N \geq N_0\) and

\[
f(N, \delta/k) \leq C \cdot \frac{\min_{i \neq j} \|M_{ij}\|}{1}
\]

### 2.5 Viterbi algorithm

The Viterbi algorithm (VA) was proposed in 1967 (Viterbi)\cite{32} as a method of estimating parameters within convolutions. In its most general form, the VA
is viewed as a solution to the problem of finding parameters that maximize a posterior probability distribution of a finite-state discrete-time Markov process (Forney 1973)[19]. In the HMM setting, we seek to find a sequence of hidden states that maximize the conditional probability \( P(\eta|y) \), that is also called global decoding. VA is an efficient programming algorithm to determine the most likely sequence of latent states. Rabiner and Juang (1986)[13] introduced this method in an HMM decoding problem. The following method is rewritten from their method.

We define

\[
\xi_{1i} = P(\eta_1 = i, Y_1 = y_1) \quad i \text{ in } 1, 2, ..., k
\]

and for \( t = 2, 3, ..., T \),

\[
\xi_{ti} = \max_{\eta_1, \eta_2, ..., \eta_{t-1}} P(\eta_1:(t-1), \eta_t = i, Y_T = y_T) \quad i \text{ in } 1, 2, ..., k
\]

Thus, the recursive process of VA can be generated

\[
\xi_{tj} = \max_i (\xi_{t-1,i}\gamma_{ij})o_j(y_t)
\]

for \( t = 1, 2, ..., T \) and \( i = 1, 2, ..., k \).

The required maximizing sequence of states \( i_1, i_2, ..., i_T \) can then be determined recursively from

\[
i_T = \arg\max_{i=1,2,...,k} \xi_{Ti}
\]

and for \( t = 1, 2, ..., T - 1 \), from

\[
i_t = \arg\max_{i=1,2,...,k} (\xi_{ti}\gamma_{it+1})
\]

It can be easily found that the VA is similar in implementation to the forward-backward calculation; while it uses a maximization over previous instead of the summing process.

### 2.6 Latent class analysis

Latent class analysis is a statistical technique for the analysis of multivariate categorical data, which is used to identify the latent (unobserved) classes underlying the observed variables. The definition of latent class analysis was first
introduced by Lazarsfeld (1968)[33] using the name ”latent structure analysis”. Our method was based on the application of R package ”poLCA” (Linzer and Lewis 2011)[34]. We introduced the following based on their work in the R package.

Suppose we observe four categorical measurements A, B, C, D, each of which contains $m_i$ ($i = 1, 2, 3, 4$) possible outcomes, for individuals $s = 1, 2, ..., N$. For simplicity, $m_i$ is considered to be binary outcomes among all measurements, denoted as $m = 0$ or $m = 1$. Denote $y_{sij}$ as the observed value such that $y_{si} = 1$ if respondent $s$ gives the positive response to the $i$th variable, and $y_{si} = 0$ otherwise, where $i = 1, 2, 3, 4$.

The likelihood of an individual’s measurement $y_{si}$ with respect to his unobserved class indicator $\eta_{si}$ is denoted as $P(y_{si}|\eta_s = j)$, where $j = 1, 2, ..., d$. We also define $\pi_j = P(\eta_s = j)$, the mixing proportions that provide the weights in latent classes. The probability density function across all classes is the weighted sum,

$$P(y_s|\pi, \eta_s) = \sum_{j=1}^{d} \pi_j \prod_{i=1}^{4} P(y_{si}|\eta_s = j)^{y_{si}}$$

The R package poLCA estimates the latent class model by maximizing the log-likelihood function,

$$\ln L = \sum_{s=1}^{N} \ln \left( \sum_{j=1}^{d} \pi_j \prod_{i=1}^{4} P(y_{si}|\eta_s = j)^{y_{si}} \right)$$

with respect to $\pi_j$ and $P(y_{si}|\eta_s = j)$, using the EM algorithm. The EM algorithm is applicable for most finite mixture model because the unknown class can be treated as missing data.

The iterative process is shown below: (modified from Linzer and Lewis 2011)[34]

- **E step:** The posterior probability that each individual belongs to each class can be calculated using Bayes’ formula:

$$\hat{P}(\eta_s = j|y_s) = \frac{\hat{\pi}_j \prod_{i=1}^{4} \hat{P}(y_{si}|\eta_s = j)^{y_{si}}}{\sum_{j=1}^{d} \hat{\pi}_j \prod_{i=1}^{4} \hat{P}(y_{si}|\eta_s = j)^{y_{si}}}$$

22
• **M step:** Update the parameter estimates by maximizing the log-likelihood function given these posterior,

\[
\hat{\pi}_j = \frac{1}{N} \sum_{s=1}^{N} \hat{P}(\eta_s = j|y_s)
\]

and,

\[
\hat{P}(y_{si}|\eta_s) = \frac{\sum_{s=1}^{N} y_{si} \hat{P}(\eta_s = j|y_s)}{\sum_{s=1}^{N} \hat{P}(\eta_s = j|y_s)}
\]

The algorithm repeats the process until the overall log-likelihood reaches a limiting point.

### 2.7 Summary

In this chapter, we introduced three computational methods for solving latent variable problems. We detailed some of the calculations for each algorithm. We mentioned some of their advantages and disadvantages. In the next chapters, we apply each algorithm to simulated and real ARI data to better understand their respective utilities.
3 Simulation

This chapter will present and discuss a small simulation study to compare the methods described in Chapter 2 with the rule-based definition of ARI used by most epidemiologists. The first section will introduce the cases from which the data are simulated. We consider data like in the motivating ARI example but with a range of underlying rates, sensitivities and specificities. For these cases, we first compare the HMM estimation procedures including the EM, MCMC and spectral algorithm. In Section 3.3, we compare the ability of the rule-based method, LCA and Viterbi algorithms to predict the hidden states.

3.1 Simulation data description

The algorithms were evaluated using simulated data generated from three different hidden Markov models. In order to mimic our motivating ARI application, each of the three models have a hidden state dimension of $K = 2$ and each state has 4 possible measurements: A, B, C, D, each with two possible outcomes: 0 or 1. With this setting, the number of possible outcomes is $2^4 = 16$. In the simulation, we assume that each observation is conditionally independent of the others given a hidden state at time point $t$.

The number of measurements in each data set (72,000 = 180 x 100 x 4) introduces an extreme computational burden when using the bootstrap in conjunction with the EM algorithm, as well as when running 5-10,000 iterations of the MCMC algorithm. Thus, only 10 replicate data sets were used in this study. A small number of replicates can identify gross bias or computational limitations. Future work with more replicates can be designed to precisely estimate and compare the bias and variance of each method.

The true positive rate (TPR) $\theta_i$ and false positive rate (FPR) $\psi_i$ for each measurement are defined as $\theta_i = P(y_i = 1|\eta = 1)$ and $\psi_i = P(y_i = 1|\eta = 0)$, where $i = 1, 2, 3, 4$. Given $\theta$, $\psi$ and latent variable $\eta$, the probability distribution for
the observations is given by

\[ P(y|\theta, \psi, \eta) = \prod_{i=1}^{4} \theta_{i}^{y_i} \psi_{i}^{1-y_i} \]

### 3.2 Hidden Markov approaches comparison

Observations are generated at \( T = 180 \) scheduled time points with equal intervals for each of \( N = 100 \) subjects by the parameters \( \lambda = (\pi, \Gamma, \theta, \psi) \) where

\[
\pi = \begin{pmatrix} 0.9 \\ 0.1 \end{pmatrix} \quad \Gamma = \begin{pmatrix} 0.95 & 0.05 \\ 0.45 & 0.55 \end{pmatrix}
\]

We consider three scenarios with \( \theta, \psi \) set as

- **Scenario 1**: \( \theta = (0.2, 0.5, 0.7, 0.5), \psi = (0.01, 0.01, 0.005, 0.005) \)
- **Scenario 2**: \( \theta = (0.5, 0.5, 0.7, 0.5), \psi = (0.01, 0.01, 0.005, 0.005) \)
- **Scenario 3**: \( \theta = (0.8, 0.5, 0.7, 0.5), \psi = (0.01, 0.01, 0.005, 0.005) \)

The EM algorithm, MCMC and spectral approach (Method of Moments, MoM) were used to estimate the model parameters. The EM code was modified from the R package HMM (Himmelmann 2010)[35] to fit a multiple-sample data. The MCMC was constructed through JAGS code and operated using the R package *rjags*. Our code for the spectral approach was written in R and was based on the algorithm by Anandkumar et al. (2012)[18]. All simulations were implemented using the platform RStudio 1.0.136 on a personal computer platform. The R programs are available at the authors Github site https://github.com/wendylin23.

For EM and MoM, the 95% confidence interval was calculated using the bootstrap method (Efron 1986)[36]. The MCMC method gives a posterior distribution of parameters directly. A 95% credible interval is used for the Bayes analogue of a confidence interval.

The estimated value and bias for the transition probabilities \( P(\eta_t = 1|\eta_{t-1} = 0) \) and \( P(\eta_t = 1|\eta_{t-1} = 1) \) are shown below (Figure 1)
Figure 1: Comparison of the accuracy of estimating $P(\eta_t = 1|\eta_{t-1} = 0)$ (top row) and $P(\eta_t = 1|\eta_{t-1} = 1)$ (bottom row) of EM algorithm, MCMC and Method of Moments for each of the three measurement error scenarios (columns 1-3).

From the plots, it is evident that MCMC produces the most accurate estimation of parameters and confidence regions that covers the true value. Compared with the EM algorithm, MoM has a wider confidence interval and larger bias. However, compared with the other two methods, MoM runs much faster. The difference in running time is more than three orders of magnitude higher for the likelihood-based methods as compared to the method of moments. Considering a large sample size of our ARI data with over 3000 objects included, we will rely upon on MoM in the next chapter’s analysis.

Figure 2 shows how the bias and variance of the MoM change as the number of subjects increases from 100 to 500. As shown in Figure 1, the statistical properties of MoM is inferior to the likelihood methods for smaller sample size. Its bias is somewhat reduced at the larger sample size (Figure 2).

3.3 Predicted hidden states comparison

In this section, we will perform experiments to test the accuracy of the predicted latent states. Prediction methods include the rule-based approach (RB), latent class analysis (LCA) and HMM-based Viterbi algorithm (VA). They were
Figure 2: Comparison of the accuracy of estimating $P(\eta_t = 1|\eta_{t-1} = 0)$ and $P(\eta_t = 1|\eta_{t-1} = 1)$ using the Method of Moments with increasing sample size $N = 100, 300, 500$

compared in terms of their ability to estimate the incidence, prevalence, false negative rate (FNR) and false positive rate (FPR). The rule-based method was to say the ARI state was 1 (infected) if any of the symptoms were present and 0 if none were. Here, we define incidence rate as the probability that there is a transition from latent state equal to 0 at time $t - 1$ to latent state equal to 1 at time $t$.

The same model parameters and data configurations are used in this section as in the previous one; we use $T = 180$ observations per subject and $N = 500$ subjects. The summarized results are presented in Table 1

Table 1 shows that the rule-based method (RB) can not accurately estimate either the prevalence or incidence except in situations where the measurement error is negligible, as in Scenario 1. The bias in the RB can be very large, as evidenced by both the prevalence and incidence in Scenario 2 and the prevalence in Scenario 3. Also, the rule-based method can have poor estimation of the underlying disease state as shown in Scenarios 2 and 3 where the false positive rates are very high. This is the reason for the over-estimation of the prevalence in these cases.
### Table 1: Prediction results and errors of Scenarios 1, 2, 3 using rule-based approach, LCA and HMM-VA, compared with simulated true values.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Scenario 1</th>
<th></th>
<th>Scenario 2</th>
<th></th>
<th>Scenario 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RB</td>
<td>LCA</td>
<td>HMM-VA</td>
<td>True Values</td>
<td></td>
</tr>
<tr>
<td>IR</td>
<td>0.069</td>
<td>0.003</td>
<td>0.039</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>Prevalence</td>
<td>0.121</td>
<td>0.074</td>
<td>0.085</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>(\hat{\eta})-FNR</td>
<td>0.006</td>
<td>0.011</td>
<td>0.010</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>(\hat{\eta})-FPR</td>
<td>0.031</td>
<td>0.003</td>
<td>0.003</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>IR</td>
<td>0.020</td>
<td>0.007</td>
<td>0.034</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>Prevalence</td>
<td>0.337</td>
<td>0.077</td>
<td>0.076</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>(\hat{\eta})-FNR</td>
<td>0.006</td>
<td>0.099</td>
<td>0.091</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>(\hat{\eta})-FPR</td>
<td>0.2243</td>
<td>0.006</td>
<td>0.005</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>IR</td>
<td>0.040</td>
<td>0.012</td>
<td>0.041</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>Prevalence</td>
<td>0.529</td>
<td>0.073</td>
<td>0.080</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>(\hat{\eta})-FNR</td>
<td>0.006</td>
<td>0.099</td>
<td>0.093</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>(\hat{\eta})-FPR</td>
<td>0.435</td>
<td>0.011</td>
<td>0.009</td>
<td>0.00</td>
<td></td>
</tr>
</tbody>
</table>

*IR=\(\hat{P}(\hat{\eta}_t = 1|\hat{\eta}_{t-1} = 0)\), Prevalence=\(\hat{P}(\hat{\eta}_t = 1)\)  
*FNR=\(\hat{P}(y_t = 0|\hat{\eta}_t = 1)\), FPR=\(\hat{P}(y_t = 1|\hat{\eta}_t = 0)\)

The LCA method ignores the time dependence. Even so, it is competitive with the HMM in estimating the unknown true state \(\eta\). However, it is poorer at estimating the incidence and prevalence. Its FNR and FPR is consistently closed to 0. This is the effect of the measurement error diluting the capacity of the method to infer transitions.

Compared with the other two approaches, VA performs best. It is able to
accurately estimate all four characteristics with small systematic bias.

### 3.4 Simulation Summary

We conducted a series of small simulation pilot studies to compare competing algorithms for estimating latent variable models. EM and MCMC are two general methods for HMM analysis. Each of them gives approximately unbiased estimates and confidence intervals. However, for large sample sizes like the Nepal ARI dataset, the running speed of these two approaches is too slow for them to be useful. MoM, on the other hand, while not as accurate as EM or MCMC, has fast convergence speed making it the only feasible choice for our motivating problem. The Viterbi algorithm is also available for estimating the $\eta$ values. We compare it to the rule-based and LCA methods with the ARI data in the next section.
4 Nepal respiratory infection data

The purpose of this chapter is to illustrate the use of hidden Markov models (HMM) with the motivating Nepal infants respiratory infection symptoms data and to compare HMM-based estimates of ARI prevalence and incidence with rule-based and latent class analysis estimates.

To meet the objectives of this thesis while avoiding overlap with past or future substantive papers, we use a subset of the original data from which the scientific team studied multiple aspects of the etiology of acute respiratory infection in this population. See Tielsch, et al (2015)[4] and references therein for details on the full data set and scientific findings.

4.1 Data description

This analysis is of 180 contiguous days of respiratory symptoms data for a total of 3171 infants observed during the period August 17, 2012 to April 30, 2014. The median duration of surveillance was 157 days (IQR 56-181). Over 310,000 daily observations of respiratory symptoms were collected in this study.

The rule-based definition of ARI we use is that ARI is present if one or more of the 4 symptoms is present: fever, cough, wheezing, or difficulty breathing. There are $2^4 = 16$ possible symptoms combinations. The most common symptom report is no symptoms, which occurred in 96.74% of infant-days. Among days with one or more symptoms, the five most common combinations are cough alone (0.67%), fever alone (0.5%), wheezing and difficulty breathing (0.32%), fever and cough (0.3%) and cough with wheezing and difficulty breathing (0.027%). The frequencies for all of the 16 symptom reports are given in Table 2.

Among the four symptoms, cough has the highest frequency (1.96%) followed by wheezing (1.44%), fever (1.35%) and difficulty breathing (1.24%). In the following analysis, we consider the six most frequent combinations of symptoms, grouping the remainder in an other category. With this approach, we
<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Indicator</th>
<th>Percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>0 0 0 0</td>
<td>96.47</td>
</tr>
<tr>
<td>Fever</td>
<td>1 0 0 0</td>
<td>0.50</td>
</tr>
<tr>
<td>Cough</td>
<td>0 1 0 0</td>
<td>0.67</td>
</tr>
<tr>
<td>Fever/Cough</td>
<td>1 1 0 0</td>
<td>0.30</td>
</tr>
<tr>
<td>Wheezing</td>
<td>0 0 1 0</td>
<td>0.15</td>
</tr>
<tr>
<td>Fever/Wheezing</td>
<td>1 0 1 0</td>
<td>0.06</td>
</tr>
<tr>
<td>Cough/Wheezing</td>
<td>0 1 1 0</td>
<td>0.22</td>
</tr>
<tr>
<td>Fever/Cough/Wheezing</td>
<td>1 1 1 0</td>
<td>0.12</td>
</tr>
<tr>
<td>Diffbreath</td>
<td>0 0 0 1</td>
<td>0.17</td>
</tr>
<tr>
<td>Fever/Diffbreath</td>
<td>1 0 0 1</td>
<td>0.04</td>
</tr>
<tr>
<td>Cough/Diffbreath</td>
<td>0 1 0 1</td>
<td>0.11</td>
</tr>
<tr>
<td>Fever/Cough/Diffbreath</td>
<td>1 1 0 1</td>
<td>0.04</td>
</tr>
<tr>
<td>Wheezing/Diffbreath</td>
<td>0 0 1 1</td>
<td>0.32</td>
</tr>
<tr>
<td>Fever/Wheezing/Diffbreath</td>
<td>1 0 1 1</td>
<td>0.06</td>
</tr>
<tr>
<td>Cough/Wheezing/Diffbreath</td>
<td>0 1 1 1</td>
<td>0.27</td>
</tr>
<tr>
<td>Fever/Cough/Wheezing/Diffbreath</td>
<td>1 1 1 1</td>
<td>0.24</td>
</tr>
</tbody>
</table>

Table 2: Frequencies of the 16 possible combinations of symptoms in the Nepal respiratory infection data

A child with no symptoms on one day, they are estimated to have a 99.15% chance of no symptoms on the following day. A child with some respiratory symptom on one day has between 50 – 70% chance of retaining the same symptom, with most of the remaining probability assigned to returning to

To look at the time patterns, we have calculated and present below the empirical transition matrix for the six top categories plus other. (Table 3).
<table>
<thead>
<tr>
<th>(t-1)/(t)</th>
<th>0000</th>
<th>1000</th>
<th>0100</th>
<th>1100</th>
<th>0011</th>
<th>0111</th>
<th>others</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0000</td>
<td>0.9915</td>
<td>0.0016</td>
<td>0.0021</td>
<td>0.0006</td>
<td>0.0008</td>
<td>0.0004</td>
<td>0.0029</td>
<td>1.00</td>
</tr>
<tr>
<td>1000</td>
<td>0.2486</td>
<td>0.6410</td>
<td>0.0019</td>
<td>0.0000</td>
<td>0.0488</td>
<td>0.0148</td>
<td>0.0449</td>
<td>1.00</td>
</tr>
<tr>
<td>0100</td>
<td>0.3760</td>
<td>0.0077</td>
<td>0.5413</td>
<td>0.0013</td>
<td>0.0340</td>
<td>0.0019</td>
<td>0.0378</td>
<td>1.00</td>
</tr>
<tr>
<td>1100</td>
<td>0.1909</td>
<td>0.0020</td>
<td>0.0000</td>
<td>0.7040</td>
<td>0.0000</td>
<td>0.0283</td>
<td>0.0747</td>
<td>1.00</td>
</tr>
<tr>
<td>0011</td>
<td>0.1941</td>
<td>0.1635</td>
<td>0.0369</td>
<td>0.0011</td>
<td>0.5622</td>
<td>0.0011</td>
<td>0.0411</td>
<td>1.00</td>
</tr>
<tr>
<td>0111</td>
<td>0.2205</td>
<td>0.0210</td>
<td>0.0012</td>
<td>0.0292</td>
<td>0.0000</td>
<td>0.6313</td>
<td>0.0968</td>
<td>1.00</td>
</tr>
<tr>
<td>others</td>
<td>0.2239</td>
<td>0.0216</td>
<td>0.0069</td>
<td>0.0219</td>
<td>0.0055</td>
<td>0.0362</td>
<td>0.6840</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 3: Empirical symptoms combination transition ($\hat{P}(y_t|y_{t-1})$) matrix

the no-symptom category. There is no obvious progression from one symptom to another apparent in the table.

A logistic regression was performed to describe the prevalence of each observed ARI symptom and of “any symptoms” by fitting a model with a natural spline function of time to produce smoothed curves as shown in Figure 3. ARI has

![Figure 3: Observed ARI prevalence in a calendar time order from August 2012 to April 2014.](image)

peaks between September and October plus February and March over the one and a half years. Among the 4 respiratory symptoms, cough is most seasonable and best tracks the pattern observed in the any symptom rule. There is also a decrease in prevalence over time for each symptom type that might reflect changes in the rate of reporting or decreasing secular trends in illness incidence.
4.2 Model fitting

In the context of this comparison, our application of the rule-based approach assumes that with one or more positive respiratory symptoms on day $t$, the latent state is $\eta_t = 1$ representing that the child had an ARI. The rules used in practice often require multiple days with symptoms. However, in our latent variable models, the observed ARI was considered as an imperfect indicator of true status and the data will be analyzed by LCA and HMM. We assume the latent variable models have a hidden state dimension $k = 2$, indicating that with each observed ARI at time point $t$, a child could be either health ($\eta_t = 0$) or sick ($\eta_t = 1$).

The LCA model is constructed with the assumption that observable respiratory symptoms are conditionally independent with each other given the latent state at that time point. Unlike HMMs, LCA also assumes that data from different days are independent; LCA does not recognize the notion of an ARI episode across multiple days. The HMM we use is a first-order model that acknowledges the time dependence of ARI by letting the probability of ARI on one day to depend on whether or not there was an ARI the prior day. Given the size of the data set, only the HMM spectral method could be used.

Table 4 compares how well the LCA and HMM using the spectral method (HMM-Spectral) approximate the observed rates of the most common symptom patterns or “phenotypes”. LCA is much better than HMM-Spectral in replicating the observed respiratory symptom combination. That LCA does well is not surprising since these frequencies do not involve the time dependence parameters that are ignored by LCA. The fact that HMM-Spectral does not replicate the marginal frequencies is a notable weakness.

Figure 4 shows that all four symptoms have near perfect specificity and that cough and wheeze have the highest sensitivities. The fever symptom is estimated to have sensitivity less than 40%. We conclude that if an infant is truly
infected with ARI, coughing and wheezing are the best indicators.

<table>
<thead>
<tr>
<th>Phenotype Indicator</th>
<th>Observed Percentage (%)</th>
<th>LCA Estimation (%)</th>
<th>HMM Estimation (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0000</td>
<td>96.743</td>
<td>96.740</td>
<td>98.523</td>
</tr>
<tr>
<td>1000</td>
<td>0.496</td>
<td>0.500</td>
<td>0.008</td>
</tr>
<tr>
<td>0100</td>
<td>0.665</td>
<td>0.670</td>
<td>0.036</td>
</tr>
<tr>
<td>1100</td>
<td>0.302</td>
<td>0.121</td>
<td>0.006</td>
</tr>
<tr>
<td>0011</td>
<td>0.316</td>
<td>0.182</td>
<td>0.019</td>
</tr>
<tr>
<td>0111</td>
<td>0.272</td>
<td>0.267</td>
<td>0.030</td>
</tr>
<tr>
<td>others</td>
<td>1.200</td>
<td>1.574</td>
<td>1.378</td>
</tr>
</tbody>
</table>

*Phenotype indicator is shown in Figure 3

Table 4: LCA and HMM estimated percentage of observed phenotypes compared with observed values.

![Figure 4: Estimated sensitivity (P(y = 1|\eta = 1)) and specificity (P(y = 0|\eta = 0)) from LCA for each symptom respectively.](image)

We then compared the latent state transition probability, prevalence and measurement errors from rule-based approach, LCA and spectral HMM (Table 5). The estimated prevalence of HMM-Spectral is the smallest among the three methods, followed by LCA estimates and the rule-based value. The rule-based
method includes both the true and false positives so is by its nature biased upward. A comparison of the LCA (or HMM) prevalence estimates for \( \eta_t \) is roughly one third lower than the biased ruled-based method. This amount is exactly what is expected given the error rates for \( y_t \) estimated from the LCA as presented in Table 5. The LCA should be able to estimate the marginal \( Pr(\eta = 1) \), so its value is likely reasonable. The HMM-Spectral estimate shows a very similar result, giving confidence in this implementation of the HMM.

The rule-based approach shows a probability of 0.7541 of staying ill in the following day for an infant who is ill the day before. LCA assumes that the transition probability is equal to the marginal one. The HMM-Spectral method estimates the incidence rate to be 0.0020\((se = 0.00008)\) and the probability of continuing an episode for another day to be 0.636\((se = .005)\). Only the HMM estimates are appropriate given the assumptions of the other two methods.

Measurement errors calculated from LCA and HMM are similar, but HMM shows a slightly smaller error rate of respiratory symptoms.

With estimated model parameters from HMM, we were able to predict the latent state distribution for a large population (Figure 5). The estimated prevalence of ARI from HMM has the same trend as that of rule-based method. However, the value of prevalence is smaller considering imperfect measurement of respiratory symptoms.

This section shows an inconsistency between the HMM-Spectral and LCA for those parameters, such as the marginal prevalence, that for which they should give similar values. Therefore, we conclude that caution is needed before or when applying spectral HMM on real data, considering possible gap between its theory and practice. Zhao and Poupart recently raised similar concerns (2014)[37].
<table>
<thead>
<tr>
<th>Variables</th>
<th>Rule-based</th>
<th>LCA (0.003)</th>
<th>HMM-Spectral (0.001)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{P}(\eta_t = 1)$</td>
<td>0.0326</td>
<td>0.0251</td>
<td>0.0243</td>
</tr>
<tr>
<td>$\hat{P}(\eta_t = 1</td>
<td>\eta_{t-1} = 0)$</td>
<td>0.0085</td>
<td>0.0251</td>
</tr>
<tr>
<td>$\hat{P}(\eta_t = 1</td>
<td>\eta_{t-1} = 1)$</td>
<td>0.7541</td>
<td>0.0251</td>
</tr>
<tr>
<td>$\hat{P}(y_t = 1</td>
<td>\eta_t = 0)$</td>
<td>0.00</td>
<td>0.0091 (0.00012)</td>
</tr>
<tr>
<td>$\hat{P}(y_t = 1</td>
<td>\eta_t = 1)$</td>
<td>1.00</td>
<td>0.9457 (0.0002)</td>
</tr>
</tbody>
</table>

* $\eta_t = 1$ represents having any symptoms at that day $t$  
* (0.003) is the standard deviation for each estimate

Table 5: Comparison of estimated prevalence, transition probability from health to sick, transition probability from sick to sick, false positive rate and true positive rate for rule-based approach, LCA and HMM.

![Graph showing predicted and observed trend](image)

Figure 5: Prevalence in calendar time of HMM predicted trend and observed trend.

## 5 Conclusions and Discussion

In this thesis, we compared the rule-based approaches with latent variable models, including latent class analysis (LCA) and hidden Markov models (HMMs), for simulated data and Nepal respiratory infection data. The rule-based method has the advantage of convenient application in data analysis. However, it is hard
to propose a standard for the rule-based approach. Prevalence and incidence estimates using rules fail to acknowledge the imperfect sensitivities and specificities in the measures. Nor do they tend to acknowledge that the different symptoms have different levels of measurement error as shown in the prior section. Both LCA and HMM address these limitations.

For HMMs, we compared the performance of the EM algorithm (Rabiner and Juang 1986)[13], Bayesian inference (Scott 2002)[15] and Spectral HMM (Hsu et al. 2012[17], Anandkumar, et al. 2012[18]) in R within a pilot simulation study. Bias and confidence interval of parameters in the transition matrix were estimated for each method. We showed that the EM algorithm and Bayesian inference have small estimation bias. However, even in this small study, the spectral approach clearly had a relatively larger bias and wider confidence intervals. On the other hand, we also found that only the spectral method had computation times that would make possible its application for the motivating example.

We also compared the performance in estimating characteristics of the latent process of the rule-based approach, LCA and HMM-Viterbi algorithm under three simulation setting with different measurement errors. The rule-based approach was highly inaccurate when there was substantial measurement error, especially a non-trivial rate of false positives. The LCA and HMM-Viterbi algorithm gave similar estimates of the marginal latent variable distributions. However, only the HMM approach could correctly estimate the dependence of the latent variable across time.

The rule-based approach, LCA and spectral HMM were implemented on the Nepal respiratory infection data. We found that the respiratory symptoms had high specificity but moderate sensitivities as estimated by the LCA method. Both the prevalence of being ill and the transition probability of getting ill was overestimated by the rule-based approach.
The Hidden Markov model (HMM) is a useful tool for time series applications with possible latents states. EM and Bayesian algorithms are general approaches for fitting HMM models that give roughly unbiased estimates and valid confidence intervals. However, current software makes their application to large studies problematic because of their long running time on desk-top computers or even moderate-sized clusters. Spectral HMM emerged only recently and is praised for its computational efficiency in discrete HMM application.[17] However, most research of spectral approaches are based on theoretical arguments; there are limited simulation studies or applications. Zhao and Poupart (2014)[37] recently remarked that HMM-Spectral can perform poorly in specific cases, for example, negative probabilities are an important problem as they occur frequently. Spectral methods are consistent as the sample size goes to infinity, however, their performance in a variety of finite sample studies still needs further investigation. In the Nepal application, the HMM gave apparently unbiased estimates of the prevalence of the latent state because it matched the estimates from LCA. Whether its estimates of incidence are similarly unbiased can not be assessed from these data.
Bibliography


Date of birth: August 29, 1992
Location of birth: Wenzhou, Zhejiang, China

Education

Johns Hopkins University, Baltimore, MD
Sc.M. in Biostatistics
09/2015—present

Zhejiang University, Hangzhou, China
B.S. in Information and Computing Science
09/2011—06/2015

Professional Experience

Johns Hopkins Bloomberg School of Public Health, Baltimore, MD
Research Assistant
Nepal Maternal Influenza Analysis with Hidden Markov Model
Advisors: Scott L. Zeger Ph.D. & Joanne Katz, Ph.D.
07/2016—present

• Applied MCMC and EM algorithms to construct Hidden Markov Model designed for time-series etiology analysis.
• Studied uncertainty and indentifiability of parameters based on Nepal respiratory symptoms time series data.
• Predicted temporal endemic trend of infants’ influenza in Nepal.

International Vaccine Access Center (IVAC), Baltimore, MD
Research Assistant
Pneumonia Etiology Research for Child Health
Supervisor: Maria Deloria-Knoll, Ph.D.
02/2016—present

• Implemented integrated Bayesian approach to estimate etiologic distribution from case-control study.
• Improved the efficiency of R package while conducting series of simulations and real-data analysis.
• Contributed to manuscript submission by reporting methods and results.

School of Mathematical Science, Research Assistant
Mathematical Modeling of Retinal Image Segmentation
Advisors: Kewei Liang, Ph.D. & Feng Han, Ph.D.
09/2014—05/2015

• Denoised retinal images with modified Gaussian blur algorithm.
• Refined Level Set Methods to segment blood vessels and delivered MATLAB realizations.
• Presented final results at departmental seminar and wrote undergraduate thesis.

Publication

**Teaching Experience**

<table>
<thead>
<tr>
<th>Course</th>
<th>Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>PH.140.621/2 Statistical Methods in Public Health I/II,</td>
<td>Fall 2016</td>
</tr>
<tr>
<td><em>Teaching Assistant</em></td>
<td></td>
</tr>
</tbody>
</table>

**Awards**

<table>
<thead>
<tr>
<th>Award</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Honorable Mention of Undergraduate Thesis</td>
<td>2015</td>
</tr>
<tr>
<td>Third Class Excellence Scholarship (Top 10%)</td>
<td>2013 &amp; 2014</td>
</tr>
</tbody>
</table>

**Skills**

- Proficient in R, MATLAB, Stata, SQL, LaTeX, C,C++,C#, Java Script