Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: http://www.tandfonline.com/loi/lssp20

# Intermittent Missing Observations in Discrete-Time Hidden Markov Models

Hung-Wen Yeh , Wenyaw Chan & Elaine Symanski

**To cite this article:** Hung-Wen Yeh , Wenyaw Chan & Elaine Symanski (2012) Intermittent Missing Observations in Discrete-Time Hidden Markov Models, Communications in Statistics - Simulation and Computation, 41:2, 167-181, DOI: 10.1080/03610918.2011.581778

**To link to this article:** http://dx.doi.org/10.1080/03610918.2011.581778

Published online: 07 Oct 2011.

*Communications in Statistics—Simulation and Computation*®, 41: 167–181, 2012
Copyright © Taylor & Francis Group, LLC
ISSN: 0361-0918 print/1532-4141 online
DOI: 10.1080/03610918.2011.581778

# Intermittent Missing Observations in Discrete-Time Hidden Markov Models

HUNG-WEN YEH[1], WENYAW CHAN[2], AND ELAINE SYMANSKI[3]

[1]Department of Biostatistics, University of Kansas Medical Center, Kansas City, Kansas, USA
[2]Division of Biostatistics, The University of Texas School of Public Health, Houston, Texas, USA
[3]Division of Epidemiology and Disease Control, The University of Texas School of Public Health, Houston, Texas, USA

*In medical and public health research, hidden Markov models (HMM) are applied in longitudinal studies to model the progression of disease based on clinical classifications that may not be accurate. While missing data are common in longitudinal studies, their impact on HMM has not been well studied. We conduct a simulation study to evaluate effects on the parameter estimates of HMM by simulating complete data, along with incomplete data with intermittent missing values generated by ignorable and non ignorable missing mechanisms. Three scenarios with different sets of parameters were simulated. For incomplete data due to an ignorable mechanism, the accuracy and precision of parameter estimates are generally similar to those obtained from complete data in all examined parameter sets. Under the non ignorable mechanism, the estimation bias is substantial for most parameters when the latent outcome is equally likely to stay at the current state or to move to other states. The bias is dramatically smaller when subjects are more likely to stay at the current state than moving to other states. An example from the mental health arena is used to illustrate the application of intermittent missing observations using HMM. Some computational issues are also discussed.*

**Keywords** Hidden Markov models; Imperfect indicator; Latent variable; Longitudinal multinomial data.

**Mathematics Subject Classification** 62M05.

## 1. Introduction

In medical and public health research, patients' health status or responses to treatment are often categorized into several states and recorded repeatedly through

Received June 21, 2010; Accepted April 11, 2011

Address correspondence to Hung-Wen Yeh, Department of Biostatistics, University of Kansas Medical Center, Mail Stop 1026, 5028K Robinson, 3901 Rainbow Boulevard, Kansas City, Kansas 66160, USA; E-mail: hyeh@kumc.edu

time, and the progression of disease or treatment effects can be studied by Markov chain models (e.g., see Salazar, 2007). However, considering the errors that occur in evaluating patient health status, researchers have taken into account imperfect diagnosis (i.e., misclassification of an outcome) using hidden Markov models (HMM) (Scott et al., 2005). In HMM, the true disease status is regarded as latent and not observable, and the observed diagnosis is treated as an imperfect indicator of true disease status. It is the latent disease status whose transition is assumed to follow Markov chains. The observed classification or diagnosis is assumed to depend only on the true disease state at the same time point (and not on the history of previously observed states). With longitudinal observations of the imperfect indicators of disease status, HMM simultaneously estimates (1) the initial distribution of disease status at baseline, (2) the transition probabilities of the latent disease status, and (3) the probabilities of misclassification (Rabiner, 1989). Because of the flexibility that allows for modeling various types of unobservable data and the ability to incorporate additional information about outcome misclassification, HMM has been adopted rapidly in health-effects research (Le Strat and Carrar, 1999; Bureau et al., 2003; Jackson et al., 2003; Altman and Petkau, 2005; Scott et al., 2005; Mass et al., 2006).

Missing values are commonly inevitable in longitudinal studies and should be handled properly to alleviate bias. For the traditional or observed Markov models (OMM) in which the observed disease status is regarded as the true status, the impact of missing observations has been well studied and corresponding methods to deal with effects due to missing data have been proposed (for example, see Albert, 2000; Bee, 2005; Deltour et al., 1999; Yeh et al., 2010). Yet, to our knowledge, there is a paucity of research that has investigated the accuracy and precision of HMM estimates when the imperfect indicator variable is completely observed at each scheduled time point. Not to mention the incomplete data where the observations on the imperfect indicator variable are intermittently missing. Therefore, we conduct a simulation study to examine the impact of intermittent missing observations in HMM. A discrete-time HMM and some computational issues are described in Sec. 2. In Sec. 3, we simulate ignorable and nonignorable intermittent missing values on the imperfect indicator and model the incomplete data as ignorable. A Schizophrenia study is used as an example in Sec. 4. Discussion is given in Sec. 5. Simulation and analysis of the Schizophrenia data are performed in SAS® version 9.2 using author-developed codes based on the Baum-Welch algorithm (Baum et al., 1970; Welch, 2003), tantamount to the well-known Expectation-Maximization (EM) algorithm (Dempster et al., 1977).

## 2. Methods

### 2.1. *The Model*

In this study, we consider a discrete-time HMM with a categorical imperfect indicator variable. Let $\boldsymbol{Y} = \{Y_1, Y_2, \ldots, Y_T\}$ denote the longitudinal outcomes observed at $T$ time points and serving as the imperfect indicator for the latent variables $\boldsymbol{S} = \{S_1, S_2, \ldots, S_T\}$ where both $Y_t$ and $S_t (t = 1, 2, \ldots, T)$ can take any one of $c$ possible states. Suppose the latent outcomes follow a discrete-time Markov chain with a $c \times c$ transition probability matrix $\boldsymbol{A} = \{a_{ij} : a_{ij} = \Pr(S_{t+1} = j \mid S_t = i), t = 1, 2, \ldots, T-1; i, j = 1, 2, \ldots, c\}$. Assume that the distribution of $Y_t$ depends on $S_t, t = 1, 2, \ldots, T$, and let $\boldsymbol{B} = \{b_{ij} : b_{ij} = \Pr(Y_t = j \mid S_t = i), t =$

$1, 2, \ldots, T; i, j = 1, 2, \ldots, c\}$ represent the misclassification probability matrix. Also, let $\boldsymbol{\pi} = \{\pi_1, \pi_2, \ldots, \pi_c\}$ denote the vector that contains the initial probabilities in true states $1, 2, \ldots, c$, i.e., $\pi_i = Pr(S_1 = i)$. Clearly, $\sum_{i=1}^{c} \pi_i = 1$. Similarly, by the property of the transition probability matrix $\sum_{j=1}^{c} a_{ij} = 1$, we can also have $\sum_{j=1}^{c} b_{ij} = 1$ because a given hidden state $i$ can be classified into one and only one of the $c$ states. It is not the observation $\boldsymbol{Y}$ but the latent variable $\boldsymbol{S}$ that is assumed to follow a Markov chain. In HMM, we often assume observations are conditionally independent given the hidden states (e.g., Eqs. (13a) and (13b) in Rabiner, 1989), so the likelihood of observing the imperfect indicator outcomes $\boldsymbol{Y} = \boldsymbol{y}$ conditioned on the parameters $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$ is

$$\Pr(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{\theta}) = \sum_{\boldsymbol{S}} \Pr(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{S} \mid \boldsymbol{\theta}) = \sum_{\boldsymbol{S}} \Pr(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{S\theta}) \Pr(\boldsymbol{S} \mid \boldsymbol{\theta}),$$

where $\Pr(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{S}, \boldsymbol{\theta}) = b_{S_1 y_1} b_{S_2 y_2} \cdots b_{S_T y_T}$ and

$$\Pr(\boldsymbol{S} \mid \boldsymbol{\theta}) = \pi_{S_1} a_{S_1 S_2} a_{S_2 S_3} \cdots a_{S_{T-1} S_T}.$$

Thus,

$$\Pr(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{\theta}) = \sum_{S_1, S_2, \cdots, S_T} \pi_{S_1} (a_{S_1 S_2} a_{S_2 S_3} \cdots a_{S_{T-1} S_T})(b_{S_1 y_1} b_{S_2 y_2} \cdots b_{S_T y_T}). \tag{1}$$

### 2.2. *The EM Algorithm*

We use the likelihood approach to analyze the HMM as suggested by Rabiner (1989). Due to the latent nature of the observations, the EM algorithm (Dempster et al., 1977) is applied. In this section, we first construct the complete-data log-likelihood, followed by the E- and the M-steps. We also discuss how the missing imperfect indicator observations are handled and how the initial values for the parameters are selected.

*2.2.1. The E- and the M-steps.* Consider a general case of multiple subjects ($N$) who may have different numbers of observations on the imperfect indicator variable and assume that these observations are independent between any two subjects. Under the HMM, the complete-data log-likelihood can be written as

$$l(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{s}) = \sum_{k=1}^{N} \left[ \ln(\pi_{s_{k,1}}) + \sum_{t=1}^{T_k - 1} \ln(a_{s_{k,t} s_{k,t+1}}) + \sum_{t=1}^{T_k} \ln(b_{s_{k,t} y_{k,t}}) \right], \tag{2}$$

where $\boldsymbol{y}$ and $\boldsymbol{s}$, respectively, denote the imperfect indicator observations and the hidden states of the $N$ subjects, $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$ denote the vector of the parameters of interest, and $T_k$ denotes the duration or the time of the last observation for the $k$th subject. In the E-step, the expectation of the complete-data log-likelihood conditioned on the observed outcomes and the $v$th iteration of the parameter estimates can be expressed as

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(v)}) = E_{\boldsymbol{S}}[\ln\Pr(\boldsymbol{y}, \boldsymbol{S} \mid \boldsymbol{\theta}) \mid \boldsymbol{y}; \boldsymbol{\theta}^{(v)}]$$

$$= \sum_{k=1}^{N}\sum_{i=1}^{c}[\ln(\pi_i) \times \gamma_{k,1}^{(v)}(i)] + \sum_{k=1}^{N}\sum_{t=1}^{T_k-1}\sum_{i=1}^{c}\sum_{j=1}^{c}[\ln(a_{ij}) \times \xi_{k,t}^{(v)}(i,j)]$$

$$+ \sum_{k=1}^{N}\sum_{t=1}^{T_k}\sum_{i=1}^{c}\sum_{j=1}^{c}[\ln(b_{ij}) \times \gamma_{k,t}^{(v)}(i) \times \delta(y_{k,t} = j)], \qquad (3)$$

where $\gamma_{k,t}^{(v)}(i) = Pr(S_{k,t} = i \mid \boldsymbol{Y_k} = \boldsymbol{y_k} = \boldsymbol{y_k}; \boldsymbol{\theta}^{(v)})$ indicates for the $k$th subject, the probability of the true outcome in state $i$ at time $t$ given the observed imperfect indicator outcomes and the $v$th iteration of the parameters; $\xi_{k,t}^{(v)}(i,j) = Pr(S_{k,t} = i, S_{k,t+1} = j \mid \boldsymbol{Y_k} = \boldsymbol{y_k}; \boldsymbol{\theta}^{(v)})$ refers to the probability that the true outcome of the $k$th subject is in state $i$ at time $t$ and in state $j$ at time $t+1$, respectively, given the imperfect indicator observations and the $v$th iteration of the parameters; $\delta(y_{k,t} = j)$ is the index function that equals 1 if the $k$th subject at time $t$ is observed to be in state $j$, and zero otherwise.

In the M-step, we maximize the $Q$ function with respect to each parameter. Because of the constraint $\sum_{i=1}^{c}\pi_i = \sum_{j=1}^{c}a_{ij} = \sum_{j=1}^{c}b_{ij} = 1$, one can apply the Lagrange multiplier with Eq. (3) and set the first derivative with respect to each parameter to zero and then obtain

$$\widehat{\pi}_i^{(v+1)} = \frac{\sum_{k=1}^{N}\gamma_{k,1}^{(v)}(i)}{\sum_{k=1}^{N}\sum_{i=1}^{c}\gamma_{k,1}^{(v)}(i)}, \quad i = 1, 2, \ldots, c \qquad (4)$$

$$\hat{a}_{ij}^{(v+1)} = \frac{\sum_{k=1}^{N}\sum_{t=1}^{T_k-1}\xi_{k,t}^{(v)}(i,j)}{\sum_{k=1}^{N}\sum_{t=1}^{T_k-1}\gamma_{k,t}^{(v)}(i)}, \quad i, j = 1, 2, \ldots, c \qquad (5)$$

where $\gamma_{k,t}^{(v)}(i) = \sum_{j=1}^{c}\xi_{k,t}^{(v)}(i,j)$, and

$$\hat{b}_{ij}^{(v+1)} = \frac{\sum_{k=1}^{N}\sum_{t=1}^{T_k}\gamma_{k,t}^{(v)}(i) \times \delta(y_{k,t} = j)}{\sum_{k=1}^{N}\sum_{t=1}^{T_k}\gamma_{k,t}^{(v)}(i)}, \quad i, j = 1, 2, \ldots, c. \qquad (6)$$

Corresponding formulas for a single sequence or subject (i.e., $N = 1$) with complete data are given in Eqs. (40a)–(40c) in Rabiner (1989).

Note that the M-step is not in a closed form. In order to calculate $\gamma_{k,t}^{(v)}(i)$ and $\xi_{k,t}^{(v)}(i,j)$, the Baum-Welch algorithm introduces two additional auxiliary forward and backward variables (Rabiner, 1989):

$$\alpha_{k,t}(i) = \Pr(Y_{k,1} = y_{k,1}, \ldots, Y_{k,t} = y_{k,t}, S_{k,t} = i \mid \boldsymbol{\theta}), \quad t = 1, 2, \ldots, T_k$$

$$\beta_{k,t}(i) = \Pr(Y_{k,t+1} = y_{k,t+1}, \ldots, Y_{k,T} = y_{k,T} \mid S_{k,t} = i, \boldsymbol{\theta}), \quad t = 1, 2, \ldots, T_k - 1,$$

where the forward variable $\alpha_{k,t}(i)$ is the probability of observing the imperfect indicator values up to time $t(Y_{k,1}, \ldots, Y_{k,t})$ and the true outcome being in state $i$ $(S_{k,t} = i)$ for the $k$th subject given the parameter values $\boldsymbol{\theta}$; the backward variable $\beta_{k,t}(i)$ is the probability of observing the imperfect indicator values from time $t+1$ to time $T_k$ given the parameter values and that the true outcome is in state $i$ for the $k$th subject.

It can be shown that

$$\alpha_{k,1}(i) = \pi_i b_{i,y_{k,1}} \quad \text{and} \quad \alpha_{k,t+1}(i) = \left[\sum_i \alpha_{k,t}(i) a_{ij}\right] b_{j,y_{k,t+1}}. \tag{7}$$

Let $\beta_{k,T_k}(i) = Pr(\cdot \mid S_{T_k} = i, \boldsymbol{\theta}) \equiv 1$,

$$\beta_{k,t}(i) = \sum_j a_{ij} b_{j,y_{k,t+1}} \beta_{k,t+1}(j) \tag{8}$$

With these auxiliary variables, $\gamma_{k,t}^{(v)}(i)$ and $\xi_{k,t}^{(v)}(i, j)$ can be expressed as

$$\gamma_{k,t}^{(v)}(i) = \frac{\alpha_{k,t}(i)\beta_{k,t}(i)}{\sum_i \alpha_{k,T_k}(i)} \tag{9}$$

$$\xi_{k,t}^{(v)}(i, j) = \frac{\alpha_{k,t}(i) a_{ij} b_{j,y_{k,t+1}} \beta_{k,t+1}(j)}{\sum_i \alpha_{k,T_k}(i)}. \tag{10}$$

In applying the Baum–Welch algorithm, we begin with a set of initial values for the parameters $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$ to calculate the auxiliary variables by Eqs. (7) and (8). These variables are then used to compute $\gamma_{k,t}^{(v)}(i)$ and $\xi_{k,t}^{(v)}(i, j)$ using Eqs. (9) and (10), which are then substituted into Eqs. (4)–(6) to obtain the parameters $\boldsymbol{\theta}$ and to evaluate $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(v)})$. These parameter values are then again used to update the auxiliary variables. The process continues until the $Q$ function converges.

*2.2.2. The Missing Imperfect Indicator Observations.* In HMM, because the parameters directly related to the observed indicator outcomes are the $b_{ij}$'s or the elements of the misclassification matrix **B** but not any element of $\pi$ or **A**, we adopt Liu's suggestion (Liu, 1997) by defining $b_{S_t y_t} = 1$ if the imperfect indicator observation is missing for the $k$th subject at time $t$ ($y_{k,t} = .$). In this way, $\ln(b_{S_t y_t}) = 0$ does not change Eq. (1), i.e., the incomplete data are modeled as ignorable. Correspondingly, Eq. (6) is modified as

$$\hat{b}_{ij}^{(v+1)} = \frac{\sum_{k=1}^{N} \sum_{t=1}^{T_k} \gamma_{k,t}^{(v)}(i) \times \delta(y_{k,t} = j)}{\sum_{k=1}^{N} \sum_{t=1}^{T_k} \gamma_{k,t}^{(v)}(i) \times \delta(y_{k,t} \neq .)} \tag{11}$$

to satisfy the condition $\sum_{j=1}^{c} \hat{b}_{ij} = 1$ for any $i = 1, 2, \ldots, c$.

*2.2.3. Selection of the Initial Values.* It is known that the EM algorithm does not guarantee convergence to global maxima, and that the obtained estimates may be sensitive to the initial values. Rabiner (1989) proposed selecting initial values arbitrarily or choosing values that were equally likely for the parameters in $\pi$ and **A.** Liu (1997) further suggested avoiding the extreme and the independent patterns for the parameters in **B** that can be described as:

(a) the extreme pattern: for a given $i$, $b_{ij} = 1$ and $b_{ij'} = 0$ for any $j \neq j'$; and
(b) the independent pattern: $b_{ij} = \Pr(y = j \mid S = i) = \Pr(y = j) = b_j^*$.

Rabiner's and Liu's suggestions are adopted in simulation and analysis (see Secs. 3.1 and 4 below).

## 3. A Simulation Study

In this section, we present a simulation study to investigate how the intermittently missing imperfect indicator observations affect parameter estimation. Complete HMM data are generated for three parameter sets and, for each set, missing values are simulated by two mechanisms. Parameters are then estimated by the procedures described in Sec. 2 for both the complete and the two incomplete data sets. Hence, in total, nine scenarios are considered.

### 3.1. *Simulation Procedure*

Observations are generated at 10 scheduled time points with equal intervals for each of 50 subjects by the parameters $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$ where

$$\boldsymbol{\pi} = \begin{pmatrix} 0.1 \\ 0.3 \\ 0.6 \end{pmatrix}, \quad \boldsymbol{B} = \begin{pmatrix} 0.8 & 0.15 & 0.05 \\ 0.1 & 0.7 & 0.2 \\ 0.05 & 0.05 & 0.9 \end{pmatrix}.$$

We consider three scenarios for the transition matrix $\boldsymbol{A}$:

$$\boldsymbol{A}_{\text{Lo}} = \begin{pmatrix} 0.34 & 0.33 & 0.33 \\ 0.33 & 0.34 & 0.33 \\ 0.33 & 0.33 & 0.34 \end{pmatrix}, \quad \boldsymbol{A}_{\text{Mo}} = \begin{pmatrix} 0.6 & 0.2 & 0.2 \\ 0.2 & 0.6 & 0.2 \\ 0.2 & 0.2 & 0.6 \end{pmatrix}, \quad \boldsymbol{A}_{\text{Hi}} = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{pmatrix}.$$

Because the $\boldsymbol{A}$ matrix provides the transition probabilities at each time point, the $\boldsymbol{A}_{\text{Hi}}$ parameter matrix specifies a relatively high probability that successive observations recur. This implies high dependency of a future outcome on the current outcome. Similarly, the $\boldsymbol{A}_{\text{Lo}}$ and the $\boldsymbol{A}_{\text{Mo}}$ matrices specify low and moderate dependency.

In simulating missing values, we consider both the ignorable and the non ignorable mechanisms. Let $\boldsymbol{Y} = (\boldsymbol{Y}^{obs}, \boldsymbol{Y}^{mis})$ be a $(N \times T)$ matrix for complete data of the imperfect indicator variable for $N$ subjects each with $T$ observations, where $\boldsymbol{Y}^{obs}$ and $\boldsymbol{Y}^{mis}$ denote the observed and the missing values, respectively, and $\boldsymbol{R}$ be a matrix with elements $R_{ij} = 1$ if $Y_{ij}$ is observed and 0 otherwise. The complete data are composed of observed values and an indication of missingness and the joint distribution of $\boldsymbol{Y}$ and $\boldsymbol{R}$ can be factored into $Pr(\boldsymbol{Y}, \boldsymbol{R} \mid \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{Y} \mid \boldsymbol{\theta}) f(\boldsymbol{R} \mid \boldsymbol{Y}, \boldsymbol{\psi})$. Then the observed-data likelihood is proportional to the joint distribution integrated over the missing values:

$$\mathscr{L}(\boldsymbol{\theta}, \boldsymbol{\psi} \mid \boldsymbol{Y}^{obs}, \boldsymbol{R}) \propto \int f(\boldsymbol{Y}^{obs}, \boldsymbol{Y}^{mis} \mid \boldsymbol{\theta}) f(\boldsymbol{R} \mid \boldsymbol{Y}^{obs}, \boldsymbol{Y}^{mis}, \boldsymbol{\psi}) d\boldsymbol{Y}^{mis}.$$

Also, the likelihood ignoring the missing mechanism is $\mathscr{L}(\boldsymbol{\theta} \mid \boldsymbol{Y}^{obs}) \propto f(\boldsymbol{Y}^{obs} \mid \boldsymbol{\theta})$. According to Little and Rubin (2002), the inference for $\boldsymbol{\theta}$ based on $\mathscr{L}(\boldsymbol{\theta} \mid \boldsymbol{Y}^{obs})$ is the same as that based on $\mathscr{L}(\boldsymbol{\theta}, \boldsymbol{\psi} \mid \boldsymbol{Y}^{obs}, \boldsymbol{R})$ only when the following two criteria are satisfied: (a) $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are separable, i.e., the joint parameter space of $(\boldsymbol{\theta}, \boldsymbol{\psi})$ is the product of individual parameter space; and (b) the missing values are missing at random (MAR), defined as $f(\boldsymbol{R} \mid \boldsymbol{Y}^{obs}, \boldsymbol{Y}^{mis}, \boldsymbol{\psi}) = f(\boldsymbol{R} \mid \boldsymbol{Y}^{obs}, \boldsymbol{\psi})$. When the chance of observing the outcome does not depend on either the observed or missing data, i.e., $f(\boldsymbol{R} \mid \boldsymbol{Y}^{obs}, \boldsymbol{Y}^{mis}, \boldsymbol{\psi}) = f(\boldsymbol{R} \mid \boldsymbol{\psi})$, the mechanism is missing completely at random (MCAR), a special case of MAR. Violation of either criterion leads to a non ignorable missing mechanism, including not missing at random (NMAR) if $\boldsymbol{R}$ depends on $\boldsymbol{Y}^{mis}$.

Here, we do not consider any covariates and simply regard the observable transition history as the "observed" values; thus, we consider an ignorable missing mechanism in which the occurrence of missing observations depends on the immediately previous "observed" state:

(1) Ignorable: for each subject, we simulate missing values at time $t$ with 60% probability if the observation at time $t-1$ is in state 3, i.e., $p_3 = Pr(r_t = 0 \mid Y_{t-1} = 3) = 0.6$ and $p_i = Pr(r_t = 0 \mid Y_{t-1} = i) = 0$ for $i = 1, 2$ and $t = 2, 3, \ldots, 10$, where the parameter $\boldsymbol{\psi} = (p_1, p_2, p_3) = (0,\ 0,\ 0.6)$ is chosen independently of $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$ thus, $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$ are separable.

We also consider a non ignorable mechanism that violates the criterion of MAR:

(2) Non ignorable: for each subject, we generate missing values at time $t$ with 60% probability only if the observation at time $t$ is in state 3, i.e., $p_3 = Pr(r_t = 0 \mid Y_t = 3) = 0.6$ and $p_i = Pr(r_t = 0 \mid Y_t = i) = 0$ for $i = 1, 2$ and $t = 1, 2, \ldots, 10$.

In this way, we generate, on average, 19–23% missing values. For each combination of parameter sets and missing mechanisms, we perform 1,000 simulation runs, each replication includes 50 subjects and each subject has 10 observations. We use the probabilities obtained from the transition of the observed imperfect indicators as the initial values for parameters in the matrix $\boldsymbol{A}$, and arbitrarily choose

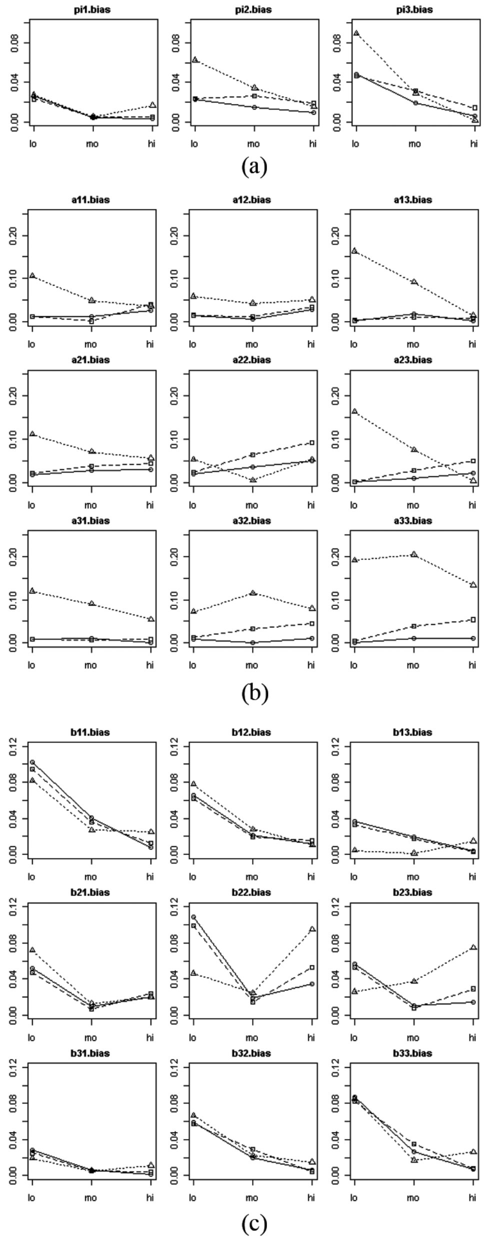
$$\boldsymbol{\pi}^{(0)} = \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix}, \quad \boldsymbol{B}^{(0)} = \begin{pmatrix} 0.7 & 0.2 & 0.1 \\ 0.2 & 0.5 & 0.3 \\ 0.1 & 0.1 & 0.8 \end{pmatrix}$$

as the initial values for the initial distribution vector and the misclassification matrix. The estimates at the 10th iteration are used because the log-likelihood often converges with 10 or fewer iterations under this simulation design.

### 3.2. *Simulation Results*

Figure 1(a) compares the estimation bias of the initial distribution parameters $\boldsymbol{\pi}$ across the three dependency levels specified in $\boldsymbol{A}_{\mathrm{Lo}}$, $\boldsymbol{A}_{\mathrm{Mo}}$, and $\boldsymbol{A}_{\mathrm{Hi}}$ among HMM fitted to (1) complete data, (2) incomplete data due to ignorable missing mechanism, and (3) incomplete data due to non ignorable missing mechanism, respectively. Figures 1(b) and (c) present respective comparisons for the parameters of the transition and misclassification matrices. When data are complete, the bias is small ($<0.02$) for most parameter estimates in the initial distribution vector $\boldsymbol{\pi}$ and the transition matrix $\boldsymbol{A}$, but is large (0.06 to 0.12) for most estimates of the misclassification matrix $\boldsymbol{B}$. The figures show a trend of decreasing bias with the dependency level ($\boldsymbol{A}_{\mathrm{Lo}}$, $\boldsymbol{A}_{\mathrm{Mo}}$, and $\boldsymbol{A}_{\mathrm{Hi}}$) in several $\boldsymbol{\pi}$ and $\boldsymbol{B}$ parameters, but this trend seems to be in the opposite direction, although less obvious, in the $\boldsymbol{A}$ parameters.

When missing data are due to the ignorable mechanism described in Sec. 3.1, the MLEs demonstrate similar bias-vs.-dependency patterns to those observed with complete data. The bias is, in general, of similar magnitude but can be slightly larger for some parameters ($\pi_2, \pi_3,\ a_{ij}, i, j = 2, 3$, and $b_{22}, b_{23}$).



—○— complete data; ---□--- incomplete data due to ignorable missing; ···Δ··· incomplete data due to non-ignorable missing. The labels of "lo", "mo", and "hi" on horizontal axis refer to the three transition parameter sets: $a_{ii} = 0.34$ and $a_{ij} = 0.33$ for "lo", $a_{ii} = 0.6$ and $a_{ij} = 0.2$ for "mo", $a_{ii} = 0.8$ and $a_{ij} = 0.1$ for "hi".

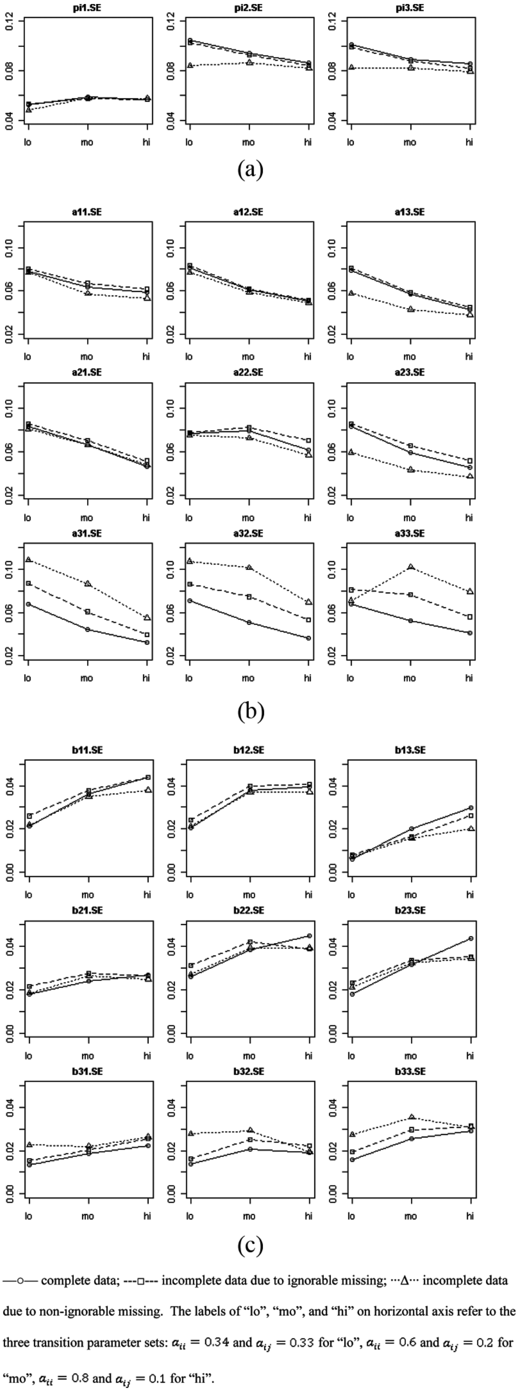
**Figure 1.** Bias of HMM estimates of (a) initial distribution (b) transition probabilities (c) misclassification probabilities vs. the dependency levels specified in $\boldsymbol{A}_{Lo}$, $\boldsymbol{A}_{Mo}$, and $\boldsymbol{A}_{Hi}$ in complete and incomplete data due to ignorable and non ignorable missing mechanism.

When the missing mechanism is non ignorable, the bias is substantially larger in most $\boldsymbol{\pi}$ and $\boldsymbol{A}$ parameters as compared to the bias of the corresponding estimates obtained from complete data or the incomplete data due to the ignorable missing mechanism that we adopt. This phenomenon is most obvious for parameters associated with transitions from and/or to state 2 or 3 ($a_{13}, a_{23}, a_{31}, a_{32}$, and $a_{33}$). The bias appears to be lower when the hidden outcome is more likely to stay at the current state (i.e., greater dependency in the matrix $\boldsymbol{A}$). The missing mechanism does not impact the bias of the estimates in the misclassification $\boldsymbol{B}$ matrix as much as in the $\boldsymbol{\pi}$ and $\boldsymbol{A}$ parameters. Also, the bias is of similar magnitude and depicts similar patterns across complete data and missing mechanisms except for the parameters of $b_{22}$ and $b_{23}$.

Figures 2 (a), (b), and (c) compare the standard errors (SE) of the parameter estimates versus various levels of dependency of the future outcome on the current outcome ($\boldsymbol{A}_{\mathrm{Lo}}$, $\boldsymbol{A}_{\mathrm{Mo}}$, and $\boldsymbol{A}_{\mathrm{Hi}}$) for complete and incomplete data. Intuitively and empirically, incomplete data lead to greater SE due to a loss of information. In this simulation, only some parameters preserve the expected pattern: the transition parameters from state 3 ($a_{31}, a_{32}$, and $a_{33}$) and the misclassification parameters with hidden state 3 ($b_{31}, b_{32}$, and $b_{33}$). Because the missingness is simulated on state 3 or depends on a previous observation with outcome in state 3 (see, Sec. 3.1), it is not surprising to find more variability in the estimates associated with state 3. Other parameter estimates from the simulation with incomplete data show SE in similar magnitude to, or slightly less than, those obtained from the analyses with complete data. Also, for complete and incomplete data, if the hidden status is more likely to remain at the current state than to move to other states (i.e., $\boldsymbol{A}_{\mathrm{Mo}}$ or $\boldsymbol{A}_{\mathrm{Hi}}$), SE decrease in the transition parameter estimates but increase in the misclassification parameter estimates. The mean squared errors, in general, show similar patterns to those obtained for the bias (results not shown).

## 4. An Example

Data from a Schizophrenia study (SCHIZX1.sas7bdat), which are publicly available (http://tigger.uic.edu/hedeker/ml.html), were analyzed. In this study, 437 patients were randomized to receive either placebo (108 patients) or drug (329 patients) treatment and observed at baseline and at equal follow-up intervals of 1 week for weeks 1–6. The severity of Schizophrenia was categorized as normal, mild, moderate, and severe. The initial observed health status for the two groups combined was 59% severe, 28% moderate, 12% mild, and less than 1% normal. More than 95% observations were missing at weeks 2, 4, and 5, and about 15% and 23% of the observations were missing at weeks 3 and 6; overall, the proportion of missing values was about 54%. The missing mechanism was assumed to be ignorable because the majority of missing data occurred at weeks 2, 4, and 5, which was probably due to the study design. The summary counts for observed baseline health status and transitions for each group are summarized in Table 5 of Yeh et al. (2010). Here, we consider the observed health status as an imperfect indicator of true status and analyze the data by HMM. The initial values for the baseline distribution and the



(a)

(b)

(c)

—○— complete data; ---□--- incomplete data due to ignorable missing; ···Δ··· incomplete data due to non-ignorable missing. The labels of "lo", "mo", and "hi" on horizontal axis refer to the three transition parameter sets: $a_{ii} = 0.34$ and $a_{ij} = 0.33$ for "lo", $a_{ii} = 0.6$ and $a_{ij} = 0.2$ for "mo", $a_{ii} = 0.8$ and $a_{ij} = 0.1$ for "hi".

**Figure 2.** Standard error of HMM estimates of (a) initial distribution (b) transition probabilities (c) misclassification probabilities vs. the dependency levels specified in $\boldsymbol{A}_{Lo}$, $\boldsymbol{A}_{Mo}$, and $\boldsymbol{A}_{Hi}$ in complete and incomplete data due to ignorable and non ignorable missing mechanism.

misclassification parameters were arbitrarily chosen as:

$$\boldsymbol{\pi}^{(0)} = \begin{pmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{pmatrix}, \quad \boldsymbol{B}^{(0)} = \begin{pmatrix} 0.850 & 0.100 & 0.030 & 0.02 \\ 0.125 & 0.700 & 0.125 & 0.05 \\ 0.050 & 0.125 & 0.700 & 0.125 \\ 0.020 & 0.050 & 0.130 & 0.800 \end{pmatrix},$$

whereas the initial values for transition parameters are estimated from transition of the observed imperfect indicator variables:

$$\boldsymbol{A}^{(0)}_{\text{Drug}} = \begin{pmatrix} 0.982 & 0.006 & 0.006 & 0.006 \\ 0.149 & 0.660 & 0.170 & 0.021 \\ 0.082 & 0.447 & 0.365 & 0.106 \\ 0.046 & 0.227 & 0.304 & 0.423 \end{pmatrix}, \quad \boldsymbol{A}^{(0)}_{\text{Plcebo}} = \begin{pmatrix} 0.250 & 0.250 & 0.250 & 0.250 \\ 0.091 & 0.636 & 0.182 & 0.091 \\ 0.001 & 0.361 & 0.472 & 0.166 \\ 0.016 & 0.078 & 0.125 & 0.781 \end{pmatrix}.$$

We terminate the algorithm at the 30th iteration according to convergence of log-likelihood.

Table 1 shows the HMM estimates for the treatment and the placebo groups, respectively. The baseline distributions ($\boldsymbol{\pi}$) indicate that the true status of most patients' Schizophrenia was moderate or severe. The next four columns (the $\boldsymbol{A}$ matrix), represent the transitions from and to the severe, moderate, mild, and normal states. In the treatment group, 63% of patients were initially in the severe state, 27% in the moderate, 10% in the mild state, and nearly no one in the normal state. Severe patients had probabilities of 29%, 12% and 4% to improve to the moderate, mild, and normal states, respectively, during a one-week period. The likelihood of improving from a moderate to the mild or normal state in a one-week period was 27% and 5%, respectively; and there was a 5% probability of regressing to the severe state. Patients transitioned from the mild to the normal state with 21% probability and from mild to the moderate state with a 2% probability. Once patients reached the normal state, they were not likely to progress to other states, implying that Schizophrenia was unlikely to recur in patients treated with the drug. The last four columns indicate the probabilities of misclassification. In the treatment group, patients who were truly severe had a 6% chance of being misclassified as moderate; patients who were actually moderate had a 15% chance of being misclassified as mild and a 4% probability of being diagnosed as severe; patients whose true state was mild had an 8% chance of being misclassified as moderate; patients in the normal state had a 78% chance of being diagnosed correctly. The placebo effects can be interpreted in a similar manner.

Generally, in comparing the unobservable true health status of patients, those individuals treated with the drug were more likely to improve and less likely to regress than patients receiving the placebo; the probabilities of correct diagnoses or misclassification were similar in the two groups except for the likelihood of a correct diagnosis for patients in the normal state, which could be due to small numbers. The variability in estimates of the HMM parameters can be estimated through bootstrapping (Efron and Tibshirani, 1994). We apply the Benjamini-Yekutieli's procedure (2001) to control for a 0.05 false discovery rate in correlated multiple tests and find significant differences in the probabilities of remaining severe and transitions from severe to the moderate and mild states.

**Table 1**
HMM estimates and 95% confidence intervals in the Schizophrenia study

| | | | Transition matrix $\boldsymbol{A}$ | | | | Misclassification matrix $\boldsymbol{B}$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group | State | Initial $\boldsymbol{\pi}$ | Severe | Moderate | Mild | Normal | Severe | Moderate | Mild | Normal |
| Drug | Severe | 0.63<br>(0.56, 0.70) | 0.56 ‡<br>(0.47, 0.65) | 0.29 ‡<br>(0.19, 0.37) | 0.12 ‡<br>(0.08, 0.20) | 0.04<br>(0.02, 0.06) | 0.93<br>(0.86, 0.98) | 0.06<br>(0.01, 0.12) | 0.01<br>(†, 0.036) | †<br>(†,†) |
| | Moderate | 0.27<br>(0.19, 0.33) | 0.05<br>(0.01, 0.11) | 0.64<br>(0.47, 0.71) | 0.27<br>(0.21, 0.42) | 0.05<br>(0.01, 0.09) | 0.04<br>(0.01, 0.09) | 0.81<br>(0.76, 0.90) | 0.15<br>(0.07, 0.21) | †<br>(†,†) |
| | Mild | 0.10<br>(0.07, 0.16) | †<br>(†, 0.004) | 0.02<br>(†, 0.12) | 0.77<br>(0.63, 0.83) | 0.21<br>(0.16, 0.29) | 0.006<br>(†, 0.033) | 0.08<br>(0.02, 0.16) | 0.92<br>(0.84, 0.97) | †<br>(†,0.003) |
| | Normal | 0.004<br>(†, 0.014) | †<br>(†,†) | †<br>(†,†) | 0.003<br>(†, 0.022) | 0.997<br>(0.978, 1) | †<br>(†,†) | †<br>(†,†) | 0.22<br>(0.09, 0.35) | 0.78<br>(0.65, 0.92) |
| Placebo | Severe | 0.58<br>(0.44, 0.69) | 0.88 ‡<br>(0.81, 0.94) | 0.09 ‡<br>(0.02, 0.15) | 0.017 ‡<br>(†, 0.066) | 0.02<br>(†, 0.08) | 0.93<br>(0.85, 0.99) | 0.05<br>(†, 0.12) | 0.02<br>(†, 0.047) | †<br>(†,†) |
| | Moderate | 0.33<br>(0.19, 0.45) | 0.09<br>(0.02, 0.17) | 0.73<br>(0.38, 0.88) | 0.18<br>(0.06, 0.54) | †<br>(†, 0.013) | 0.03<br>(†, 0.17) | 0.86<br>(0.76, 0.99) | 0.11<br>(†, 0.18) | †<br>(†,†) |
| | Mild | 0.09<br>(0.01, 0.19) | †<br>(†, 0.02) | 0.02<br>(†, 0.25) | 0.89<br>(0.49, 1.0) | 0.09<br>(†, 0.30) | †<br>(†,†) | 0.11<br>(†, 0.22) | 0.89<br>(0.72, 0.99) | †<br>(†, 0.17) |
| | Normal | †<br>(†, 0.167) | 0.08<br>(†, 0.48) | 0.003<br>(†, 0.70) | 0.23<br>(†, 0.98) | 0.69<br>(†, 0.98) | †<br>(†, 0.49) | 0.02<br>(†, 0.82) | 0.01<br>(†, 0.76) | 0.97<br>(†, 1.0) |

†Extreme small values

‡Significantly different between groups according to the Benjamini-Yekutieli's procedure

## 5. Discussion

Hidden Markov models (HMM) offer advantages in analyzing longitudinal latent disease progression or improvement in health status because they account for imperfect measurement of the disease status. In this paper, we study the intermittent missing in HMM by simulation. Generally, we found that incomplete data due to ignorable missing mechanisms lead to estimates with similar bias and SE to the complete data. In contrast, incomplete data due to ignorable mechanisms result in substantially greater bias, but the magnitude of bias reduces when the diagonal elements in the transition matrix approach 1. The SE is greater in estimates of transitions or misclassification associated with latent state subject to missing, but is similar to or slightly smaller in other estimates as compared to those obtained from complete data.

The algorithm described in Sec. 2.2 assumes the missing mechanism is ignorable. Whether this assumption is plausible depends on how accurately the observed data can predict the missing values. In a bivariate normal setting, for example, with one variable completely observed and the other only partially observed, the ignorable method may introduce substantial bias in the mean of the partially observed variable under not missing at random if the correlation between the two variables is low (Schafer, 1997, Sec. 2.5). Low correlation suggests the completely observed variable does not provide much information in predicting the missing values. However, if the correlation is strong, the missing values can be more accurately predicted by the observed data and the bias can be dramatically less. In HMM, the true latent state a subject will move to depends only on the current state and not on the previous transition history. Therefore, when subjects are more likely to stay at the current state than moving to other states, the missing values can be predicted more accurately by the immediately prior or immediately following observation, and parameter estimation leads to less bias. Our simulation shows, for some parameters, this trend. When this is not the case, it is probably because the likelihood function is not smooth and the achieved estimates are likely local maxima or saddle points.

The ignorable method can be a reasonable approximation for incomplete data due to non ignorable mechanisms when $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are separable. In our simulation work, missing values are generated only for state 3 in the non ignorable mechanism. In other situations though, missing values are possible for any state with certain proportions. For a non ignorable mechanism $Pr(r_t = 0 \mid Y_t = i) = p_i$, the missing proportion $p_i$ varies from state to state. It is intuitive that the more homogeneous these missing proportions are, the less the missing will depend on its actual state. An extreme case is if each state has the same missing proportion, then $Pr(r_t = 0 \mid Y_t = 1) \approx Pr(r_t = 0 \mid Y_t = 2) \approx \cdots \approx Pr(r_t = 0) \approx$ constant, and the missingness is virtually completely at random, a special case of ignorable mechanisms.

When ignorability is not a plausible assumption, methods for non ignorable missing are required. Established statistical methods for non ignorable missing in models other than HMM commonly need knowledge about the missing mechanism (informative missing), and require joint modeling the data $f(\boldsymbol{Y} \mid \boldsymbol{\theta})$ and the missing mechanism $f(\boldsymbol{R} \mid \boldsymbol{Y}, \boldsymbol{\psi})$, and estimating parameters $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ simultaneously (Troxel et al., 1998; Albert, 2000). These methods usually introduce additional parameters other than $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ and make the computation more intensive. Further work is required to cope with general non ignorable missing mechanisms.

Finally, like many other iterative procedures, the convergence in likelihood in the EM algorithm does not guarantee convergence in estimates. In some simulation runs, the overall bias (defined as the sum of squared bias over all the 21 parameters) diminishes and then increases after a few iterations. When we examine the bias for individual parameters in either case, some parameters converge to the true values with increasing iterations while others diverge (results not shown). Additionally, different sets of initial values may produce estimates that vary in magnitude. Therefore, we recommend researchers applying HMM to report the initial values of parameters and the number of iterations used to achieve the final parameter estimates so that the analysis can be reproduced.

## Acknowledgments

The authors Dr. Donald Hedeker of University of Illinois at Chicago for his permission to use the Schizophrenia dataset. Appreciation also goes to Drs. John M. Davis and Nina Schooler, the original data owners. We also thank the anonymous reviewer for comments and suggestions that greatly help improve this article.

## References

Albert, P. S. (2000). A transitional model for longitudinal binary data subject to nonignorable missing data. *Biometrics* 56(2):602–608.

Altman, R. M., Petkau, A. J. (2005). Application of hidden Markov models to multiple Sclerosis lesion count data. *Statistics in Medicine* 24(15):2335–2344.

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

Bee, M. (2005). Estimating rating transition probabilities with missing data. *Statistical Methods and Applications* 14:127–141.

Benjamini, Y., Yekutieli D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annuals of Statistics* 29(4):1165–1188.

Bureau, A., Shiboski S., Hughes, J. P. (2003). Applications of continuous time hidden Markov models to the study of misclassified disease outcomes. *Statistics in Medicine* 22(3):441–462.

Deltour, I., Richardson, S., Le Hesran, J. (1999). Stochastic algorithms for Markov models estimation with intermittent missing data. *Biometrics* 55:565–573.

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

Efron, B., Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. Chapman & Hall/CRC. 1st edition.

Jackson, C. H., Sharples, L. D., Thompson, S. G., Duffy, S. W., Couto, E. (2003). Multistate Markov models for disease progression with classification error. *Statistician* 52(2):193–209.

Le Strat, Y., Carrar F. (1999). Monitoring epidemiologic surveillance data using hidden Markov models. *Statistics in Medicine* 18(24):3463–3478.

Little, R. J., Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. 2nd ed. New York: Wiley-Interscience. pp. 118–119.

Liu, L. (1997). *Hidden Markov Models for Precipitation in a Region of Atlantic Canada*. Master's Report. University of New Brunswick.

Mass, H. J., Danhof, M., Della Pasqua, O. E. (2006). Prediction of headache response in migraine treatment. *Cephalalgia* 26(4):416–22.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77(2):257–286.

Salazar, J. C., Schmitt, F. A., Yu, L., Mendiondo, M. M., Kryscio, R. J. (2007). Shared random effects analysis of multi-state Markov models: application to a longitudinal study of transitions to dementia. *Statistics in Medicine* 26(3):568–580.

Schafer, J. (1997). *Analysis of incomplete multivariate data*. Chapman and Hall.

Scott, S. L., James, G. M., Sugar, C. S. (2005). Hidden Markov models for longitudinal comparisons. *Journal of the American Statistical Association* 100(470):565–576.

Troxel, A. B., Harrington, D. P., Lipsitz, S. R. (1998). Analysis of longitudinal data with non ignorable non-monotone missing values. *Journal of the Royal Statistical Society*. Series C, 47(3):425–438.

Welch, L. R. (2003). Hidden markov models and the Baum-Welch Algorithm. *IEEE information theory society Newsletter* 53(4):1, 10–13.

Yeh, H. W., Chan, W. Y., Symanski, E., Davis, B. R. (2010). Estimating transition probabilities for ignorable intermittent missing data in a discrete-time Markov chain. *Communications in Statistics—Simulation and Computation* 39(2):433–448.