

Correcting for Misclassification for a Monotonic Disease Process with an Application in Dental Research

María José García-Zattera

Department of Statistics, PUC (Chile) & L-BioStat, KUL (Belgium)

Financed by CONICYT, Chile & KUL grant OT/05/60

Prague, ISCB 2009

Joint work with: T. Mutsvari (L-Biostat), A. Jara (UdeC), D. Declerck (KUL) and E. Lesaffre (Erasmus MC & L-Biostat)

Outline

- 1 Introduction
- 2 Simple Hidden Markov Model (HMM)
- 3 Regression Hidden Markov Model
- 4 Concluding Remarks

Outline

- 1 Introduction
- 2 Simple Hidden Markov Model (HMM)
- 3 Regression Hidden Markov Model
- 4 Concluding Remarks

Misclassification

- Often a diagnosis is a **binary process**: diseased ($Y = 1$) or healthy ($Y = 0$)
- Wrong classification \implies **misclassification**

Classification (Y^*)	True (Y)	
	0	1
0	τ_{00}	$1 - \tau_{11}$
1	$1 - \tau_{00}$	τ_{11}

- Drawbacks:
 - Prevalence and incidence estimated with error
 - The impact of risk factors may be incorrectly estimated in the presence of classification errors
- **Correction for misclassification is necessary**

Correction for Misclassification

- Cross-sectional study:
 - expert knowledge
 - **internal validation**
- Practical problems:
 - geographical spread
 - small sample size
- Methods for prevalence estimation in the presence of diagnostic error (Magder and Hughes, 1997; Mwalili et al., 2005).

Progressive Disease

- Progressive disease
 - monotonic
 - systemic lupus, AIDS, **caries experience (CE)**...
- Models for longitudinal studies to estimate (Espeland et al., 1989; Albert et al., 1997)
 - prevalence
 - incidence
 - misclassification parameters
- Properties
 - identifiability
 - efficiency

Objectives

- Caries experience



- Prevalence and incidence in presence of misclassification

Objectives

- Caries experience



- Prevalence and incidence in presence of misclassification
- Simple HMM:
 - evaluate its performance
 - compare it with some early approaches

Objectives

- Caries experience



- Prevalence and incidence in presence of misclassification
- Simple HMM:
 - evaluate its performance
 - compare it with some early approaches
- Extension to include covariates

The Signal-Tandmobiël[®] Study - 1

- Longitudinal dental study
- Flanders, Belgium: 1996-2001
- 4468 children followed from 7 until 12 years old
- Annual dental examinations
- Sixteen dental examiners
- Clinical information
- Data on oral hygiene and dietary habits

The Signal-Tandmobiel[®] Study - 2

- The **diagnosis** of CE is **difficult**
- **Three** calibration exercises for CE:
⇒ **sensitivity** and **specificity** of each examiner
- Validation data **is not a random sample**

Estimate the prevalence and incidences of CE and the evaluation of risk factors.

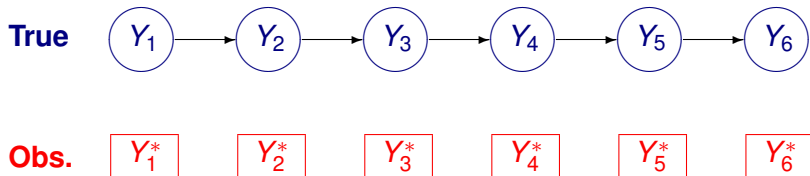
Outline

- 1 Introduction
- 2 Simple Hidden Markov Model (HMM)**
- 3 Regression Hidden Markov Model
- 4 Concluding Remarks

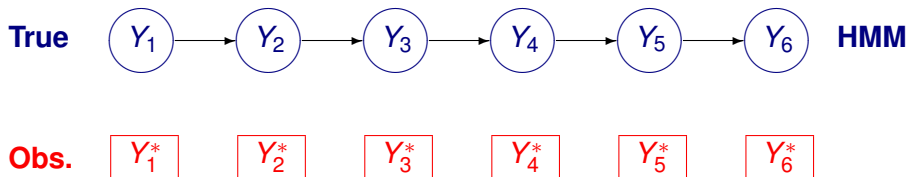
Model Representation - 1

- Suppose that each subject is examined at n time points (t_1, \dots, t_n)
- Let $\mathbf{Y}^* = (Y_1^*, \dots, Y_n^*)$ be the vector of possibly corrupted binary responses
- Let $\mathbf{Y} = (Y_1, \dots, Y_n)$ be the vector of true (unobserved) binary responses

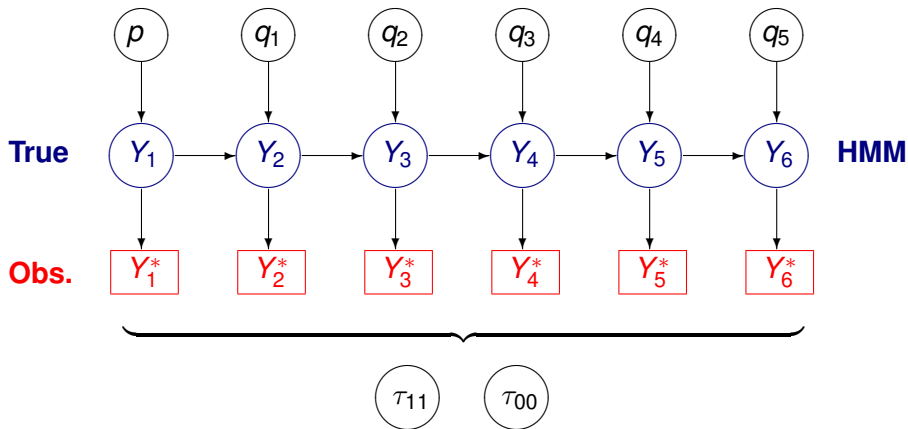
Model Representation - 2



Model Representation - 2



Model Representation - 2



Model Representation - 3

- Thus we assume that the transition matrices Q_{j-1} at time t_{j-1} are

$$Q_{j-1} = \begin{pmatrix} 1 - q_{j-1} & q_{j-1} \\ 0 & 1 \end{pmatrix}, \quad j = 2, \dots, n$$

Model Representation - 3

- Thus we assume that the transition matrices Q_{j-1} at time t_{j-1} are

$$Q_{j-1} = \begin{pmatrix} 1 - q_{j-1} & q_{j-1} \\ 0 & 1 \end{pmatrix}, \quad j = 2, \dots, n$$

\implies Monotonic disease

Early approaches

- Reversals Excluded
- Reversals Included

Early approaches

- Reversals Excluded
- Reversals Included
- Carlos-Senning (1968)
- Lu (1968)
- Poole (1973)

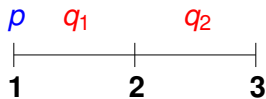
Early approaches

- Reversals Excluded
- Reversals Included
- Carlos-Senning (1968)
- Lu (1968)
- Poole (1973)

These approaches only allow the **incidence** estimation and are designed for only 2 time points.

Simulation Settings

Examinations:



Simulation Settings

Examinations:



Simulation Settings



Prevalence

0.03
0.10
0.15

Incidences

0.04
0.10
0.15

Sensitivity = 0.85, 0.90

Specificity = 0.85, 0.90

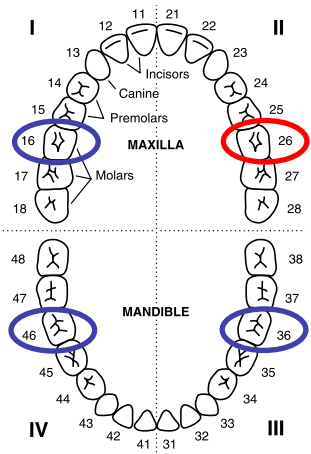
$n = 2000, 5000$

Simulation Results

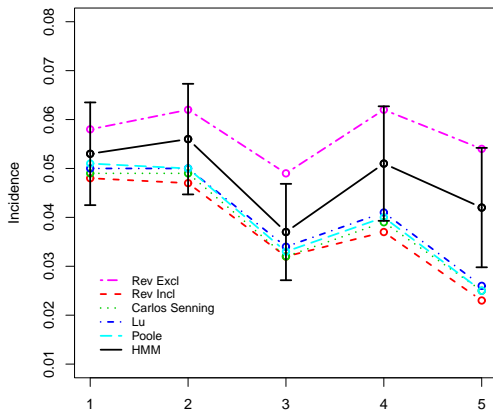
- **MLE for HMM better** with respect to bias and MSE
- Sensitivity and specificity are **well estimated**
- The bigger the sample size, the smaller the MSE

In the HMM the parameters of interest are **identified** and can be estimated using the **main data**.

ST Study: Molar 26



Molar 26



Outline

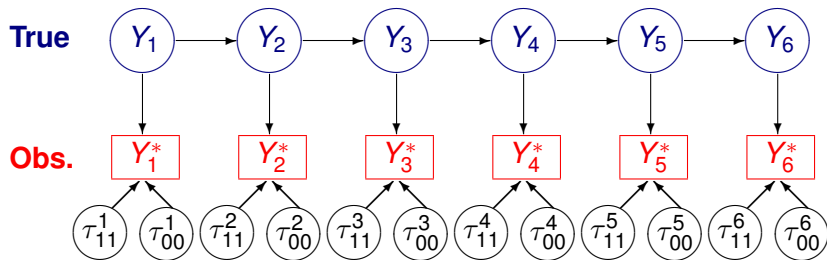
- 1 Introduction
- 2 Simple Hidden Markov Model (HMM)
- 3 Regression Hidden Markov Model**
- 4 Concluding Remarks

Motivation

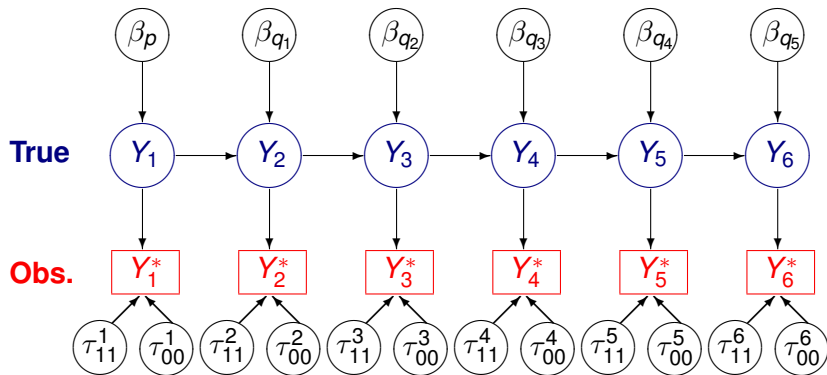
- Different characteristic at the beginning of the study
- Timing of examination was irregular
- Examiners switched over time
- Time dependent covariates

The conclusions drawn from the Simple HMM simulation study, can not be immediately extended to the application setting.

Model Representation



Model Representation



Bayesian Implementation

- Zellner's (1983) g-priors for logistic regression coefficients with $g = 2n$
- Independent $Beta(1, 1)$ prior for the examiners misclassification parameters
- Data augmentation - Metropolis within Gibbs
- MH steps based on weighted least squares normal proposals (Gamerman, 1997)
- R-package

Simulation Settings

- Same structure as the application:
 - covariates
 - sample size
- True regression coefficients from the application
- Misclassification parameters:

Examiner	Sensitivity	Specificity
Ex1 - Ex4	0.95	0.90
Ex5 - Ex8	0.90	0.95
Ex9 - Ex12	0.85	0.92
Ex13 - Ex16	0.93	0.85

- Different prior specifications

Simulation Results

- **Good estimation:**
 - Regression coefficients
 - Misclassification parameters
- **More variability** for the incidences in the latest states
- **Robust** results under different priors

Regression Results: Molar 26

- Significant covariates for the ST study

Prevalence : age

Incidence 1:

Incidence 2: days between examinations

Incidence 3:

Incidence 4: x-ordinate & age at start brushing

Incidence 5: in-between meals

Outline

- 1 Introduction
- 2 Simple Hidden Markov Model (HMM)
- 3 Regression Hidden Markov Model
- 4 Concluding Remarks**

Conclusions

- When the response is **monotone** the **misclassification parameters** can be estimated **without using validation data**
- Test to evaluate the existence of proper validation data can be performed
- The availability of real **internal validation** data could lead to **more efficient estimates** of the misclassification parameters
- HMM:
 - performs **better** than the early dental approaches
 - allows the estimation of the **prevalence, incidence, SE, SP**
 - allows the inclusion of **covariates**

Future Research

- Modeling of several teeth jointly
- Explore the connection with survival models