Time-Varying Exposure Subject to Misclassification

Bias Characterization and Adjustment

by

Eric Cormier

B.Sc. (Honors), The University of Victoria, 2009

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate Studies

(Statistics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

August 2010

Abstract

Measurement error occurs frequently in observational studies investigating the relationship between exposure variables and a clinical outcome. Error-prone observations on the explanatory variable may lead to biased estimation and loss of power in detecting the impact of an exposure variable. When the exposure variable is time-varying, the impact of misclassification is complicated and significant. This increases uncertainty in assessing the consequences of ignoring measurement error associated with observed data, and brings difficulties to adjustment for misclassification.

In this study we considered situations in which the exposure is time-varying and nondifferential misclassification occurs independently over time. We determined how misclassification biases the exposure outcome relationship through probabilistic arguments and then characterized the effect of misclassification as the model parameters vary. We show that misclassification of time-varying exposure measurements has a complicated effect when estimating the exposure-disease relationship. In particular the bias toward the null seen in the static case is not observed.

After misclassification had been characterized we developed a means to adjust for misclassification by recreating, with greatest likelihood, the exposure path of each subject. Our adjustment uses hidden Markov chain theory to quickly and efficiently reduce the number of misclassified states and reduce the effect of misclassification on estimating the disease-exposure relationship.

The method we propose makes use of only the observed misclassified exposure data and no validation data needs to be obtained. This is achieved by estimated switching probabilities and misclassification probabilities from the observed data. When these estimates are obtained the effect of misclassification can be determined through the characterization of the effect of misclassification presented previously. We can also directly adjust for misclassification by recreating the most likely exposure path using the Viterbi algorithm.

The methods developed in this dissertation allow the effect of misclassification, on estimating the exposure-disease relationship, to be determined. It accounts for misclassification by reducing the number of misclassified states and allows the exposure-disease relationship to be estimated significantly more accurately. It does this without the use of validation data and is easy to implement in existing statistical software.

Table of Contents

Abstract ii

Table of Contents iv

List of Tables vi

List of Figures vii

Acknowledgments ix

Dedication x

1 Introduction 1
1.1 Problem Formulation 2

2 Bias Determination 4
2.1 Bias Calculation 4

3 Bias Characterization 6
3.1 Bias Characterization for Linear Outcome Model 6
3.1.1 Effect of the Exposure Switching Probability 8
3.1.2 Effect of the Sensitivity and Specificity 18
3.2 Bias Characterization for Linear Outcome Model with Lag Term 25
3.2.1 Effect of the Exposure Switching Probability 26
3.2.2 Effect of the Sensitivity and Specificity 31
3.3 The Effect of the Number of Exposure Measurements 41

4 Discrete-Time Hidden Markov Adjustment 43
4.1 Discrete-Time Hidden Markov Process 44
4.2 Inference for Discrete-Time Hidden Markov Process 44
4.3 Recreating the Path Through True Exposure States 45

4.4 Adjusting for Misclassification: True Exposure Path Recreation . . . 46
4.4.1 Data Simulation . . . 47
4.4.2 Simulation Results . . . 48

5 Continuous-Time Hidden Markov Adjustment 51
5.1 Continuous-Time Markov Process . . . 52
5.2 Maximum Likelihood Estimation . . . 53
5.3 Continuous Time Hidden Markov Process . . . 53
5.4 Inference for Continuous-Time Hidden Markov Process . . . 54
5.5 Recreating the Exposure Path for Continuous Time Hidden Markov Processes . . . 55
5.6 Adjusting for Misclassification: True Exposure Path Recreation in Continuous Time . . . 56
5.6.1 Data Simulation . . . 57
5.6.2 Simulation Results . . . 58

6 Conclusion and Future Work . . . 66

Bibliography . . . 69

Appendix

A R Code for Bias Determination . . . 72

List of Tables

4.1 Simulation results for reducing misclassification using discrete adjustment . . . 48
4.2 Simulation results for discrete adjustment with no lag term . . . 48
4.3 Simulation results for discrete adjustment with misspecified no lag term . . . 49
4.4 Simulation results for discrete adjustment with lag term . . . 49
4.5 Simulation results for discrete adjustment with misspecified lag term . . . 50

5.1 Simulation results of continuous time hidden Markov parameters: Case 1 . . . 58
5.2 Simulation results for continuous recreation: Case 1 . . . 59
5.3 Simulation results for continuous time adjustment with no lag term . . . 59
5.4 Simulation results of continuous time hidden Markov parameters: Case 2 . . . 60
5.5 Simulation results for continuous recreation: Case 2 . . . 60
5.6 Simulation results for continuous time adjustment with misspecified no lag term . . . 61
5.7 Simulation results of continuous time hidden Markov parameters: Case 3 . . . 61
5.8 Simulation results for continuous recreation: Case 3 . . . 62
5.9 Simulation results for continuous time adjustment with lag term . . . 62
5.10 Simulation results of continuous time hidden Markov parameters: Case 4 . . . 63
5.11 Simulation results for continuous recreation: Case 4 . . . 64
5.12 Simulation results for continuous time adjustment with misspecified lag term . . . 64

List of Figures

3.1 Coefficient Magnitudes 7
3.2 Common Switching Probability when $(SN = 0.8, SP = 0.95)$ 9
3.3 Switching Probabilities for β_4 when $(SN = 0.8, SP = 0.95)$. 10
3.4 Switching Probabilities for β_3 when $(SN = 0.8, SP = 0.95)$. 11
3.5 Switching Probabilities for β_{34} when $(SN = 0.8, SP = 0.95)$. 12
3.6 Switching Probabilities for Bias when $(SN = 0.8, SP = 0.95)$ 13
3.7 Switching Probabilities for β_4 when $(SN = 0.95, SP = 0.8)$. 14
3.8 Switching Probabilities for β_4 when $(SN = 0.9, SP = 0.9)$ . . 15
3.9 Switching Probabilities for β_3 when $(SN = 0.95, SP = 0.8)$. 16
3.10 Switching Probabilities for β_3 when $(SN = 0.9, SP = 0.9)$ . . 17
3.11 Effect of SN and SP on Bias when $\phi = 0.2$ 18
3.12 Effect of SN and SP on Bias when $\phi = 0.5$ 19
3.13 Effect of SN and SP on Determination of β_4 when $\phi = 0.2$ . . 20
3.14 Effect of SN and SP on Determination of β_4 when $\phi = 0.8$ . . 21
3.15 Effect of SN and SP on Determination of β_3 when $\phi = 0.2$ . . 22
3.16 Effect of SN and SP on Determination of β_3 when $\phi = 0.8$ . . 23
3.17 Effect of SN and SP on Determination of β_{34} when $\phi = 0.2$. 24
3.18 Coefficient Magnitudes for Lagged Model 26
3.19 Effect of Switching Probabilities on β_4 for Lagged Model . . . 28
3.20 Effect of Switching Probabilities on β_3 for Lagged Model . . . 29
3.21 Effect of Switching Probabilities on β_2 for Lagged Model . . . 30
3.22 Effect of Switching Probabilities on Bias for Lagged Model . 31
3.23 Effect of SN and SP on Determination of β_3 for Lagged Model when $\phi = 0.2$ 32
3.24 Effect of SN and SP on Determination of β_3 for Lagged Model when $\phi = 0.5$ 33
3.25 Effect of SN and SP on Determination of β_3 for Lagged Model when $\phi = 0.8$ 34
3.26 Effect of SN and SP on Determination of β_4 for Lagged Model when $\phi = 0.2$ 35

3.27 Effect of SN and SP on Determination of β_4 for Lagged Model when $\phi = 0.5$ 36
3.28 Effect of SN and SP on Determination of β_4 for Lagged Model when $\phi = 0.8$ 37
3.29 Effect of SN and SP on Determination of Model Bias for Lagged Model when $\phi = 0.2$ 38
3.30 Effect of SN and SP on Determination of Model Bias for Lagged Model when $\phi = 0.5$ 39
3.31 Effect of SN and SP on Determination of β_2 for Lagged Model when $\phi = 0.2$ 40
3.32 Effect of the Number of Exposure Measurements on Bias, β_n and β_{n-1} 42

Acknowledgments

I would like to express my supreme gratitude to my supervisor Professor Paul Gustafson and my co-supervisor Dr. Nhu Le whose support and suggestions have been an immeasurable help throughout this entire process.

I would also like to thank Professors John Petkau, Lang Wu and Ruben Zamar for their support throughout my masters program.

I am also grateful to my fellow graduate students Corrine and Sky whose collaboration throughout my masters program has been very beneficial.

Eric Cormier

The University of British Columbia
August 2010

To My Family and Friends Who Enrich My Life Every Single Day.

Chapter 1

Introduction

In many epidemiological and clinical studies we wish to model a health related outcome, Y, dependent on an explanatory variable corresponding to some exposure status, X, and certain measured potential confounders Z. Sometimes the measured exposure status, denoted by X^*, is an imperfect surrogate for the actual exposure X. This is known as exposure misclassification and it is very important to account for in these studies. Carroll, Ruppert, Stefanski and Crainiceanu (2006) found that measurement error in the explanatory variable:

- causes bias in parameter estimation for statistical models;
- masks the features of the data;
- leads to a profound loss of power for detecting relationships between variables.

An example of this is when a prescription is dispensed to a patient but the medication is not taken. If the exposure measure is taken from the prescription records, then our data would assume that the patient was exposed to the treatment when actually no exposure occurred. This will cause our estimates to be biased and is a serious problem in many studies. Hence the goal of adjustment for mis-measurement is to achieve roughly unbiased estimates to reveal the relationship between Y and X indirectly, based on the measurements of Y, X^* and perhaps other correctly recorded covariates Z. In binary contexts the degree of misclassification is determined by the sensitivity and specificity, (SN_i, SP_i) respectfully:

$$SN_i = Pr(X^* = 1|X = 1, Y = i), \quad SP_i = Pr(X^* = 0|X = 0, Y = i), \tag{1.1}$$

for $i = 0, 1$.

In simple contexts it is reasonable to assume the actual exposure and the recorded exposure are both binary and X^* depends on (X, Y, Z) only through X, which is know as nondifferential misclassification. Then

$$SN = Pr(X^* = 1|X = 1), \qquad SP = Pr(X^* = 0|X = 0). \tag{1.2}$$

In this situation it is known that exposure misclassification biases the attenuation factor:

$$Attenuation factor = \frac{X^{*}\text{coefficient in the } (Y|X^{*}, Z) \text{ regression}}{X\text{coefficient in the } (Y|X, Z) \text{ regression}} \quad (1.3)$$

toward the null ($0 < AF < 1$) when estimating the exposure-disease association (Gustafson 2004). Therefore there is a tendency to report an artificially weak association between the exposure and response in ignoring measurement error on the exposure. This bias increases when either the sensitivity or specificity of the classification decreases or if the correlation between X and Z increases. Furthermore, when the exposure prevalence approach 0 or 1, measurement error induces serious and unstable attenuation toward the null. It is also known that in these contexts adjusting for exposure misclassification has little to no effect on the 'nearer to null' endpoint of the interval estimate for the coefficient of X^{*}. That is, misclassification adjustment will not strengthen the evidence for the existence of an exposure-disease association (Gustafson 2004). The simple contexts defined above for modeling misclassification bias are very restrictive and most medical studies do not fall into this category. Greeland and Gustafson (2006) found that no general conclusion could be made regarding the direction of estimated association when the nondifferential misclassification requirement on a binary exposure is not satisfied, or when the exposure variable is polychotomous. Further discussion of measurement in continuous or polychotomous exposure variables or when misclassification is differential is found in Gustafson (2004) and Carrol et al. (2006).

To account for measurement error in all the situations described above complete information on the outcome variable Y, true exposure X and surrogate exposure X^{*} is needed for a small proportion of the data (validation sample). The true exposure status for the majority of study subjects (main study) remains unobservable or cannot be precisely measured.

1.1 Problem Formulation

In this thesis we restrict our attention to misclassification on a time-varying binary exposure variable with no other measured covariates and we assume no measurement error arising in the outcome of interest Y. When the actual binary exposure status is time-varying all of the rules about how the bias affects our results no longer hold.

We assume that the actual binary exposure status across time $X_1, X_2, X_3, \ldots$

is a Markov chain with switching probabilities:

$$P(X_i = 1|X_{i-1} = 0) = \phi_1, \qquad P(X_i = 0|X_{i-1} = 1) = \phi_2, \tag{1.4}$$

and that these time-varying exposures are misclassified ($X_j \rightarrow X_j^*$) independently over time. This misclassification is assumed to be nondifferential and can be characterized by (SN, SP). To characterize the effect of exposure misclassification we consider a linear outcome model of the form

$$E(Y_i|X_1, \ldots, X_i) = \alpha + \psi X_i, \tag{1.5}$$

and a linear outcome model including lag terms such as

$$E(Y_i|X_1, \ldots, X_i) = \alpha + \gamma X_{i-1} + \psi X_i. \tag{1.6}$$

Even in this simple case when very strong assumptions are made about the dependence of exposure-status over time, misclassification of the time-varying exposure will have a significant effect. This effect will be different then the attenuation toward the null seen in the static case.

In this thesis we characterize and adjust for the effect of time-varying exposure misclassification and thereby increase the accuracy of the estimates and allow for correct inference to be made about the effect of time-varying exposure. This adjustment is obtained without the use of a validation study to determine the misclassification model, $(X_{1:n}^*|X_{1:n})$, and is easily implemented in statistical software.

The thesis is organized as follows. Chapter 2 provides general results on how the effect of misclassification was determined. Chapter 3 characterizes the effect of misclassification for specific examples and depicts general trends that result. Chapters 4 and 5 describe how to adjust for misclassification using Markov chain theory and display results from simulation studies. Chapter 6 provides overall conclusions and remarks on further research.

Chapter 2

Bias Determination for Time-Varying Exposure Misclassification

Let the outcome variable $\mathbf{Y}$ be dependent on exposure variable $\mathbf{X}$. If $\mathbf{Y}$ is modeled on the misclassified exposure variable $\mathbf{X}^*$ then bias results. In the simple case this bias and how misclassification affects the exposure-outcome relationship can be computed exactly.

2.1 Bias Calculation

Assume the actual binary exposure-status across time is $\mathbf{X_{1:n}} = (X_1, X_2, X_3, \ldots, X_n)$ and that these time-varying exposures are misclassified ($X_j \rightarrow X_j^*$) as $\mathbf{X^*_{1:n}} = (X_1^*, X_2^*, X_3^*, \ldots, X_n^*)$. Assume this misclassification occurs independently over time and is nondifferential. Then for a linear outcome model

$$E(Y_n|\mathbf{X_{1:n}}) = \mathbf{X_{1:n}}\mathbf{\Psi}, \tag{2.1}$$

the outcome variable $\mathbf{Y}$ depends on $\mathbf{X^*_{1:n}}$ through the relationship

$$\begin{aligned} E(Y_n|\mathbf{X^*_{1:n}}) &= E\left\{E(Y_n|\mathbf{X_{1:n}}, \mathbf{X^*_{1:n}})|\mathbf{X^*_{1:n}}\right\} \\ &= E\left\{(\mathbf{X_{1:n}}\mathbf{\Psi})|\mathbf{X^*_{1:n}}\right\} \\ &= E\left\{\mathbf{X_{1:n}}|\mathbf{X^*_{1:n}}\right\}\Psi. \end{aligned}$$

The joint distribution of $\mathbf{X}$ and $\mathbf{X}^*$ is calculated by

$$Pr(\mathbf{X_{1:n}}, \mathbf{X^*_{1:n}}) = \left[Pr(x_1)\prod_{i=2}^{n} Pr(x_i|x_1 \ldots x_{i-1})\right]\prod_{i=1}^{n} Pr(x_i^*|x_i), \tag{2.2}$$

where

$$Pr(x_i^*|x_i) = \begin{cases} (SN)^{x_i^*}(1-SN)^{1-x_i^*} \text{ if } x_i = 1, \\ (1-SP)^{x_i^*}(SP)^{1-x_i^*} \text{ if } x_i = 0, \end{cases} \tag{2.3}$$

and $Pr(x_i|x_1 \ldots x_{i-1})$ is determined by the switching probabilities.

In this thesis we are interested in the special case where the exposure is governed by a Markov chain, $Pr(x_i|x_1 \ldots x_{i-1}) = Pr(x_i|x_{i-1})$ and $Pr(x_1)$ is taken to be the stationary probability distribution of the ergodic Markov chain.

To determine the effect of misclassification, the joint probability distribution of $E\left\{\mathbf{X_{1:n}}|\mathbf{X^*_{1:n}}\right\}$ is tabulated for all 2^n possible values of $X^*_{1:n}$. The 2^n values of $E\left\{\mathbf{X_{1:n}}|\mathbf{X^*_{1:n}}\right\}\Psi$ are then expressed via the binary expansion with 2^n coefficients. This one-to-one correspondence determines the relationship between $\mathbf{Y}$ and $\mathbf{X}^*$,

$$E(Y_n|\mathbf{X^*_{1:n}}) = \beta_n X^*_n + \beta_{n-1} X^*_{n-1} + \ldots + \beta_{12\ldots n}(X^*_1 X^*_2 \ldots X^*_n). \tag{2.4}$$

This calculation allows us to determine the coefficients and the bias that results from using misclassified time-varying exposure measurements and how quantities such as sensitivity, specificity and Markov switching probabilities affect these quantities. Code for this calculation is presented in Appendix A.

Chapter 3

Bias Characterization for Time-Varying Exposure Misclassification

When examining a time-varying exposure that is subject to misclassification the normal rules of attenuation toward the null shown for the static case in Gustafson (2004) do not apply. This is true even in the most simplistic case.

Assume that the actual binary exposure-status across time is $X_1, X_2, X_3, \dots$ and that these time-varying exposures are misclassified ($X_j \to X_j^*$) independently over time. This misclassification is assumed to be nondifferential and can be characterized by (SN, SP). To characterize the effect of exposure misclassification we consider a linear outcome model and a linear outcome model including a lag term. Even in this simple case misclassification of the time-varying exposure will have significant effect. This is shown for specific examples in the next two sections.

3.1 Bias Characterization for Linear Outcome Model

To show the effect of time-varying exposure misclassification we consider a specific example. Let the binary exposure $X_1, X_2, X_3, \dots$ be a Markov chain with switching probabilities:

$$P(X_i = 1 | X_{i-1} = 0) = \phi_1, \qquad P(X_i = 0 | X_{i-1} = 1) = \phi_2, \tag{3.1}$$

for $i = 2, 3, \dots n$. We assume that the Markov chain is in its stationary distribution which implies that

$$P(X_1 = 0) = \frac{\phi_2}{\phi_1 + \phi_2}, \qquad P(X_1 = 1) = \frac{\phi_1}{\phi_1 + \phi_2}. \tag{3.2}$$

We further assume that the outcome variable only depends on the current exposure variable, with $\psi = 1$. Then at the fourth exposure observation we have

$$E(Y_4|X_1,\ldots,X_4) = X_4. \tag{3.3}$$

The choice of $\psi = 1$ is made without loss of generality and all results hold for arbitrary ψ. The choice of the fourth observation is also arbitrary and was chosen for computational convenience. Results hold for n exposure observations.

Equation (3.4) implies that the relationship of the outcome variable dependent on the misclassified exposure will have the form

$$E(Y_4|X_1^*,\ldots,X_4^*) = \beta_4 X_4^* + \beta_3 X_3^* + \ldots + \beta_{1234}(X_1^* X_2^* X_3^* X_4^*), \tag{3.4}$$

where if $(SN = 1, SP = 1)$ then $X_i^* = X_i$ and $\beta_{(\mathbf{-4})} = \vec{\mathbf{0}}$ and $\beta_4 = 1$.

In the linear outcome model the largest coefficients are β_4, β_3 and β_{34}. All other coefficients are estimated to be close to zero. Figure 3.1 displays the coefficient magnitudes when $(SN = 0.8, SP = 0.95)$ and common switching probability $\phi_1 = \phi_2 = \phi = 0.2$.

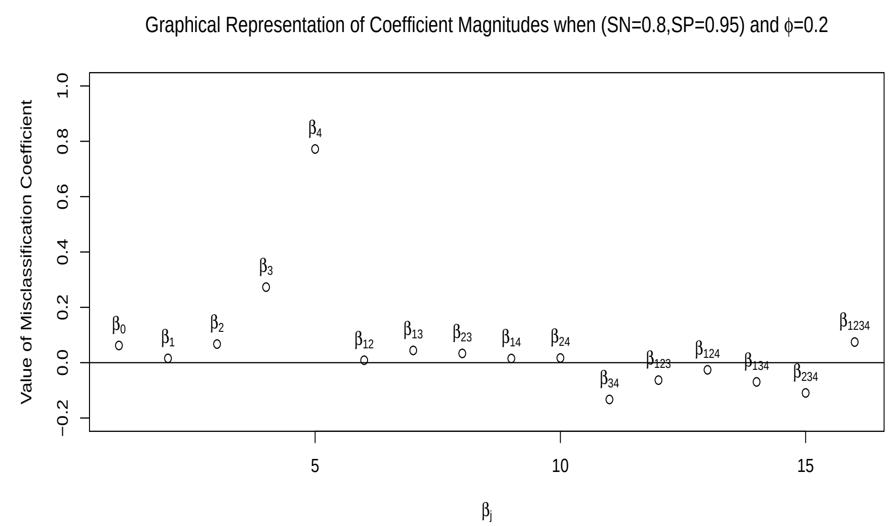

Figure 3.1: Coefficient magnitudes, when (SN=0.8, SP=0.95) and $\phi = 0.2$, showing β_4, β_3 and β_{34} to be the largest coefficients.

3.1.1 Effect of the Exposure Switching Probability

To determine how the switching probabilities of the exposure Markov chain affect the misclassification coefficients and model bias, we calculate these quantities for fixed (SN, SP) as ϕ_1 and ϕ_2 vary. We define model bias as the sum of the absolute differences between the coefficients (main effects and all interaction terms) of the linear outcome model when misclassification is present (β_j) and when misclassification is not present (ψ_j),

$$Bias = \sum_i |\psi_i - \beta_i|. \tag{3.5}$$

In all calculations $Pr(X_1)$ is taken to be the stationary probability distribution of the ergodic Markov chain defined in equation 3.2.

If the two exposure switching probabilities are equal ($\phi_1 = \phi_2 = \phi$), Figure 3.2 shows how the coefficient determination changes with ϕ when $(SN = 0.8, SP = 0.95)$. From Figure 3.2 we can see that β_4 is determined to be closest to the coefficient in the correctly classified exposure model when $\phi \approx 0.3$. The other misclassification coefficients are closest to the coefficient in the correctly classified exposure model at the same point as the bias is minimized at $\phi = 0.5$.

When $\phi = 0.5$ the next state is not affected by the previous state so there is no dependence in the exposure Markov chain. This means that for this model only the last exposure is important and we are essentially in the static case. This causes the familiar attenuation toward the null to occur. This is reflected by $\beta_{(-\mathbf{4})} \approx \vec{\mathbf{0}}$ and $\beta_4 = 0.77$.

When both switching probabilities (ϕ_1, ϕ_2) are allowed to vary independently, similar effects are shown. Figure 3.3 shows a determination surface for the misclassification coefficient β_4 when both switching probabilities vary and $(SN = 0.8, SP = 0.95)$. We can see that β_4 increases as switching probability 1, ϕ_1, increases until $\phi_1 \approx 0.3$ and then decreases after that. We can also see that β_4 remains almost constant as ϕ_2 varies except when ϕ_2 gets very small. When ϕ_2 is small it causes a strong negative effect on the determination of β_4. Figure 3.4 shows that the magnitude of β_3 remains close to zero unless ϕ_2 is small. The magnitude of β_3 becomes greatest when $\phi_1 \approx 0.3$ and ϕ_2 is low. The magnitude of β_{34} also becomes greatest when $\phi_1 \approx 0.3$ and ϕ_2 is low. This is shown in Figure 3.5. This implies that as ϕ_2 approaches its lower limit the effect of exposure misclassification becomes greatest. This is reflected in Figure 3.6 where we can see that the bias increases significantly when ϕ_2 is close to zero. The bias is most apparent when both switching probabilities approach zero or one.

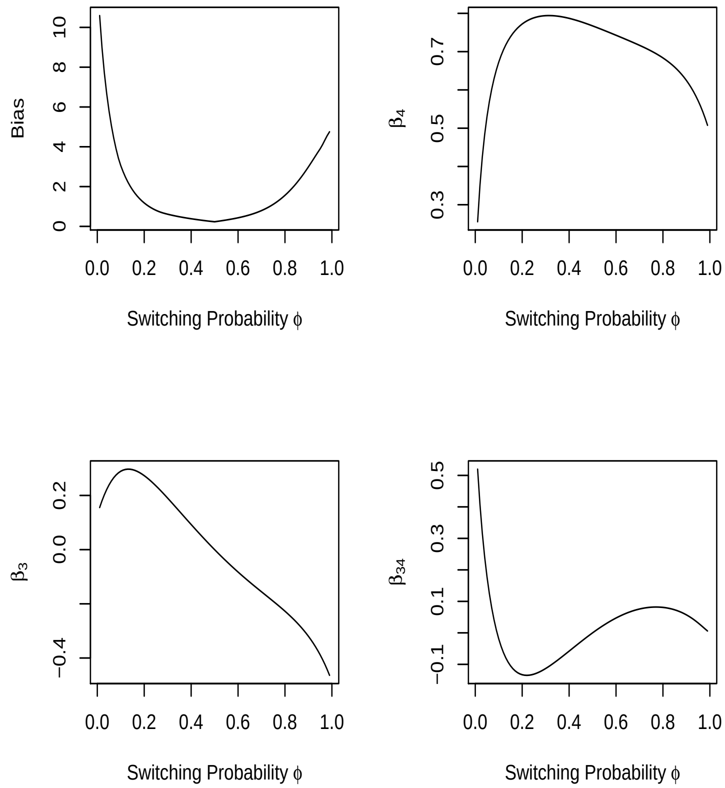

Figure 3.2: Effect of switching probability, ϕ, on determination of β_4, β_3, β_{34} and overall Bias for time-varying exposure misclassification when $(SN = 0.8, SP = 0.95)$.

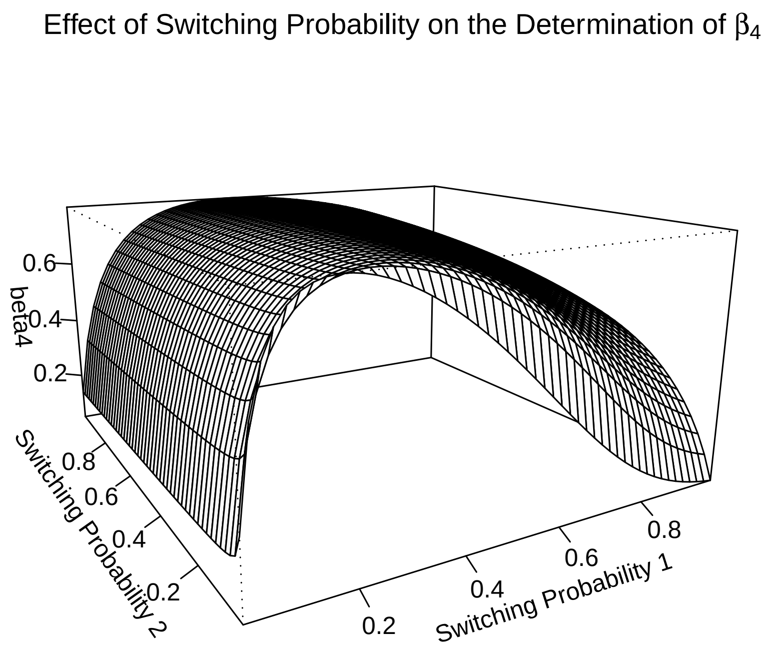

Figure 3.3: Effect of switching probabilities, (ϕ_1, ϕ_2), on determination of β_4 for time-varying exposure misclassification when $(SN = 0.8, SP = 0.95)$.

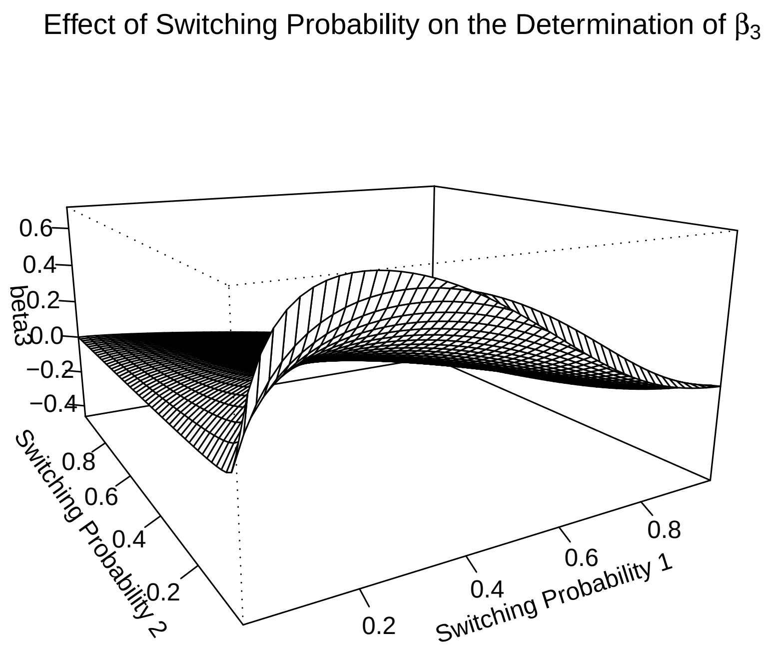

Figure 3.4: Effect of switching probabilities, (ϕ_1, ϕ_2), on determination of β_3 for time-varying exposure misclassification when $(SN = 0.8, SP = 0.95)$.

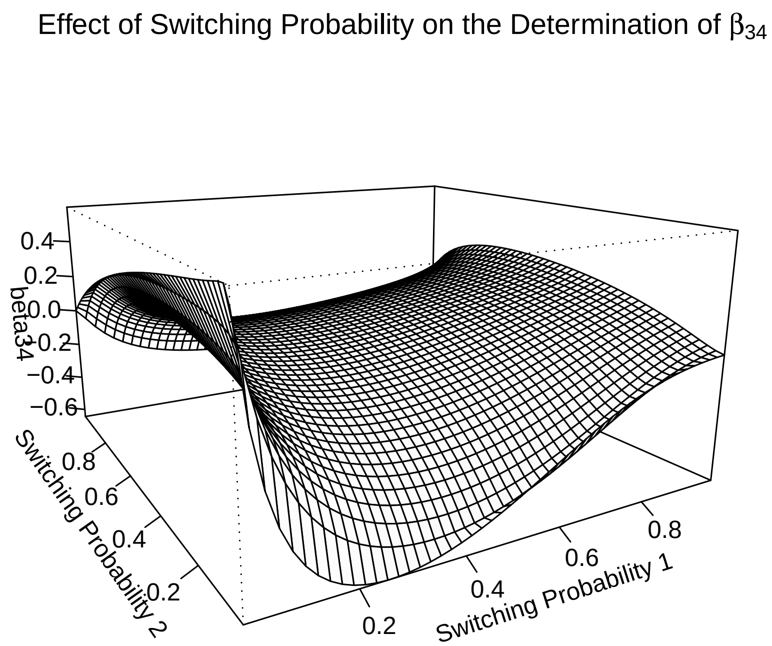

Figure 3.5: Effect of switching probabilities, (ϕ_1, ϕ_2), on determination of β_{34} for time-varying exposure misclassification when $(SN = 0.8, SP = 0.95)$.

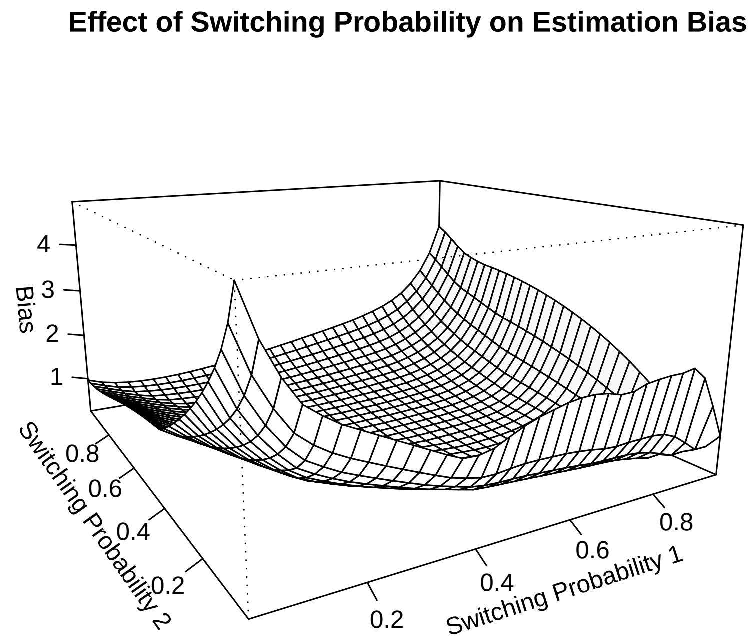

Figure 3.6: Effect of switching probabilities, (ϕ_1, ϕ_2), on bias for time-varying exposure misclassification when $(SN = 0.8, SP = 0.95)$.

When the values of (SN, SP) change, the determination surface of the misclassification coefficients differ with respect to the switching probabilities. The shapes of the determination surface for bias remains mostly unaffected by change in (SN, SP) but the determination surface for β_4, β_3 and β_{34} shifts depending on the relative magnitude of (SN, SP). When SN is small relative to SP, such as $(SN = 0.8, SP = 0.95)$, Figure 3.3 shows that the maximum of β_4 occurs around $\phi_1 \approx 0.3$. When SN is large relative to SP, such as $(SN = 0.95, SP = 0.8)$, the maximum of β_4 is shifted right and occurs around $\phi_1 \approx 0.7$, as shown in Figure 3.7. When SN and SP are approximately equal then the determination surface for β_4 is roughly symmetric and the maximum occurs when $\phi_1 \approx 0.5$, as shown in Figure 3.8. Similar asymmetric behavior is displayed by β_3 and β_{34} as shown by Figure 3.9 and Figure 3.10.

This asymmetric behavior is believed to occur because by changing the (SN, SP) we are just relabeling the states that are measured with precision. When SN is smaller than SP, the unexposed state is measured with more precision therefore misclassification has the least effect when ϕ_1 is low. When SN is larger then SP misclassification has the least effect when ϕ_1 is high.

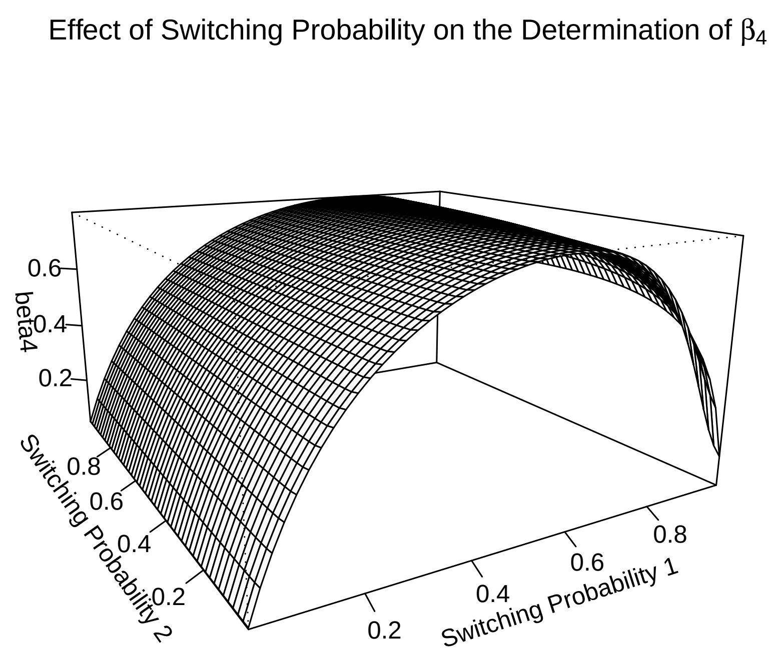

Figure 3.7: Effect of switching probabilities, (ϕ_1, ϕ_2), on determination of β_4 for time-varying exposure misclassification when $(SN = 0.95, SP = 0.8)$.

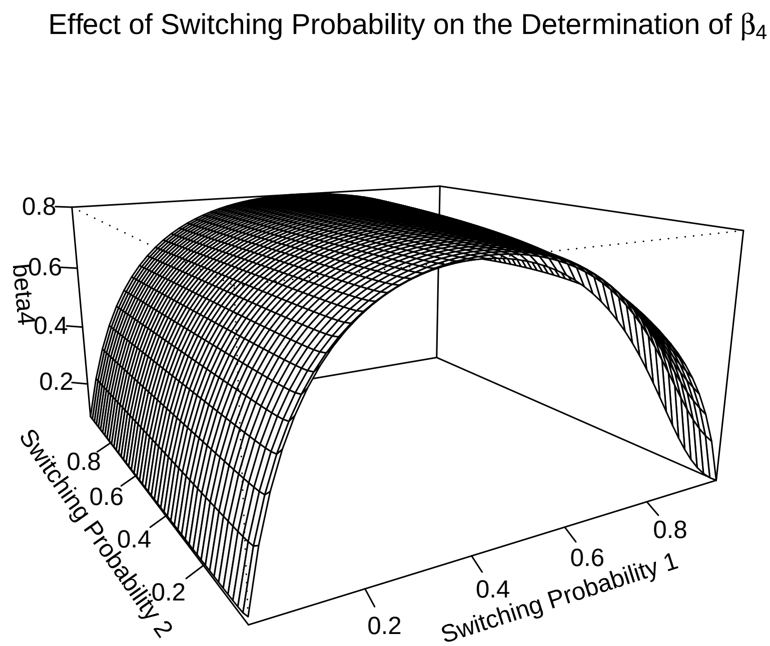

Figure 3.8: Effect of switching probabilities, (ϕ_1, ϕ_2), on determination of β_4 for time-varying exposure misclassification when $(SN = 0.9, SP = 0.9)$.

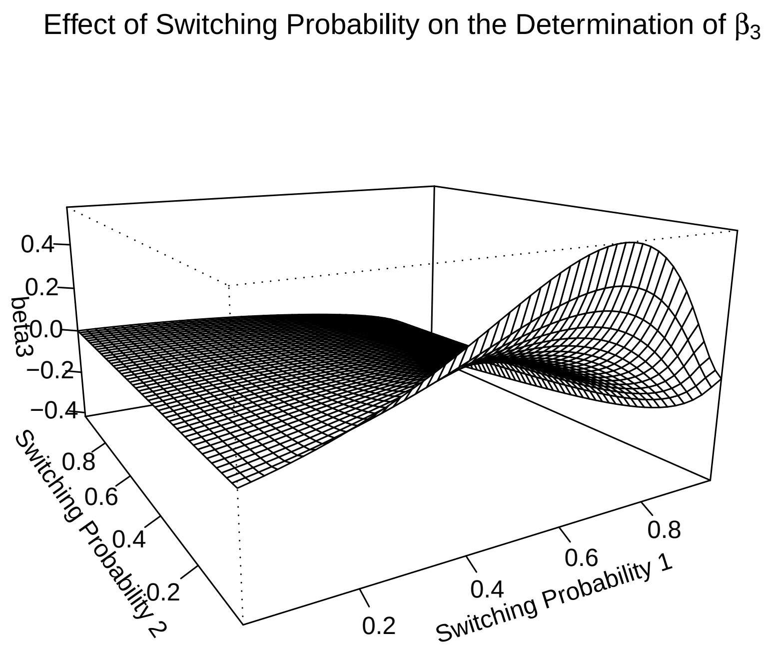

Figure 3.9: Effect of switching probabilities, (ϕ_1, ϕ_2), on determination of β_3 for time-varying exposure misclassification when $(SN = 0.95, SP = 0.8)$.

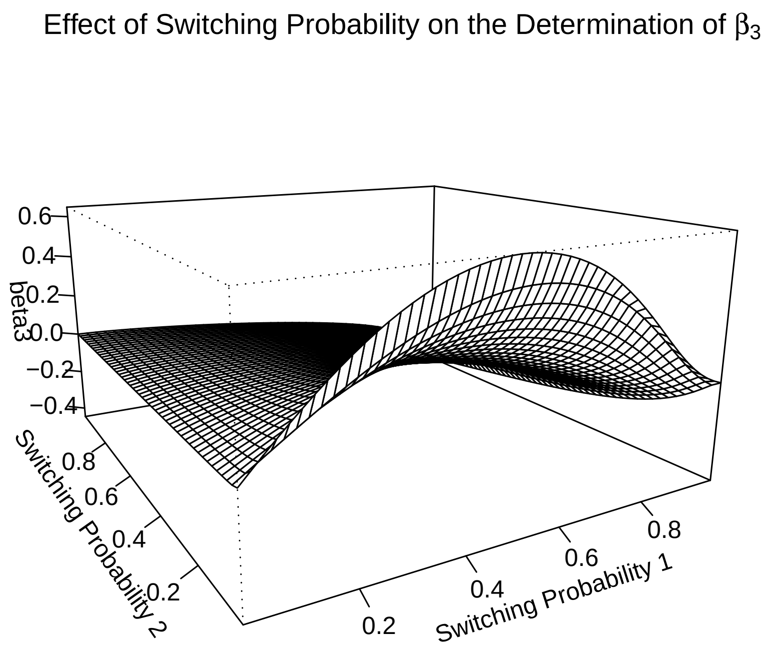

Figure 3.10: Effect of switching probabilities, (ϕ_1, ϕ_2), on determination of β_3 for time-varying exposure misclassification when $(SN = 0.9, SP = 0.9)$.

3.1.2 Effect of the Sensitivity and Specificity

To determine how SN and SP affect the impact of misclassification, for fixed switching probability ϕ, determination surfaces are created as SN and SP vary. Figure 3.11 shows that when $\phi = 0.2$ at low SN and high SP the effect of misclassification is greatest. A similar surface is produced when $\phi = 0.8$. This shows that when there is a switching probability that is not around 0.5, bias does not behave in a linear fashion with respect to (SN, SP) and high sensitivity is necessary for bias minimization. Figure 3.12 shows that when $\phi = 0.5$ bias does behave in a linear way, decreasing as either SN or SP increases.

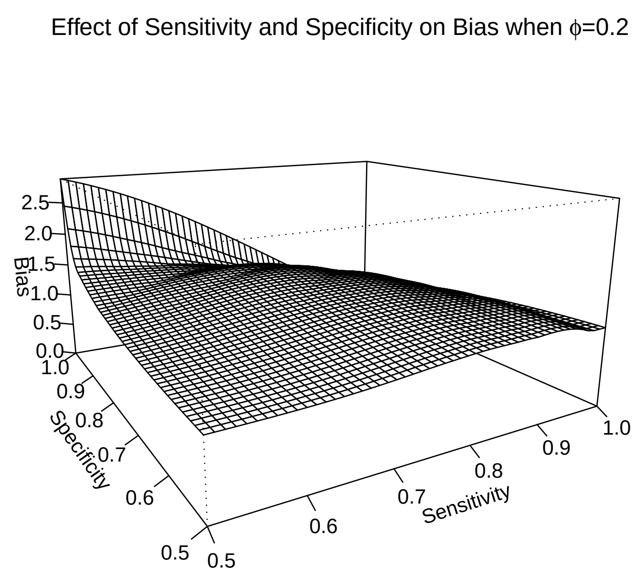

Figure 3.11: Effect of sensitivity and specificity on bias for time-varying exposure misclassification when $\phi = 0.2$.

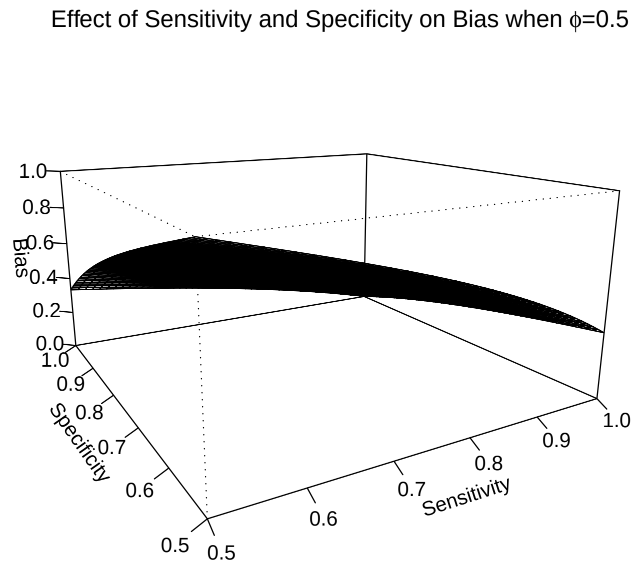

Figure 3.12: Effect of sensitivity and specificity on bias for time-varying exposure misclassification when $\phi = 0.5$.

Sensitivity and specificity have a linear effect on the determination of β_4. When $\phi = 0.2$, high values of specificity cause misclassification to have the least impact on the determination. When $\phi = 0.5$, sensitivity and specificity have an equal effect on the determination of β_4, and when $\phi = 0.8$, high values of sensitivity cause misclassification to have the least impact on the determination of β_4. This is shown in Figure 3.13 and Figure 3.14.

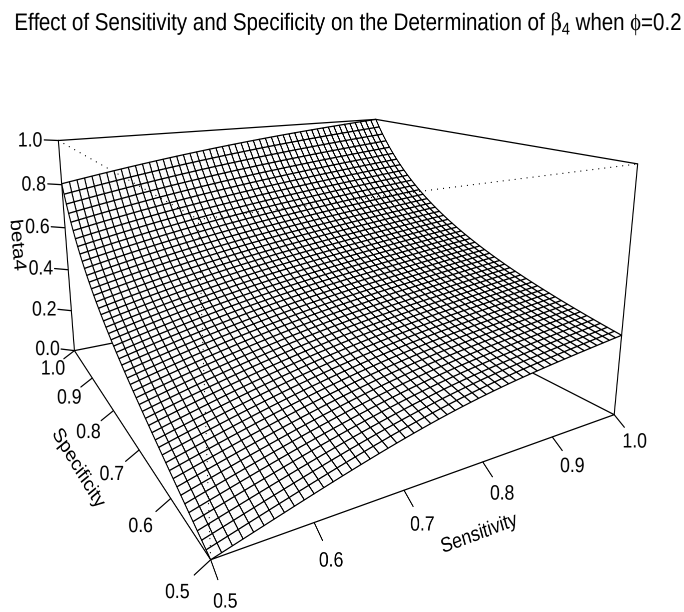

Figure 3.13: Effect of sensitivity and specificity on determination of β_4 for time-varying exposure misclassification when $\phi = 0.2$.

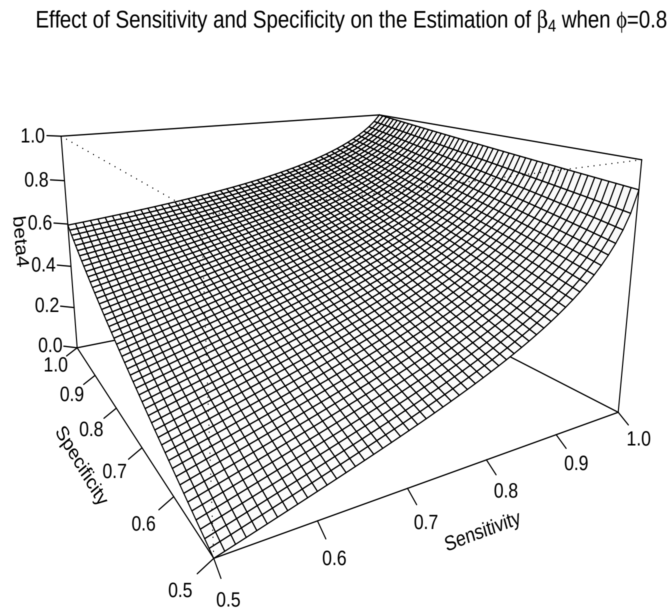

Figure 3.14: Effect of sensitivity and specificity on determination of β_4 for time-varying exposure misclassification when $\phi = 0.8$.

The determination of β_3 is affected by SN and SP in a non-linear way. Figure 3.15 shows that when $\phi = 0.2$, β_3 is positive and approaches zero as SN approaches one. Figure 3.16 shows that when $\phi = 0.8$ β_3 is negative and also approaches zero as SN approaches one. When $\phi = 0.5$ β_3 is approximately zero and (SN, SP) have no effect on the determination.

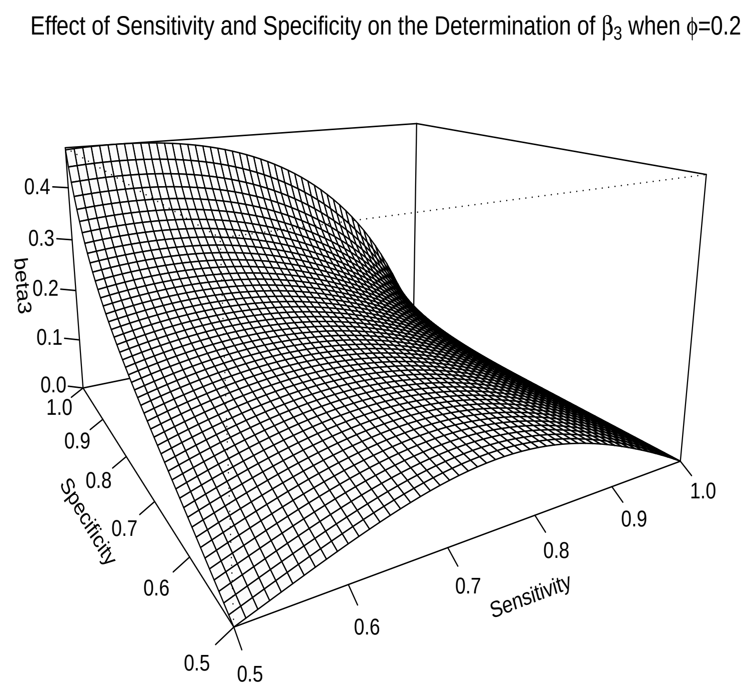

Figure 3.15: Effect of sensitivity and specificity on determination of β_3 for time-varying exposure misclassification when $\phi = 0.2$.

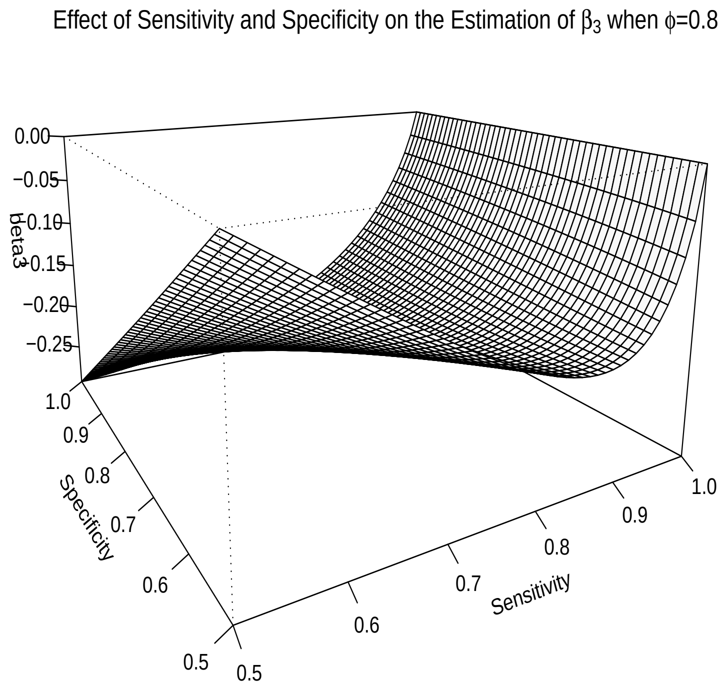

Figure 3.16: Effect of sensitivity and specificity on determination of β_3 for time-varying exposure misclassification when $\phi = 0.8$.

The determination surface for β_{34} formed as (SN, SP) vary is hyperbolic and the value of β_{34} cannot easily be predicted based on values of (SN, SP). This is seen in Figure 3.17. When $\phi = 0.5$, β_{34} is approximately zero and (SN, SP) have no effect on its determination.

When determining the effect of sensitivity and specificity, symmetric behavior as ϕ moves away from 0.5 is observed. The switching probabilities and (SN, SP) interact in a reciprocal way. This is can be explained by the symmetry that is obtained by just relabeling the exposed and unexposed states.

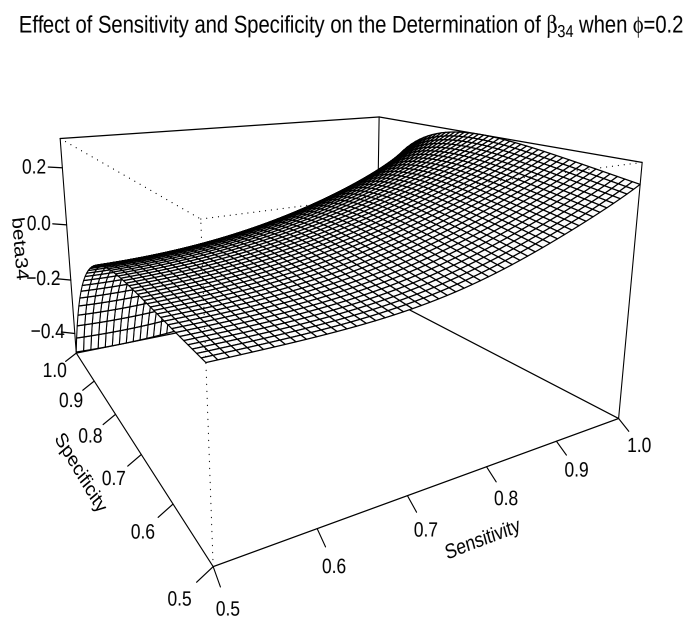

Figure 3.17: Effect of sensitivity and specificity on determination of β_{34} for time-varying exposure misclassification when $\phi = 0.2$.

3.2 Bias Characterization for Linear Outcome Model with Lag Term

To determine the effect of misclassification on a linear outcome model with lag term we consider a specific example. We again consider the binary exposure $X_1, X_2, X_3, \dots$ to be a Markov chain with switching probabilities ϕ_1 and ϕ_2, and independent nondifferential misclassification determined by (SN, SP). We further assume that the outcome variable depends on the current exposure variable with coefficient one and the previous lag term with coefficient 0.5. Therefore at the fourth exposure observation we have

$$E(Y_4|X_1, \dots, X_4) = X_4 + 0.5X_3. \tag{3.6}$$

This relationships implies that the outcome variable dependent on the misclassified exposure will have the form

$$E(Y_4|X_1^*, \dots, X_4^*) = \beta_4 X_4^* + \beta_3 X_3^* + \dots + \beta_{1234}(X_1^* X_2^* X_3^* X_4^*), \tag{3.7}$$

where if $(SN = 1, SP = 1)$ then $X_i^* = X_i$ and $\beta_4 = 1$, $\beta_3 = 0.5$ and $\beta_j = 0$ otherwise. In this case we study the behavior of the main effects β_4, β_3 and β_2. Figure 3.18 shows that when $(SN = 0.8, SP = 0.95)$ and $\phi = 0.2$ most of the other coefficients are estimated to be close to zero with the exception of some interaction terms. Figure 3.18 shows that under these condition the value of β_3 is overestimated. This means that when including lag terms misclassification can overestimate the effect of previous exposure.

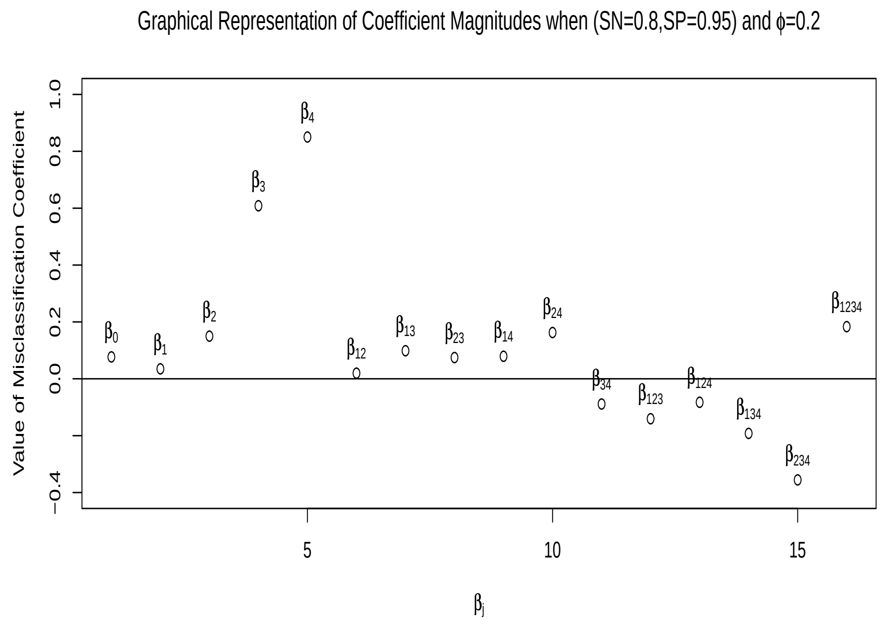

Figure 3.18: Coefficient magnitudes when $(SN = 0.8, SP = 0.95)$ and $\phi = 0.2$

3.2.1 Effect of the Exposure Switching Probability

To determine how the Markov switching probabilities affect the determination of misclassification coefficients and model bias, we again calculate these quantities for fixed (SN, SP) as ϕ_1 and ϕ_2 vary. Through this characterization many similarities can be seen between linear outcome models with lagged terms and linear model outcome models without lagged terms. Figure 3.19 shows that when lagged terms are included the determination of the leading coefficient (β_4) behaves similarly as when there is no lagged term present. When $(SN = 0.8, SP = 0.95)$ the familiar determination surface is seen with a sharp increase to the maximum obtained at $\phi_1 \approx 0.3$ and then slow descent as ϕ_1 increases. This shape is also observed for the determination surface of β_3 except that ϕ_2 has a greater effect. As ϕ_2 get small we see that the value of β_3 increases rapidly causing its effect to be overestimated. This can be seen in Figure 3.20. From Figure 3.3, Figure 3.19 and Figure 3.20 we can see that the determination of misclassification coefficients for exposure measurements that are present in the correctly specified linear outcome model behave similarly regardless of whether lagged terms are present or not.

The last misclassification coefficient that was examined was β_2. The random variable X_2 is not in the linear outcome model so if the exposure status is correctly classified then it should be zero. We can see from Figure 3.21 that the determination of β_2 behaves similarly to the determination to β_3 in the linear outcome model without lagged term. The determination of these coefficients correspond because they are both coefficients for the first random variable that is not included in the true linear outcome model.

The determination of the effect of switching probabilities on linear outcome models with lagged term have shown many similarities with the effect of switching probabilities on linear outcome models without lagged terms. Figure 3.22 shows that this correspondence is also displayed in the determination of bias.

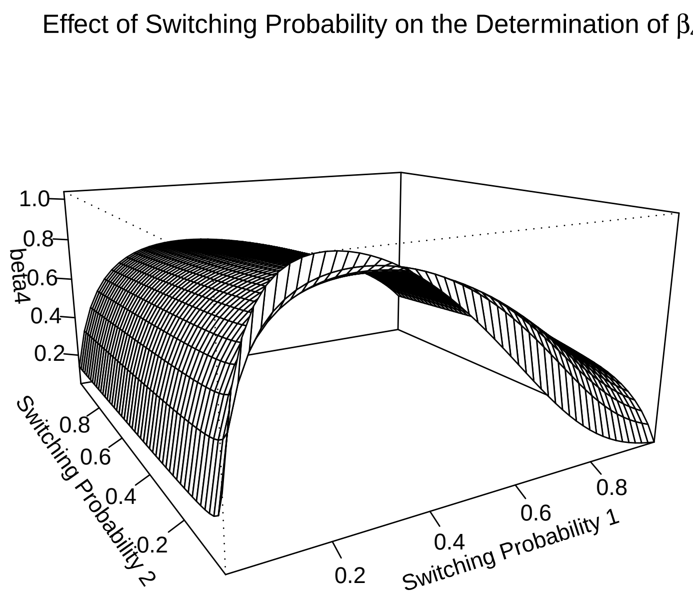

Figure 3.19: Effect of switching probabilities, (ϕ_1, ϕ_2), on determination of β_4 for time-varying exposure misclassification when $(SN = 0.8, SP = 0.95)$.

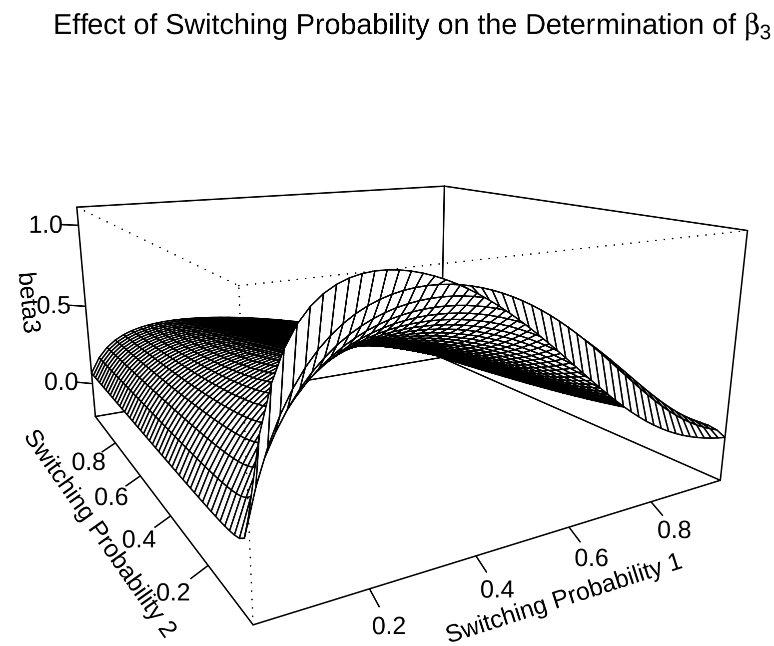

Figure 3.20: Effect of switching probabilities, (ϕ_1, ϕ_2), on determination of β_3 for time-varying exposure misclassification when $(SN = 0.8, SP = 0.95)$.

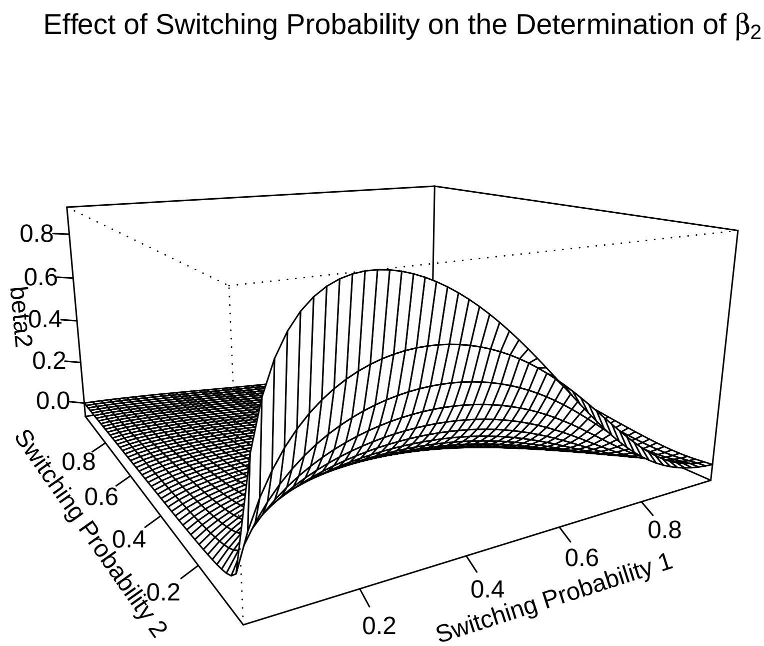

Figure 3.21: Effect of switching probabilities, (ϕ_1, ϕ_2), on determination of β_2 for time-varying exposure misclassification when $(SN = 0.8, SP = 0.95)$.

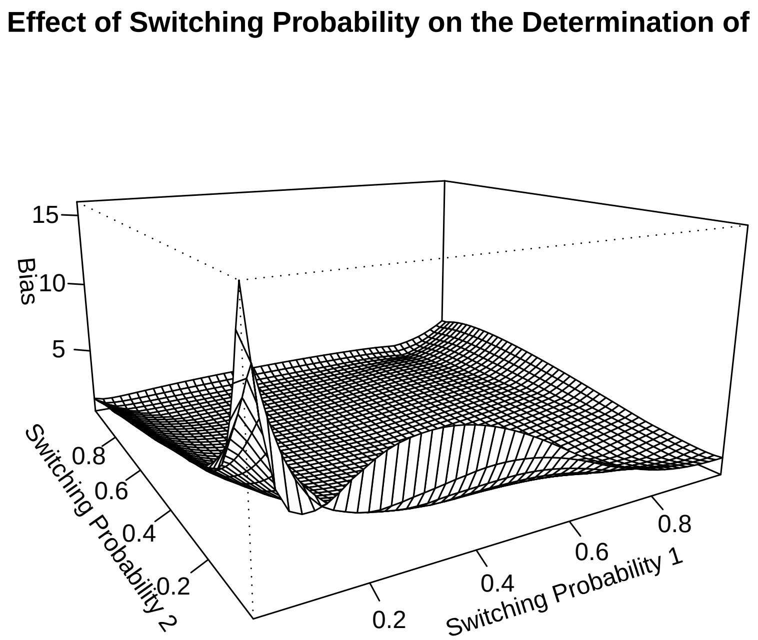

Figure 3.22: Effect of switching probabilities, (ϕ_1, ϕ_2), on Bias for time-varying exposure misclassification when $(SN = 0.8, SP = 0.95)$.

3.2.2 Effect of the Sensitivity and Specificity

To determine how SN and SP affect the impact of misclassification in a linear outcome model with a lag term, determination surfaces are created as SN and SP vary and ϕ is fixed. The coefficient of most interest in the outcome model with a lag term is that of β_3. Figure 3.24 and Figure 3.25 show that when ϕ is large β_3 behaves similar to β_4 in the linear model with no lag term. When ϕ is small something quite different occurs. When misclassification is present random variables that are in the true model almost always have their effect on the outcome variable underestimated. This is not the case for β_3 when ϕ is small. Figure 3.23 shows that depending on sensitivity and specificity β_3 can either be under or over estimated. This can cause a big problem in analysis because of the uncertainty of whether the lagged exposure effect is truly more important when misclassification is present or if it is truly less important.

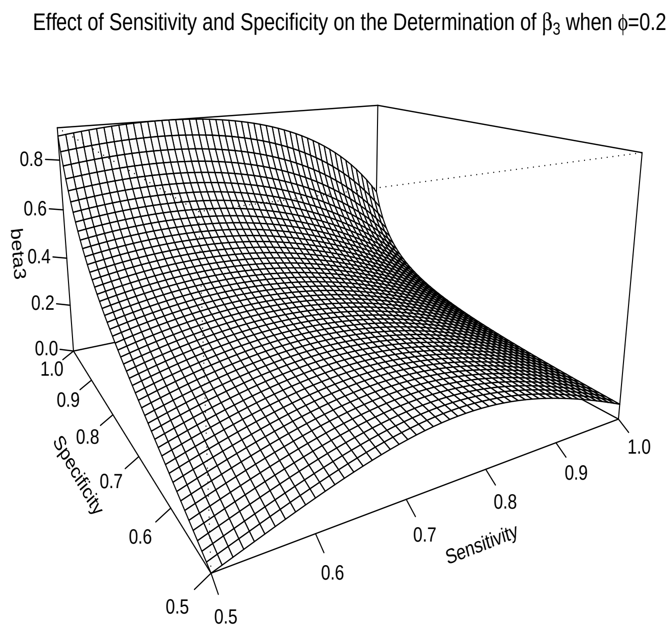

Figure 3.23: Effect of sensitivity and specificity on determination of β_3 for time-varying exposure misclassification with lagged term when $\phi = 0.2$.

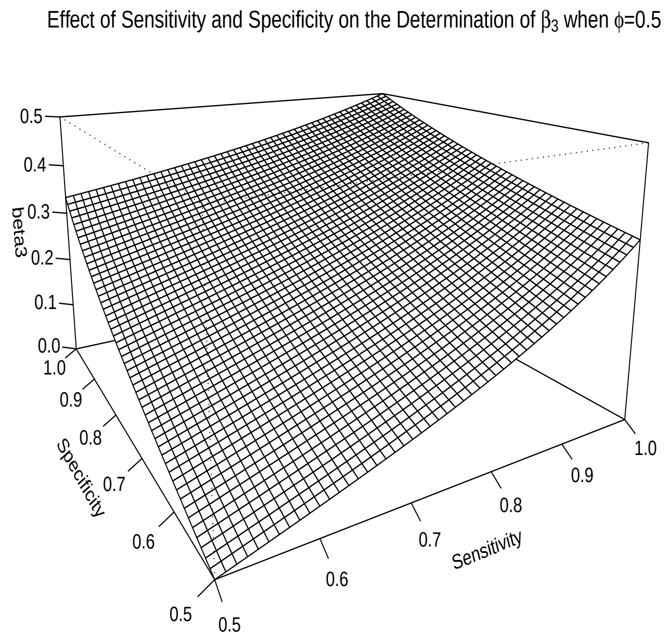

Figure 3.24: Effect of sensitivity and specificity on determination of β_3 for time-varying exposure misclassification with lagged term when $\phi = 0.5$.

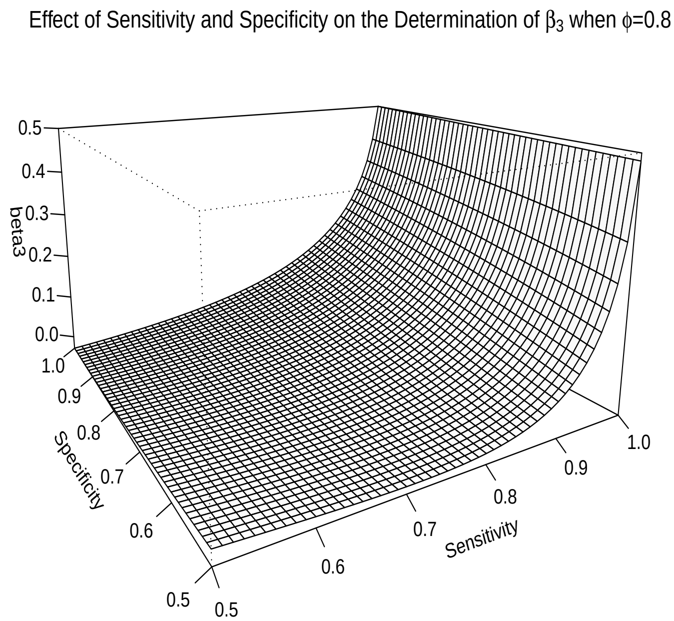

Figure 3.25: Effect of sensitivity and specificity on determination of β_3 for time-varying exposure misclassification with lagged term when $\phi = 0.8$.

The determination surfaces for β_4 and bias behave the same as in the previous model, as shown in Figure 3.26, Figure 3.27, Figure 3.28, Figure 3.29 and Figure 3.30. For β_2 we can see from Figure 3.31 that the determination surface behaves similarly to the determination to β_3 in the linear outcome model without lagged term. The determination of these coefficients correspond because they are both coefficients for the closest exposure term that is not included in the true linear outcome model.

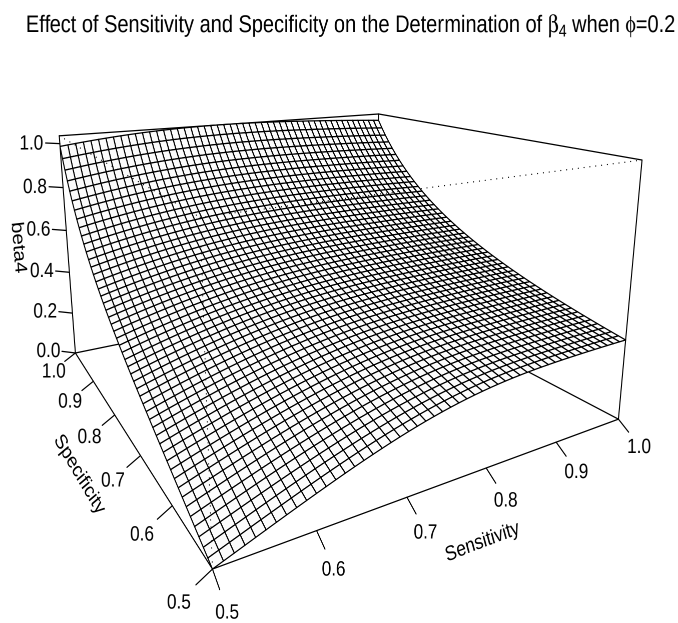

Figure 3.26: Effect of sensitivity and specificity on determination of β_4 for time-varying exposure misclassification with lagged term when $\phi = 0.2$.

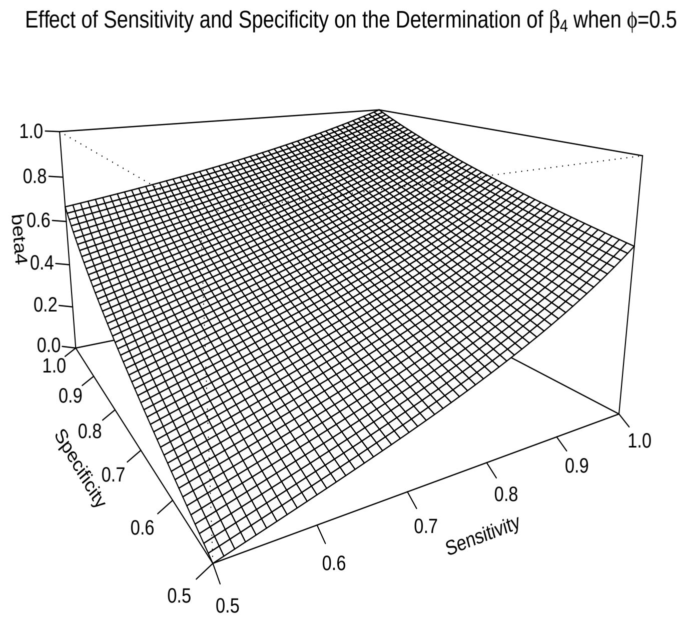

Figure 3.27: Effect of sensitivity and specificity on determination of β_4 for time-varying exposure misclassification with lagged term when $\phi = 0.5$.

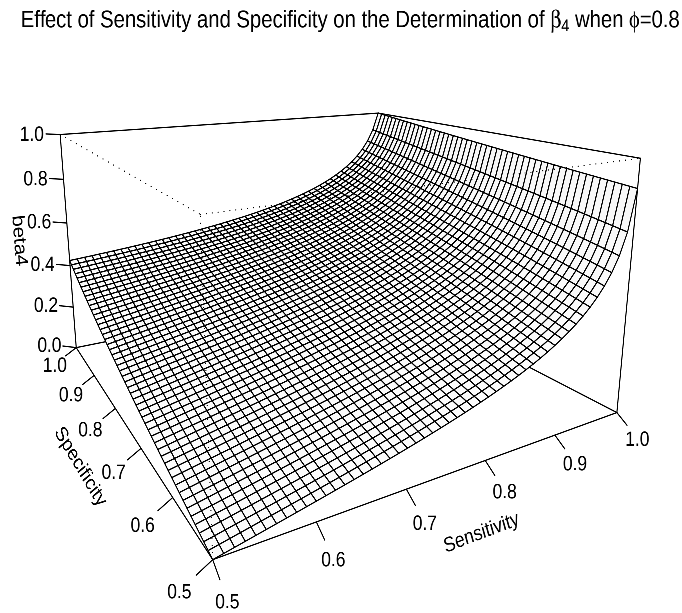

Figure 3.28: Effect of sensitivity and specificity on determination of β_4 for time-varying exposure misclassification with lagged term when $\phi = 0.8$.

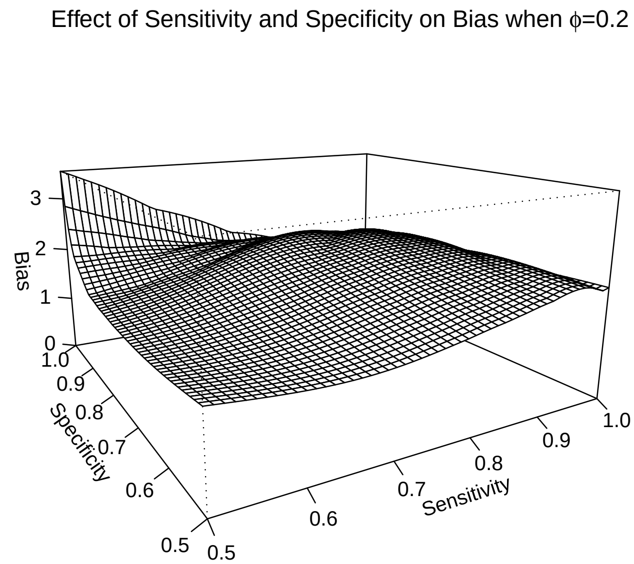

Figure 3.29: Effect of sensitivity and specificity on determination of bias for time-varying exposure misclassification with lagged term when $\phi = 0.2$.

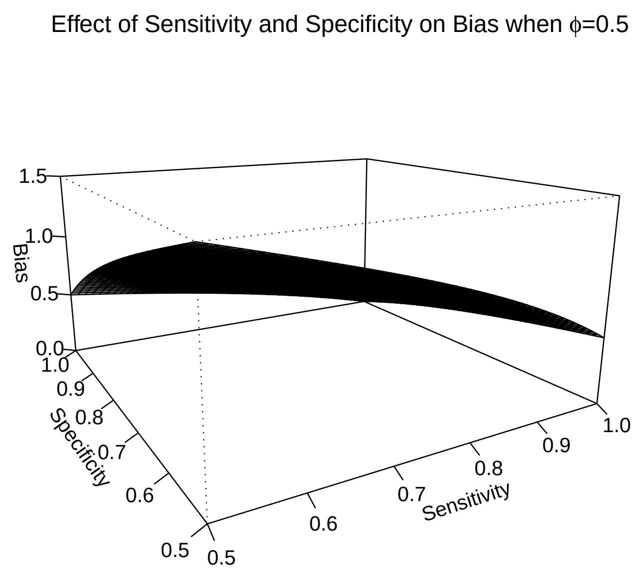

Figure 3.30: Effect of sensitivity and specificity on determination of bias for time-varying exposure misclassification with lagged term when $\phi = 0.5$.

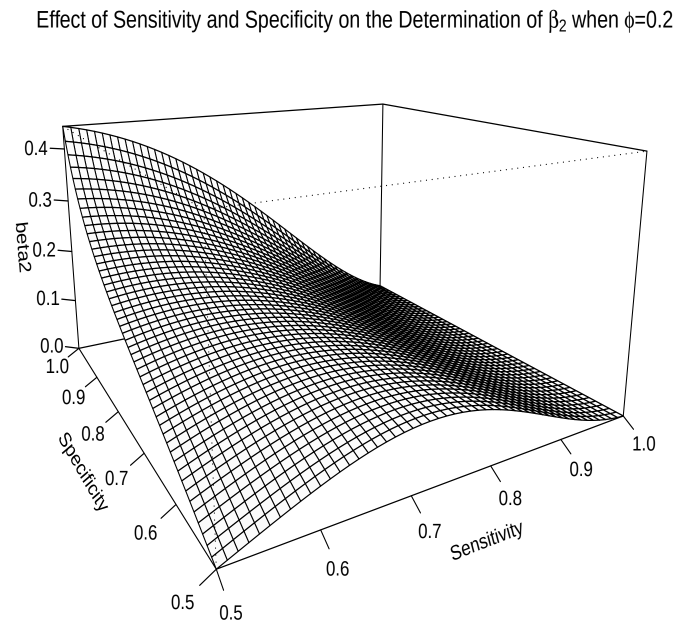

Figure 3.31: Effect of sensitivity and specificity on determination of β_2 for time-varying exposure misclassification with lagged term when $\phi = 0.2$.

3.3 The Effect of the Number of Exposure Measurements

In the previous two section we have arbitrarily chosen to study the effect of time-varying misclassification at the fourth exposure measurement. This choice was made due to computational efficiency, not for any reasons pertaining to misclassification. The results derived for the fourth time point will also hold, approximately, for any other number of time points greater then two as long as the exposure Markov chain is in its stationary distribution. This is because when the Markov chain is stationary the misclassified Markov chain will also be stationary because SN and SP are constant over time. The results are not exactly equivalent because as the number of time points increases so does the number of misclassification coefficients. This causes β_j to vary slightly but for all intents and purposes the results are the same. In our analysis we have chosen $Pr(X_1)$ to be the Markov chain's stationary distribution so the results are not dependent on the number of exposure measurements that are taken. This can be seen in Figure 3.32 where we examine the relationship:

$$E(Y_n|\mathbf{X}^*_{\mathbf{1:n}}) = \beta_n X^*_n + \beta_{n-1} X^*_{n-1} + \ldots + \beta_{12\ldots n}(X^*_1 X^*_2 \ldots X^*_n). \tag{3.8}$$

The lines for both β_n and β_{n-1} are horizontal indicating effectively no change with respect to number of exposure measurements. The curve for bias increases but this is believed to be solely the result of the exponentially increasing number of misclassification coefficients. If the exposure Markov chain is far from its stationary distribution then the number of exposure measurements might have an impact on the effect of misclassification and this should be investigated.

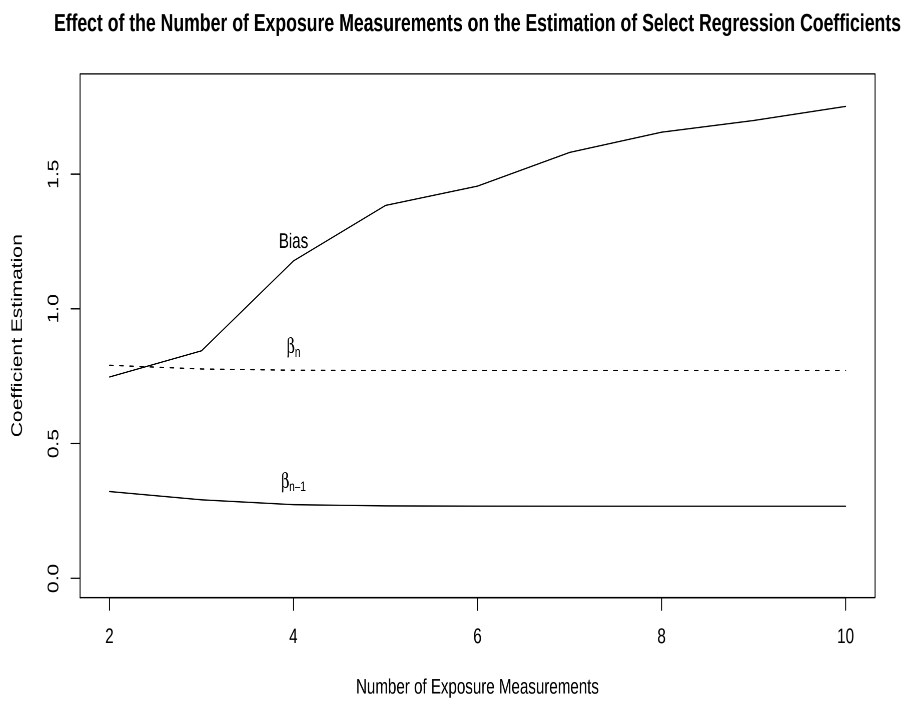

Figure 3.32: Effect of the number of exposure measurements on Bias, β_n and β_{n-1} for linear outcome model when $(SN = 0.8, SP = 0.95)$ and $\phi = 0.2$.

Chapter 4

Adjustment for Misclassification Using Discrete-Time Hidden Markov Process

Time-varying misclassification, where the underlying exposure is assumed to be a Markov chain, is an example of a hidden Markov chain. A hidden Markov model is a statistical model in which the system being modeled is assumed to be a Markov process with unobserved states. A hidden Markov model can be considered as a simple dynamic Bayesian network. The hidden Markov model and situations analyzed via the Kalman filter can be considered the most simple dynamic Bayesian networks.

In a regular Markov model, the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. In a hidden Markov model, the state is not directly visible, but output, dependent on the state, is visible. Each state gives rise to a probability distribution over the possible output symbols. Therefore the sequence of symbols generated by a hidden Markov model gives some information about the sequence of states.

When time-varying misclassification is present the possible output symbols are the misclassified exposure measurements while the states of the hidden Markov chain are the true, correctly classified, exposure status. By modeling time-varying misclassification using hidden Markov chains, estimates can be obtained for the Markov transition probabilities as well as the misclassification probabilities (SN, SP). After estimates for transition probabilities and misclassification probabilities are obtained, the most likely path through the underlying states can be recreated using the Viterbi Algorithm. This recreated path is an estimate of the underlying exposure status and thus can be used to adjust for misclassification using this estimated path as the time-varying exposure status in an exposure-disease model.

4.1 Discrete-Time Hidden Markov Process

We denote the observed sequence as $\{X_i^*\}$, for $i = 1, \ldots, n$ and the hidden Markov chain as $\{X_i\}$, for $i = 1, \ldots, n$. The history of the observed process up to time i is denoted by,

$$X_{1:i} = (X_1, \ldots, X_i), \tag{4.1}$$

where $i = 1, \ldots, n$. We define $X_{1:i}^*$ similarly.

The hidden Markov chain has m states denoted by $0, 1, \ldots, m-1$ and the underlying Markov chain has transition probability matrix denoted by Π, where the $(j,k)^{th}$ element is

$$\pi_{jk} = Pr(X_{i+1} = k | X_i = j) \quad \text{for } i = 1, \ldots, n; \;\; j, k = 0, \ldots, m-1. \tag{4.2}$$

In this analysis the Markov chain is assumed to be homogeneous, which means that for each j and k, π_{jk} is constant over time. The Markov chain can be stationary or non-stationary. A Markov chain is said to be stationary if the marginal distribution is the same over time, i.e. for each j, $\delta_{ij} = Pr(X_i = j)$ is constant for all i. The stationary marginal distribution is denoted by $\delta^s = (\delta_1, \ldots, \delta_m)$.

4.2 Inference for Discrete-Time Hidden Markov Process

To conduct inference for hidden Markov chains we must maximize the likelihood function. To do this the EM algorithm is used, since we only know the observations and not the sequence of states producing them (Durbin 1998). For this expectation algorithm the complete data likelihood, L_c, is

$$L_c = Pr(X_1^* = x_1^*, ..., X_n^* = x_n^*, X_1 = x_1, ..., X_n = x_n). \tag{4.3}$$

This can be shown to be

$$\begin{aligned} L_c &= Pr(X_1^* = x_1^* | X_1 = x_1) Pr(X_1 = x_1) \\ &\quad \prod_{i=2}^{n} Pr(X_i^* = x_i^* | X_i = x_i) Pr(X_i = x_i | X_{i-1} = x_{i-1}), \end{aligned}$$

and hence, substituting model parameters, we get

$$L_c = \delta_{1,x_1} \pi_{x_1 x_2} \pi_{x_2 x_3} \ldots \pi_{x_{n-1} x_n} \prod_{i=1}^{n} Pr(X_i^* = x_i^* | X_i = x_i), \tag{4.4}$$

so

$$\log L_c = \log \delta_{1,x_1} + \sum_{i=2}^{n} \log \pi_{x_{i-1}x_i} + \sum_{i=1}^{n} \log Pr(X_i^* = x_i^* | X_i = x_i). \quad (4.5)$$

Hence the complete data likelihood is split into three terms: the first relates to parameters of the marginal distribution of the Markov chain, the second to the transition probabilities, and the third to the distribution parameters of the observed random variable (MacDonald and Zucchini 1997).

When the hidden Markov chain is assumed to be non-stationary, the complete data likelihood has a neat structure, in that delta only occurs in the first term, Π only occurs in the second term, and the parameters associated with the observed probabilities only occur in the third term. Hence, the likelihood can easily be maximized by maximizing each term individually. In this situation, the estimated parameters using EM algorithm will be the exact maximum likelihood estimates.

When the hidden Markov chain is assumed to be stationary, $\delta = \delta\Pi$, and then the first two terms of $\log L_c$ determine the transition probabilities Π. This raises more complicated numerical problems, as the first term is effectively a constraint. This is dealt with in a slightly ad-hoc manner by effectively disregarding the first term, which is assumed to be relatively small. In the M-step, the transition matrix is determined by the second term, then δ is estimated using the relation $\delta = \delta\Pi$ (Harte 2010). Both these methods give maximum likelihood estimates for the transition probabilities and the misclassification probabilities (SN, SP).

4.3 Recreating the Path Through True Exposure States

To adjust for time-varying misclassification we want to estimate the true time-varying exposure states of each subject. To do this we can predict the most likely sequence of the true Markov exposure states given the observed misclassified states using the Viterbi algorithm. The purpose of the Viterbi algorithm is to globally decode the underlying hidden Markov state at each time point. It does this by determining the sequence of states $(k_1^*, \ldots, k_n^*)$ which maximizes the joint distribution of the hidden states given the entire observed process,

$$(k_1^*, \ldots, k_n^*) = argmax_{k_1,\ldots,k_n \epsilon \{1,2,\ldots,m\}} Pr(X_1 = k_1, \ldots, X_n = k_n | X_{1:n}^* = x_{1:n}^*).$$

The algorithm has been taken from Zucchini (2005).

Determining the a posteriori most probable state at time i is referred to as local decoding,

$$k_i^* = argmax_{k\epsilon\{1,2,\ldots,m\}} Pr(X_i = k|X_{1:n}^* = x_{1:n}^*).$$

Once the sequence of states $(k_1^*, \ldots, k_n^*)$ which maximizes the joint distribution of the hidden states is determined, this can be used as an estimate of the true time-varying exposure status for each subject. With this estimated exposure Markov chain, inference can be done using a plug-in method. This is done by using this estimate as the time-varying exposure status and determining the exposure-outcome relationship using this estimated Markov chain.

4.4 Adjusting for Misclassification: True Exposure Path Recreation

In order to demonstrate the performance of adjusting for time-varying misclassification using the most-likely path through true exposure states, we conduct a simulation study under four cases. In each case the underlying time-varying exposure and misclassification are generated under the same conditions. The exposure-outcome model differs in each situation and the effectiveness of the adjustment is evaluated.

The R package 'HiddenMarkov' contains functions for the analysis of discrete time hidden Markov models, Markov modulated GLMs and the Markov modulated Poisson process. It includes functions for simulation, parameter estimation, and the Viterbi algorithm. The package is currently under development and it is designed for a single long Markov chain not a series of longitudinal data. This means that transition probabilities and misclassification probabilities cannot be accurately estimated when there only exist short hidden Markov chains. In the time-varying exposure case this is the type of data we are interested in, so the transition probabilities and misclassification probabilities need to be estimated by a different means. The package can still accurately calculate the most likely path through hidden states using the Viterbi algorithm when estimates for transition probabilities and misclassification probabilities are available. This means that when accurate estimates of transition probabilities and misclassification probabilities exist, such as from a validation study, then the most likely path through true exposure states can be determined. This allows misclassification to be adjusted for by recreating the most likely true exposure path for each subject. The effectiveness of this adjustment is shown through the simulation study in the next section.

4.4.1 Data Simulation

To demonstrate the performance of misclassification adjustment, 500 Monte Carlo samples were simulated as follows:

1. The total number of subjects is $N = 1,000$ and the number of exposure measurements on each subject is $n = 10$

2. For each subject i generate an exposure Markov chains $X_{i1}, X_{i2}, \ldots X_{i,n}$ where the first exposure measurement X_{i1} is generated from the stationary distribution of the Markov chain. The Markov chain is defined by its transition probabilities.

   - $(\phi_1 = 0.1, \phi_2 = 0.3)$
   - $Pr(X_{i1} = x) = 0.25^x(1 - 0.25)^{1-x} \qquad \text{for } x = 0, 1.$

3. Generate the exposure outcome model in two cases:

   - $Y_{ij} \sim N(X_{ij}, 0.1)$ for $i = 1, \ldots, N$ and $j = 1, \ldots, n.$
   - $Y_{ij} \sim N(X_{ij} + 0.5X_{i,j-1}, 0.1)$ for $i = 1, \ldots, N$ and $j = 2, \ldots, n.$

4. Misclassify the Markov chain $X^*_{i1}, X^*_{i2}, \ldots X^*_{i,n}$ with $(SN = 0.85, SP = 0.95)$.

5. Recreate the most likely exposure path $X^{est}_{i1}, X^{est}_{i2}, \ldots X^{est}_{i,n}$ using the Viterbi algorithm with estimates $(SN_0 = 0.85, SP_0 = 0.95, \phi_{1_0} = 0.1, \phi_{2_0} = 0.3)$.

6. Consider two different exposure outcome models:

   - $E(Y_{ij}|X_{i,1:j}) = \beta_1 X_{ij}$
   - $E(Y_{ij}|X_{i,1:j}) = \beta_1 X_{ij} + \beta_2 X_{i,j-1}$

7. Fit each exposure outcome model using

$$L(\beta) = \prod_{i=1}^{N} \prod_{j=1}^{n} f(Y_{ij}|d_{i,1:j}), \tag{4.6}$$

   where $d_{i,1:j} = X_{i,1:j}$, $X^*_{i,1:j}$, or $X^{est}_{i,1:j}$ when estimating β_{true}, $\beta_{misclassified}$, or $\beta_{estimate}$ respectively.

4.4.2 Simulation Results

The simulation can be broken into four cases. Cases are determined by the form of the linear outcome model (lagged term or no lagged term) and whether the model has been correctly specified. In each case the coefficient estimates for the exposure outcome model are presented based on the true data, misclassified data and estimated data. The simulation standard deviation for the 500 Monte Carlo Samples and Monte Carlo 95% confidence interval are also presented in the tables below. The percentage of Misclassified states, before and after the most likely path recreation, are presented in Table 4.1.

	% of States	Std. Dev.	95% CI
Misclassified	6.88	0.26	(6.86, 6.90)
Misclassified after Recreation	6.00	0.29	(5.97, 6.03)

Table 4.1: Comparison of the percentage of misclassified states before and after Viterbi path recreation when a sample of 1,000 subjects was taken with 10 observations per subject. $(SN = 0.85, SP = 0.95)$, $(\phi_1 = 0.1, \phi_2 = 0.3)$.

Case 1 .

True Model: $E(Y_{ij}|X_{i,1:j}) = X_{ij}$
Assumed Model: $E(Y_{ij}|X_{i,1:j}) = \beta X_{ij}$

Coeff.	Coeff.est	Std. Dev.	95% CI
β_{true}	1.000	0.0008	(0.9999, 1.0001)
$\beta_{estimate}$	0.823	0.0035	(0.8227 0.8233)
$\beta_{misclassified}$	0.762	0.0033	(0.7617, 0.7623)

Table 4.2: Simulation results for misclassification adjustment in discrete time for linear outcome model with no lag term when $(SN = 0.85, SP = 0.95)$, $(\phi_1 = 0.1, \phi_2 = 0.3)$ and a sample of 1,000 subjects was taken with 10 observations each.

Case 2 .

True Model: $E(Y_{ij}|X_{i,1:j}) = X_{ij} + 0.5X_{i,j-1}$
Assumed Model: $E(Y_{ij}|X_{i,1:j}) = \beta X_{ij}$

Coeff.	Coeff.est	Std. Dev.	95% CI
β_{true}	1.269	0.0018	(1.2688, 1.2692)
$\beta_{estimate}$	1.095	0.0047	(1.0946, 1.0954)
$\beta_{misclassified}$	0.987	0.0047	(0.9866, 0.9874)

Table 4.3: Simulation results for misclassification adjustment in discrete time for a misspecified linear outcome model with no lag term when ($SN = 0.85, SP = 0.95$), ($\phi_1 = 0.1, \phi_2 = 0.3$) and a sample of 1,000 subjects was taken with 10 observations each.

Case 3 .

True Model: $E(Y_{ij}|X_{i,1:j}) = X_{ij} + 0.5X_{i,j-1}$
Assumed Model: $E(Y_{ij}|X_{i,1:j}) = \beta_1 X_{ij} + \beta_2 X_{i,j-1}$

Coeff.	Coeff.est	Std. Dev.	95% CI
$\beta 1_{true}$	1.000	0.0010	(0.9999, 1.0001)
$\beta 1_{estimate}$	0.791	0.0049	(0.7906, 0.7914)
$\beta 1_{misclassified}$	0.806	0.0031	(0.8057, 0.8063)
$\beta 2_{true}$	0.499	0.0010	(0.4989, 0.4991)
$\beta 2_{estimate}$	0.472	0.0051	(0.4716, 0.4724)
$\beta 2_{misclassified}$	0.539	0.0031	(0.5387, 0.5393)

Table 4.4: Simulation results for misclassification adjustment in discrete time for linear outcome model with lagged term when ($SN = 0.85, SP = 0.95$), ($\phi_1 = 0.1, \phi_2 = 0.3$) and a sample of 1,000 subjects was taken with 10 observations each.

Case 4 .

True Model: $E(Y_{ij}|X_{i,1:j}) = X_{ij}$
Assumed Model: $E(Y_{ij}|X_{i,1:j}) = \beta_1 X_{ij} + \beta_2 X_{i,j-1}$

Coeff.	Coeff.est	Std. Dev.	95% CI
$\beta 1_{true}$	0.999	0.0010	(0.998, 1.000)
$\beta 1_{estimate}$	0.768	0.0045	(0.7676, 0.7684)
$\beta 1_{misclassified}$	0.712	0.0032	(0.7117, 0.7123)
$\beta 2_{true}$	0.000	0.0010	(-0.0001, 0.0001)
$\beta 2_{estimate}$	0.073	0.0046	(0.0726, 0.0734)
$\beta 2_{misclassified}$	0.195	0.0029	(0.1947, 0.1953)

Table 4.5: Simulation results for misclassification adjustment in discrete time for linear outcome model with misspecified lagged term when $(SN = 0.85, SP = 0.95)$, $(\phi_1 = 0.1, \phi_2 = 0.3)$ and a sample of 1,000 subjects was taken with 10 observations each.

The results above show that recreation of the true exposure path using the Viterbi algorithm, with the true values of (SN, SP) and (ϕ_1, ϕ_2), reduces the number of misclassified states and positively adjusts for misclassification. The coefficient estimates of the exposure outcome model are significantly closer to the true exposure outcome model. This adjustment is least effective when we are fitting an exposure outcome model with lag term that matches the data generating model. Although there are significantly less misclassified states this is not reflected in the coefficient estimates. This model does however adjust the estimate for β_2 so that an artificially strong association between the outcome and the lagged exposure term is no longer observed. The adjustment is most effective in adjusting for misclassification when a misspecified model is fit including extra lagged terms. This is very beneficial because when model selection procedures are employed most procedures start with a saturated/partially saturated model and then remove covariates that are not significant. We see that in Case 4 when the exposure status X_{n-1} had no association with the outcome variable $\beta_{misclassified}$ was still large, indicating a relationship between X_{n-1} and Y that did not exist. When the most likely exposure path was recreated the effect of X_{n-1} dropped dramatically and $\beta_{estimate}$ was close to zero, indicating no association. This is very helpful for model selection and allows artificial associations that result from misclassification to be minimized.

Chapter 5

Adjustment for Misclassification Using Continuous-Time Hidden Markov Process

Continuous time Markov chains have found a wide application in the medical and social sciences, especially in studies that consist of data that record life history of exposures for individuals. In this chapter we consider continuous time Markov process with panel data. The panel data consists of the states occupied by the individuals under study at a sequence of discrete time points but no information is available about the timing of events between observation times. One of the most useful properties of continuous-time Markov chains is their ability to model multi-state processes under this type of panel data.

In practice most time-varying exposures will occur according to a continuous time process. Modeling time-varying exposure by a continuous time Markov process allows transitions between exposure states to happen at any time and allows for transitions to happen between observations. Observation times of the process are arbitrary and they no longer need to be equally spaced. The ability for continuous time Markov chains to model panel data means exact transition times do not need to be observed. These advantages of continuous time Markov chains make them a very important extension of discrete time Markov processes. All the theory for hidden Markov chains in discrete time can be extended to continuous time hidden Markov chains, making continuous time hidden Markov processes a very powerful tool for modeling and adjusting for misclassification. By modeling time-varying misclassification using a continuous time hidden Markov chain, estimates can be obtained for the Markov transition probabilities as well as the misclassification probabilities (SN, SP). This can be done using only the observed data so a validation sample does not need to be obtained. This estimation can be extended to allow transition intensities and misclassification probabilities to depend on accurately measured covariates. Due to the fact the misclassification probabilities can depend on measured covariates, differential misclassification can be modeled by letting outcome status be a covariate. In this thesis we only consider time-varying exposure and misclassification that does not depend on covariates.

When estimates for transition probabilities and misclassification probabilities are obtained the path through the underlying states can be recreated with highest probability using the Viterbi Algorithm. This recreated path is an estimate of the true underlying exposure status and thus can be used to adjust for misclassification using this estimated Markov chain as the time-varying exposure status of each subject.

5.1 Continuous-Time Markov Process

Suppose individuals move independently among m states according to a continuous-time Markov process. Let $X(t)$ be the state occupied at time t by a randomly chosen individual. For $0 \leq s \leq t$, let $P(s,t)$ be the $m \times m$ transition probability matrix with entries

$$p_{ij}(s,t) = Pr(X(t) = j | X(s) = i), \tag{5.1}$$

for $i, j = 0, 1, ..., m-1$. This process can be specified in terms of the transition intensities,

$$q_{ij}(t) = lim_{\Delta t \to 0} \frac{p_{ij}(t, t+\Delta t)}{\Delta t}, \quad i \neq j \tag{5.2}$$

and

$$q_{ii}(t) = -\sum_{i \neq j} q_{ij}(t), \quad i = 1, ..., k, \tag{5.3}$$

and let $Q(t)$ be the $m \times m$ transition intensity matrix with entries $q_{ij}(t)$. In this chapter only the time homogeneous models are investigated which implies $q_{ij}(t) = q_{ij}$. In the time homogeneous case the process is stationary and

$$P(t) = P(s, s+t) = P(0,t), \tag{5.4}$$

in this case we can write

$$P(t) = e^{Qt} = \sum_{h=0}^{\infty} Q^h t^h / h!. \tag{5.5}$$

We can also allow $q_{ij} = q_{ij}(\beta)$ to depend on b functionally independent parameters $\beta_1, \beta_2, ..., \beta_b$. If each individual's multi-state transitions depend on measured covariates then adjustments can be made using, a generalized Cox proportional hazards model. This entails setting the transition rate functions of an individual's covariates as follows:

$$q(x) = \psi(x;\beta)q_0, \tag{5.6}$$

where x is a vector of patient covariates, β the corresponding coefficients, and q_0 the baseline transition rate. $\psi()$ can take several parameterizations including the exponential, linear, logistic, and augmented family forms.

To be able to estimate the transition probabilities we must determine the likelihood function. Suppose that a random sample of N individuals is observed at times $t_0, t_1, ..., t_n$. If N_{ijl} denotes the number of individual in state i at t_{l-1} and j at t_l then conditioning on the distribution of individuals among states at t_0 the likelihood function for β is

$$L(\beta) = \prod_{l=1}^{n} \left\{ \prod_{i,j=1}^{m} p_{ij}(t_{l-1}, t_l)^{N_{ijl}} \right\}. \tag{5.7}$$

5.2 Maximum Likelihood Estimation

Maximum likelihood estimation can be conducted to estimate transition probabilities when these transitions depend on measured covariates or when they do not. This is done using an efficient quasi-Newton procedure that uses first derivatives of $\log L(\beta)$ to compute $P(t;\beta) = exp(Q(\beta))$. For a given β, $Q(\beta)$ is decomposed into $Q = ADA^{-1}$. Here $D = diag(d_1, ..., d_k)$, where $d_1, ..., d_k$ are the distinct eigenvalues of Q, and A is the $k \times k$ matrix whose jth column is the right eigenvector corresponding to d_j. Then

$$P(t) = A\, diag(e^{d_1 t}, ..., e^{d_k t})A^{-1}. \tag{5.8}$$

Using this expression for $P(t)$ in our likelihood function and using the quasi-Newton (or scoring) procedure the MLE can be determined. This procedure is implemented in the R package msm.

5.3 Continuous Time Hidden Markov Process

In a hidden continuous Markov model the states of the Markov chain are not observed. The observed data are governed by some probability distribution conditionally on the unobserved state. The evolution of the underlying Markov chain is governed by a transition intensity matrix Q. Hidden Markov models are mixture models, where observations are generated from a certain number of unknown distributions. However the distribution changes through time according to states of a hidden Markov chain. Multi-state models with misclassification are hidden Markov models. Here the observed data are states, assumed to be misclassification of the true, underlying states. As an extension to the simple multi-state model, the msm package can fit a general multi-state model with misclassification. For patient i, and observation time t_{ij} , observed states X^*_{ij} are generated conditionally on true states X_{ij} according to a misclassification matrix E. This is a $m \times m$ matrix, whose (r, s) entry is

$$e_{rs} = Pr(X^*(t_{ij}) = s|X(t_{ij}) = r) \tag{5.9}$$

which we assume to be independent of time t. When the exposure misclassification is binary then this matrix is completely determined by (SN, SP). Analogously to the entries of Q, some of the e_{rs} may be fixed to reflect knowledge of the diagnosis process. For example, the probability of misclassification may be negligibly small for non-adjacent states. Thus the progression through underlying states is governed by the transition intensity matrix Q, while the observation process of the underlying states is governed by the misclassification matrix E. Both Q and E can depend on accurately measured covariates but in this thesis we only consider the case where both do not.

5.4 Inference for Continuous-Time Hidden Markov Process

Consider now a hidden Markov model in continuous time. The true state of the model X_{ij} evolves as an unobserved Markov process. Observed data x^*_{ij} are generated conditionally from true states $X_{ij} = 1, 2, \ldots, m$ according to a set of distributions $f_1(x^*|\theta_1), f_2(x^*|\theta_2), \ldots, f_n(x^*|\theta_m)$ respectively, where θ_r is a vector of parameters for the state r distribution.

A type of EM algorithm known as the Baum-Welch or forward-backward algorithm is commonly used for hidden Markov model estimation in continuous time (Bureau et al 2000).

To develop the likelihood for a continuous time hidden Markov process we start at looking at each subject separately. The i^{th} subject's contribution to the likelihood is

$$\begin{aligned} L_i &= Pr(x_{i1}^*, \ldots, x_{im_i}^*) \\ &= \sum Pr(x_{i1}^*, \ldots, x_{im_i}^* | X_{i1}, \ldots, X_{im_i}) Pr(X_{i1}, \ldots, X_{im_i}), \end{aligned}$$

where the sum is taken over all possible paths of underlying states $X_{i1}, \ldots, X_{im_i}$ (Jackson 2009) . Assume that the observed states are conditionally independent given the values of the underlying states. Also assume the Markov property, $Pr(X_{ij}|X_{i,j-1}, \ldots, X_{i1}) = Pr(X_{ij}|X_{i,j-1})$. Then the contribution L_i can be written as a product of matrices by decomposing the overall sum in equation 5.11 into sums over each underlying state. The sum is accumulated over the unknown first state, the unknown second state, and so on until the unknown final state, so

$$\begin{aligned} L_i = & \sum_{X_{i1}} Pr(x_{i1}^*|X_{i1}) Pr(X_{i1}) \sum_{X_{i2}} Pr(x_{i2}^*|X_{i2}) Pr(X_{i2}|X_{i1}) \\ & \ldots \sum_{X_{im_i}} Pr(x_{im_i}^*|X_{im_i}) Pr(X_{im_i}|X_{im_{i-1}}), \end{aligned}$$

where $Pr(x_{ij}^*|X_{ij})$ is the misclassification probability density, in the binary case determined by (SN, SP). For general hidden Markov models, this is the probability density $f_{X_{ij}}(x_{ij}^*|\theta_{X_{ij}})$.

$Pr(X_{i,j+1}|X_{ij})$ is the $(X_{ij}, X_{i,j+1})$ entry of the Markov chain transition matrix $P(t) = (p_{ij}(t))_{1 \leq i,j \leq n}$ evaluated at $t = t_{i,j+1} - t_{ij}$. Let f be the vector with r^{th} element the product of the initial state occupation probability $Pr(X_{i1} = r)$ and $Pr(x_{i1}^*|r)$, and let $\mathbf{1}$ be a column vector consisting of ones. For $j = 2, \ldots, m_i$ let T_{ij} be the $n \times n$ matrix with (r, s) entry

$$Pr(x_{ij}^*|s) p_{rs}(t_{ij} - t_{i,j-1}). \tag{5.10}$$

Then the likelihood contribution for subject i is

$$L_i = f T_{i2} T_{i3} \ldots T_{im_i} \mathbf{1}. \tag{5.11}$$

5.5 Recreating the Exposure Path for Continuous Time Hidden Markov Processes

The most common method of reconstructing a continuous time hidden Markov chain is the Viterbi algorithm. The Viterbi algorithm is a dynamic programming algorithm for finding the most likely sequence of hidden states.

Originally proposed by Viterbi (1967), it is also described by Durbin et al. (1998) and Macdonald & Zucchini (1997). For continuous-time models it proceeds as follows. Suppose that a hidden Markov model has been fitted and a Markov transition matrix $P(t)$ and misclassification matrix E are known. Let $v_k(t_i)$ be the probability of the most probable path ending in state k at time t_i.

1. Estimate $v_k(t_1)$ using known or estimated initial-state occupation probabilities.

2. For $i = 1, \ldots, N$, calculate $v_l(t_i) = e_{l,X^*_{ti}} \max_k v_k(t_{i-1})P_{kl}(t_i - t_{i-1})$. Let $K_i(l)$ be the maximizing value of k.

3. At the final time point t_N, the most likely underlying state $\widehat{X}^*_N$ is the value of k which maximizes $v_k(t_N)$.

4. Retrace back through the time points, setting $\widehat{X}^*_{i-1} = K_i(\widehat{X}^*_i)$.

5.6 Adjusting for Misclassification: True Exposure Path Recreation in Continuous Time

To demonstrate the performance of adjusting for time-varying misclassification using the most-likely path through true exposure states, we conduct a simulation study under four cases. In each case the underlying time-varying exposure and misclassification are generated under the same conditions, then differing exposure-outcome models are developed and the effectiveness of the misclassification adjustment is evaluated through the estimation of transition and misclassification probabilities, the number of misclassified states and the estimation of the exposure-outcome model.

The R package ‘msm’ consists of functions for fitting general continuous time Markov and hidden Markov multi-state models to longitudinal data. Both Markov transition rates and the hidden Markov output process can be modeled in terms of covariates. A variety of observation schemes are supported, including processes observed at arbitrary times, completely-observed processes, and censored states. The package can estimate transition probabilities as well as misclassification probabilities from the observed data. This allows adjustment for misclassification to be done with only the observed misclassified data so a validation study is not needed. This package can also calculate the most likely path through hidden states using the Viterbi algorithm. This allows misclassification to be adjusted for by recreating the true exposure path of each individual in continuous time. The effectiveness of this adjustment is shown through the simulation study below.

5.6.1 Data Simulation

To demonstrate the performance of misclassification adjustment, 500 Monte Carlo samples were simulated as follows:

1. The total number of subjects is set to $N = 1,000$ and the number of exposure measurements on each subject is set to$n = 10$.
2. For each subject i generate a continuous time exposure Markov chain $X_i(t)$ for $t \geq 0$ where the first exposure measurement $X_i(0)$ is generated with equal probability of being exposed ($X_i(0) = 1$) or unexposed ($X_i(0) = 0$) and the transition intensities are:
   - ($q_{01} = 0.2, q_{10} = 0.3$).
3. Censor the continuous Markov chain by observing the process state of subject i at time points $t_{i1}, t_{i2}, \ldots, t_{i,n}$ to obtain the observation $X_i(t_{i1}), X_i(t_{i2}), \ldots, X_i(t_{i,n})$ (panel data where there is no information about the process between t_{ij}).
4. Generate the exposure-outcome model from $X_i(t_{ij})$ in two cases:
   - $Y_{ij} \sim N(X_i(t_{ij}), 0.1)$ for $i = 1, \ldots, N$ and $j = 1, \ldots, n$.
   - $Y_{ij} \sim N(X_i(t_{ij}) + 0.5X_i(t_{i,j-1}), 0.1)$ for $i = 1, \ldots, N$ and $j = 2, \ldots, n$.
5. Misclassify the Markov chain $X_i^*(t_{i1}), X_i^*(t_{i2}), \ldots, X_i^*(t_{i,n})$ with ($SN = 0.8, SP = 0.95$).
6. Estimate the transition intensities (q_{01}, q_{10}) and misclassification probabilities (e_{11}, e_{00}) using the R function msm.
7. Recreate the most likely exposure path $X_i^{est}(t_{i1}), X_i^{est}(t_{i2}), \ldots X_i^{est}(t_{i,n})$ using the Viterbi algorithm with the estimates obtained in previous step.
8. Consider two different exposure-outcome models:
   - $E(Y_{ij}|X_i(t_{i,1:j})) = \beta_1 X_i(t_{i,j})$
   - $E(Y_{ij}|X_i(t_{i,1:j})) = \beta_1 X_i(t_{i,j}) + \beta_2 X_i(t_{i,j-1})$

9. Fit each exposure-outcome model using

$$L(\beta) = \prod_{i=1}^{N} \prod_{j=1}^{n} f(Y_{ij}|d_i(t_{i,1:j})), \tag{5.12}$$

where $d_i(t_{1,1:j}) = X_i(t_{i,1:j})$, $X_i^*(t_{i,1:j})$, or $X_i^{est}(t_{i,1:j})$ when estimating β_{true}, $\beta_{misclassified}$, or $\beta_{estimate}$ respectively.

5.6.2 Simulation Results

The simulation can be broken into four cases. Cases are determined by the form of the linear outcome model (lagged term or no lagged term) and whether the model has been correctly specified. In each case the coefficient estimates for the exposure outcome model are presented based on the true data, misclassified data and estimated data. The simulation standard deviation from the 500 Monte Carlo samples and 95% Monte Carlo confidence intervals are also presented in the tables below. The percentage of misclassified states, before and after the most likely path recreation, are presented as well as estimates for transitions intensities (q_{01}, q_{10}), corresponding transition probabilities (p_{01}, p_{10}) and misclassification probabilities (e_{11}, e_{00}). The misclassification probabilities corresponding to (SN, SP) respectively. The transition probabilities are calculated using the matrix exponential of Q.

Case 1 .

True Model: $E(Y_{ij}|X_i(t_{i,1:j})) = X_i(t_{i,j})$
Assumed Model: $E(Y_{ij}|X_i(t_{i,1:j})) = \beta X_i(t_{i,j})$

Parameter	Param.est	Std. Dev.	95% CI
q_{01}	0.200	0.0152	(0.199, 0.201)
q_{10}	0.301	0.0282	(0.299, 0.304)
p_{01}	0.157	0.0094	(0.156, 0.158)
p_{10}	0.237	0.0168	(0.235, 0.239)
e_{11}	0.800	0.0214	(0.798, 0.802)
e_{00}	0.950	0.0607	(0.945, 0.955)

Table 5.1: Simulation results for continuous time hidden Markov parameters when ($SN = 0.8, SP = 0.95$), ($q_{01} = 0.2, q_{10} = 0.3$) and a sample of 1,000 subjects was taken with 10 observations per subject.

	% of States	Std. Dev.	95% CI
Misclassified	9.50	0.30	(9.47, 9.53)
Misclassified after Recreation	8.70	0.42	(8.66, 0.874)

Table 5.2: Comparison of the percentage of misclassified states before and after Viterbi path recreation when a sample of 1,000 subjects was taken with 10 observations per subject. The parameters used for this recreation are shown in Table 5.1.

Coeff.	Coeff.est	Std. Dev.	95% CI
β_{true}	1.000	0.0023	(0.9998, 1.0002)
$\beta_{estimate}$	0.820	0.0112	(0.819, 0.821)
$\beta_{misclassified}$	0.790	0.0071	(0.789, 0.791)

Table 5.3: Simulation results for misclassification adjustment in continuous time for linear outcome model with no lag term when $(SN = 0.8, SP = 0.95)$, $(q_{01} = 0.2, q_{10} = 0.3)$ and a sample of 1,000 subjects was taken with 10 observations each.

Case 2 .
True Model: $E(Y_{ij}|X_i(t_{i,1:j})) = X_i(t_{ij}) + 0.5X_i(t_{i,j-1})$
Assumed Model: $E(Y_{ij}|X_i(t_{i,1:j})) = \beta_1 X_i(t_{i,j})$

Parameter	Param.est	Std. Dev.	95% CI
q_{01}	0.201	0.0121	(0.201, 0.202)
q_{10}	0.301	0.0276	(0.299, 0.303)
p_{01}	0.158	0.0076	(0.157, 0.159)
p_{10}	0.237	0.0178	(0.235, 0.239)
e_{11}	0.800	0.0179	(0.798, 0.802)
e_{00}	0.951	0.0057	(0.950, 0.952)

Table 5.4: Simulation results for continuous time hidden Markov parameters when $(SN = 0.8, SP = 0.95)$, $(q_{01} = 0.2, q_{10} = 0.3)$ and a sample of 1,000 subjects was taken with 10 observations per subject.

	% of States	Std. Dev.	95% CI
Misclassified	9.48	0.30	(9.45, 9.51)
Misclassified after Recreation	8.64	0.38	(8.61, 8.67)

Table 5.5: Comparison of the percentage of misclassified states before and after Viterbi path recreation when a sample of 1,000 subjects was taken with 10 observations per subject. The parameters used for this recreation are shown in Table 5.4.

Coeff.	Coeff.est	Std. Dev.	95% CI
β_{true}	1.281	0.0049	(1.2806, 1.2814)
$\beta_{estimate}$	1.092	0.0212	(1.090, 1.094)
$\beta_{misclassified}$	1.015	0.0097	(1.014, 1.016)

Table 5.6: Simulation results for misclassification adjustment in continuous time for linear outcome model with misspecified no lag term when ($SN = 0.8, SP = 0.95$), ($q_{01} = 0.2, q_{10} = 0.3$) and a sample of 1,000 subjects was taken with 10 observations per subject.

Case 3 .

True Model: $E(Y_{ij}|X_i(t_{i,1:j})) = X_i(t_{ij}) + 0.5X_i(t_{i,j-1})$
Assumed Model: $E(Y_{ij}|X_i(t_{i,1:j})) = \beta_1 X_i(t_{i,j}) + \beta_2 X_i(t_{i,j-1})$

Parameter	Param.est	Std. Dev.	95% CI
q_{01}	0.200	0.0142	(0.199, 0.201)
q_{10}	0.296	0.0272	(0.293, 0.299)
p_{01}	0.158	0.0093	(0.157, 0.159)
p_{10}	0.233	0.0177	(0.231, 0.235)
e_{11}	0.798	0.0210	(0.796, 0.800)
e_{00}	0.950	0.0060	(0.949, 0.951)

Table 5.7: Simulation results for continuous time hidden Markov parameters when ($SN = 0.8, SP = 0.95$), ($q_{12} = 0.2, q_{21} = 0.3$) and a sample of 1,000 subjects was taken with 10 observations per subject.

	% of States	Std. Dev.	95% CI
Misclassified	9.50	0.33	(9.47, 9.53)
Misclassified after Recreation	8.68	0.42	(8.64, 8.72)

Table 5.8: Comparison of the percentage of misclassified states before and after Viterbi path recreation when a sample of 1,000 subjects was taken with 10 observations per subject. The parameters used for this recreation are shown in Table 5.7.

Coeff.	Coeff.est	Std. Dev.	95% CI
$\beta 1_{true}$	1.000	0.0031	(0.9997, 1.0003)
$\beta 1_{estimate}$	0.837	0.0163	(0.836, 0.838)
$\beta 1_{misclassified}$	0.712	0.0089	(0.711, 0.713)
$\beta 2_{true}$	0.499	0.0027	(0.4987, 0.4993)
$\beta 2_{estimate}$	0.479	0.0287	(0.476, 0.482)
$\beta 2_{misclassified}$	0.576	0.0081	(0.575, 0.577)

Table 5.9: Simulation results for misclassification adjustment in continuous time for linear outcome model with lag term when $(SN = 0.8, SP = 0.95)$, $(q_{12} = 0.2, q_{21} = 0.3)$ and a sample of 1,000 subjects was taken with 10 observations per subject.

Case 4 .

True Model: $E(Y_{ij}|X_i(t_{i,1:j})) = X_i(t_{ij})$
Assumed Model: $E(Y_{ij}|X_i(t_{i,1:j})) = \beta_1 X_i(t_{i,j}) + \beta_2 X_i(t_{i,j-1})$

Parameter	Param.est	Std. Dev.	95% CI
q_{01}	0.199	0.0126	(0.198, 0.200)
q_{10}	0.301	0.0241	(0.299, 0.303)
p_{01}	0.157	0.0084	(0.156, 0.158)
p_{10}	0.237	0.0157	(0.236, 0.238)
e_{11}	0.801	0.0195	(0.799, 0.803)
e_{00}	0.950	0.0058	(0.949, 0.951)

Table 5.10: Simulation results for continuous time hidden Markov parameters when $(SN = 0.8, SP = 0.95)$, $(q_{01} = 0.2, q_{10} = 0.3)$ and a sample of 1,000 subjects was taken with 10 observations per subject.

	% of States	Std. Dev.	95% CI
Misclassified	9.49	0.33	(9.46, 9.52)
Misclassified after Recreation	8.67	0.40	(8.63, 8.71

Table 5.11: Comparison of the percentage of misclassified states before and after Viterbi path recreation when a sample of 1,000 subjects was taken with 10 observations per subject. The parameters used for this recreation are shown in Table 5.10.

Coeff.	Coeff.est	Std. Dev.	95% CI
$\beta 1_{true}$	1.000	0.0025	(0.9998, 1.0002)
$\beta 1_{estimate}$	0.750	0.0163	(0.749, 0.751)
$\beta 1_{misclassified}$	0.707	0.0079	(0.706, 0.708)
$\beta 2_{true}$	0.000	0.0030	(-0.0003, 0.0003)
$\beta 2_{estimate}$	0.087	0.0275	(0.085, 0.089)
$\beta 2_{misclassified}$	0.238	0.0077	(0.237, 0.239)

Table 5.12: Simulation results for misclassification adjustment in continuous time for linear outcome model with misspecified lag term when ($SN = 0.8, SP = 0.95$), ($q_{01} = 0.2, q_{10} = 0.3$) and a sample of 1,000 subjects was taken with 10 observations per subject.

The results above show that we can accurately determine the estimates of transition probabilities and misclassification probabilities. This is done without any other information available besides the observed panel data so a validation study is not needed. When estimates for transition probabilities and misclassification probabilities are obtained then we can use the results of Ch.3 to see how this form of misclassification affects the coefficients. This allows us to have some intuition on how estimation of the exposure-outcome model will be affected and allows us to better interpret our results. It can also be seen that the Viterbi algorithm, supplied with the estimates of Q and E, is effective in reducing the number of misclassified states. This reduction allows the coefficient estimates of the exposure-outcome model to be more accurate and reduces the effect of misclassification. The adjustment is most effective in adjusting for misclassification when a misspecified model is fit including extra lagged terms. This is very beneficial because when model selection procedures are employed most procedures start with a saturated/partially saturated model and then remove covariates that are not significant. We see that in Case 4 when the exposure status $X(t_{n-1})$ had no association with the outcome variable, $\beta 2_{misclassified}$ was still large, indicating a relationship between $X(t_{n-1})$ and Y that did not exist. When the most likely true exposure path was recreated the effect of X_{n-1} dropped dramatically from $\beta 2_{misclassified} = 0.238$ to $\beta 2_{estimate} = 0.087$. This is very helpful for model selection and allows artificial associations that result from misclassification to be minimized.

Chapter 6

Conclusion and Future Work

In this dissertation we concentrate on time-varying exposure misclassification. We determine and characterize the bias that results from time-varying misclassification for various misclassification parameters and time-varying exposure parameters. We have also determined two separate, easily-implemented adjustment methods that allow for estimation of the misclassification parameters and exposure parameters while reducing the effect of misclassification. When potential measurement error on the time-varying exposure is not accounted for, statistical assessment of the impact of the exposure variable on a health related outcome is misleading. The direction in which the association between the actual but unobserved explanatory variable and the response is biased, unpredictable and substantial.

The bias that results from misclassification is determined by tabulating the discrete probability distribution of $Pr(X_{1:n}, X^*_{1:n})$ and then the one-to-one correspondence between the discrete probability distribution and all the combinations of the random variables $(X^*_{1:n})$ is used to determine the coefficients in $E(Y_n|X^*_{1:n})$. This allows us to determine how misclassified time-varying exposure measurements affect the exposure outcome association. This development is presented in Chapter 2. Code for this determination is presented in Appendix A. This calculation allows us to characterize the effect of misclassification on the estimation of the regression coefficients and model bias. This is done by determining the effect of misclassification while (SN, SP) vary as well as when the exposure switching probabilities, (ϕ_1, ϕ_2), vary. The results of this characterization are presented in Chapter 3. It can be seen that (SN, SP) and (ϕ_1, ϕ_2) cause the effect of misclassification to change in a complicated way and no general rule can be determined to describe how misclassification affects the association between the actual explanatory variable and the response. This characterization allows the effect of misclassification to be determined for certain values of the misclassification parameters. Therefore we can use this to develop intuition on how the exposure effect is attenuated.

To adjust for the effect of misclassification we use algorithms for discrete time hidden Markov chains and continuous time hidden Markov chains to try and determine the true exposure by recreating the most likely path through the unobserved true exposure states (Chapter 4 and 5). These algorithms are easy to use and are already implemented in standard software. By using inferential techniques for hidden Markov chains in continuous time we are able to estimate the misclassification parameters (SN, SP) as well the switching probabilities (ϕ_1, ϕ_2) with only the observed data and no validation study is needed. These parameter estimates allow us to use the misclassification characterization of Chapter 3 to determine how the coefficients in the exposure outcome model are affected. When these parameter values are determined we can also adjust for the effect of misclassification in a more direct way by determining the most likely exposure path using the Viterbi algorithm. We can then determine the exposure-outcome relationship with this recreated path. It is shown through simulation that this recreated path allows a more accurate exposure-outcome relationship to be estimated and diminishes the effect of misclassification.

One analogous technique that parallels the adjustment method used in this thesis is that of regression calibration. Regression calibration reconstructs X using X^* and then regresses Y on this reconstruction. In regression calibration Y is regressed on $E(X|X^*)$ (Carroll et al. 2006). Our adjustment method regresses Y on the $mode(X|X^*)$. In principal $E(X|X^*)$ could also be used in our analysis and possible advantages could be achieved.

The adjustment for misclassification presented in this paper can be implemented in a quick and easy way. It does not, however, remove the effect of misclassification entirely. We can see that when using the recreated exposure path the coefficient estimates of the exposure outcome model are still biased but they are much closer then the coefficients obtain using the observed misclassified data. The adjustment is most effective when the assumed model has more lagged terms than the true model. This allows proper model selection to be conducted by including many lag terms and then removing the ones that are not significant. The adjustment that we have proposed only makes use of the hidden Markov chain and techniques associated with hidden Markov theory. It only uses the observed data X^* and the outcome variable Y is not used to predict the true exposure status X. A more complete adjustment for misclassification would use a Bayesian framework that would include both X^* and Y to predict X. The use of Y might enable more accurate prediction of X and adjustment would be more efficient but it would also be much more complicated to implement. The adjustment presented in this dissertation can be easily used by clinicians and epidemiologists and is effective at reducing the effect of misclassification.

The misclassification adjustment in this thesis can be extended in many ways. The inferential procedure implemented in the R package 'Hidden-Markov' for discrete time hidden Markov chains can be extended to use longitudinal data so that accurate estimation of transition probabilities and misclassification probabilities can be calculated. This would allow for adjustment for misclassification to be done without a validation study in the discrete time case, as it is done in the continuous time case. In the continuous time case the function 'msm' allows transition probabilities to depend on covariates. Therefore, if there is reason to believe the exposure status of a subjects switches based on accurately measured covariates, this can be accounted for. Our adjustment method can also be easily extended to adjust for differential misclassification. This can be done by using the outcome variable Y as a covariate for misclassification probabilities. The R package 'msm' also allows for misclassification probabilities to depend on covariates so existing software has the ability to account for differential misclassification. If the outcome variable Y is binary then the hidden Markov adjustment is still effective but how misclassification affects the results is unknown. The measurement error problems arising from combinations of the above scenarios are worth exploring. Further research should be conducted to improve the validity of scientific findings in epidemiological studies.

Bibliography

Baum L. E. and Petrie T.(1966) Statistical inference for probabilistic functions of finite state Markov chains, *Annals of Mathematical Statistics* **37**: 1554-1563

Baum L. E., Petrie T., Soules G., and Weiss N. (1970). A maximisation technique occurring in the statistical analysis of probabilistic functions of Markov chains, *Annals of Mathematical Statistics*
41: 164-171.

Bureau A., Hughes J. P., and Shiboski S. C. (2000). An S-Plus implementation of hidden Markov models in continuous time. *Journal of Computational and Graphical Statistics*, **9**: 621-632.

Carroll, R. J., Ruppert, D., Stefanski, L.A. and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models*, Vol. 105 of *Monographs on Statistics and Applied Probability,* second edn, Chapman & Hall/CRC, Boca Raton.

Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998). *Biological sequence analysis*, Cambridge University Press.

Greenland, S. and Gustafson, P. (2006). Accounting for independent non-differential misclassification does not increase certainty that an observed association is in the correct direction, *American Journal of Epidemiology* **164**: 63-68.

Gustafson, P. (2004) *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustment,* Vol. 13 of *Interdisciplinary Statistics*, Chapman & Hall/CRC, Boca Raton.

Harte D. (2010). *HiddenMarkov* [Computer program].

Jackson C. (2009) *Multi-state modeling with R: the msm package*, Cambridge, United Kingdom.

Jackson, C.H. and Sharples, L.D. (2002). Hidden Markov models for the onset and progression of bronchiolitis obliterans syndrome in lung transplant recipients, *Statistics in Medicine*, **21**: 113-128.

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

Kalbfleisch, J., Lawless, J.F. (1985). The analysis of panel data under a Markov assumption, *Journal of the American Statistical Association* **80**: 863-871.

MacDonald, I.L. and Zucchini, W. (1997). *Hidden Markov and Other Models for Discrete-valued Time Series*, Chapman & Hall/CRC, Boca Raton.

Pan S. L. and Wu H. M. (2007). A Markov regression random-effects model for remission of functional disability in patients following a first stroke:A Bayesian approach, *Statistics in Medicine* **26**: 5335-5353

Satten, G.A. and Longini, I.M. (1996). Markov chains with measurement error: estimating the 'true' course of a marker of the progression of human immunodeficiency virus disease (with discussion), *Applied Statistics* **45**: 275-309.

Viterbi J. (1967). Error bounds for convolutional codes and an asymptotically optimal decoding algorithm, *IEEE Transactions on Information Theory* **13**: 260-269.

Zucchini, W. (2005). *Hidden Markov Models Short Course, 34 April 2005.* Macquarie University, Sydney.

Appendix A

R Code for Bias Determination

```
#defining the total number of random variables (X,X*), (SN,SP) and
$(\phi_1, \phi_2)$
n=8
SN<-0.8
SP<-0.95
swit1<-Switch1[k]
swit2<-Switch2[l]

# creating all combination of true data and misclassified data
com<-t(rep(0,n))
for(i in 1:n){
com<-rbind(com,t(combn(1:n, i, tabulate, nbins = n)))
}
d<-data.frame(com)

# creating all combination of misclassified data
com2<-t(rep(0, n/2))
for(i in 1:I(n/2)){
com2<-rbind(com2,t(combn(1:I(n/2), i, tabulate, nbins = I(n/2))))
}

# determining stationary distribution for Markov chain
marMat<-cbind(c(-swit1,1), c(swit2,1))
sol<-solve(marMat)%*%c(0,1)

# determining Pr(X, X*)
prob<-NULL
px4<-NULL

for(i in 1:2^n){
```

```
#determining $Pr(X)$
r<-NULL
r[1]<-sol[I(d[i,1]+1)]
for(j in 2:I(n/2)){
if(d[i,j-1]==0){
r[j]<-(1-abs(d[i,j]-d[i,j-1]))*(1-swit1)+abs(d[i,j]-d[i,j-1])*swit1
}
if(d[i,j-1]==1){
r[j]<-(1-abs(d[i,j]-d[i,j-1]))*(1-swit2)+abs(d[i,j]-d[i,j-1])*swit2
}
}
px14<-prod(r)

# determine $Pr(X^*|X)$
x<-com[i,1:I(n/2)]
xs<-com[i,I(n/2+1):n]
pStar<-prod(x*(SN^xs*(1-SN)^(1-xs)) + (1-x)*((1-SP)^xs*SP^(1-xs)))

# determing $Pr(X^*,X)$
prob[i]<-px14*pStar
}

# determing $Pr(X_4==1|X)$
x4<-NULL
xstar<-NULL
for(i in 1:length(com2[,1])){
x4[i]<-sum(prob[d[,4]==1 & d[,5]==com2[i,1] & d[,6]==com2[i,2] &
 d[,7]==com2[i,3] & d[,8]==com2[i,4] ])
xstar[i]<-sum(prob[d[,5]==com2[i,1] & d[,6]==com2[i,2] &
 d[,7]==com2[i,3] & d[,8]==com2[i,4] ])
}
Eprob<-x4/xstar
#Creating the one-to-one correspondence
mat<-data.frame(com2)
Emat2<-model.matrix(~ X1*X2*X3*X4, mat)
# Determining $\beta$
E2<-solve(Emat2)%*%$Eprob
row.names(E2)<-names(as.data.frame(Emat2))
E2
```