**JIRSS** (2004)
**Vol. 3, No. 1, pp 39-57**



# Parameter Identifiability Issues in a Latent Markov Model for Misclassified Binary Responses

**Rhonda J. Rosychuk[1], Mary E. Thompson[2]**

[1]Department of Pediatrics, University of Alberta, 9423 Aberhart Centre, Edmonton, Alberta, Canada T6G 2J3. (rhonda.rosychuk@ualberta.ca)
[2]Department of Statistics and Actuarial Science University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada N2L 3G1. (methompso@setosa.uwaterloo.ca)

**Abstract.** Medical researchers may be interested in disease processes that are not directly observable. Imperfect diagnostic tests may be used repeatedly to monitor the condition of a patient in the absence of a gold standard. We consider parameter identifiability and estimability in a Markov model for alternating binary longitudinal responses that may be misclassified. Exactly two distinct sets of parameter values are shown to generate the distribution for the data in a common situation and we propose a restriction to distinguishes the two. Even with the restriction, parameters may not be estimable. Issues of sampling and correct model specification are discussed.

## 1 Introduction

Diagnostic tests are often used repeatedly to monitor the condition of a patient. However, a diagnostic test may not perfectly reflect

Received: May 2003, Revised: October 2003
*Key words and phrases:* Estimability, hidden Markov models, identifiability, longitudinal data, misclassification.

the patient's true condition. This situation commonly arises in clinical medicine where gold standards may be invasive, expensive and/or dangerous. For example, individuals may be repeatedly infected with a parasite and the only way to detect the parasite may be through an imperfect assay. When such a test is not definitive, researchers and clinicians cannot be certain of the start or end of infection. This uncertainty can affect the discovery of the source of infection and the therapy delivered to the patient. Models need to incorporate this uncertainty and the identifiability and estimability of model parameters are important considerations for valid estimation and inference.

This paper focuses on the situation where a subject may alternate between two states (eg. uninfected/infected, remitting/relapsing) and each response collected over time may either be correctly classified or misclassified. The two states of the true process, labelled 0 and 1, are not directly observable. Instead, data are collected on an observable process thought to be related to the true unobservable process. The observable process may misclassify the state of the true process. We assume misclassification can occur in two ways: an observed 0 may truly be a 1 or an observed 1 may truly be a 0. Misclassification of longitudinal data arising from an alternating binary response has been addressed in the literature by Nagelkerke, Chunge and Kinoti (1990) (hereafter NCK), Cook, Ng and Meade (2000) and Rosychuk and Thompson (2003). Only one type of misclassification is allowed in NCK, where underlying true states are assumed to follow a Markov model with constant infection and cure rates. Cook, Ng and Meade (2000) proposed hidden Markov models for several diagnostic tests applied repeatedly over time which are discrete-time first order, second order and time-nonhomogeneous Markov. Rosychuk and Thompson (2003) considered the impact of misclassification on maximum likelihood transition probability estimates.

Problems in measurement error models and hidden Markov models include parameter nonidentifiability. Unknown parameters in a model are nonidentifiable if more than one set of parameter values gives the same distribution function for the observation. In measurement error problems, normal distributions are generally assumed for all continuous explanatory variables, leading to nonidentifiability of the regression parameters (Bekker, 1986). In practice, the problem of nonidentifiability is solved by incorporating supplementary data such as validation studies as well as by adding constraints to the parameters (see for example Corroll and Stefanski, 1990). In hidden

Markov models, since only a function of the state in a finite state Markov chain is observed, parameter identifiability is a basic problem (Ito, Amari and Kobayashi, 1992) and states may be labeled to correspond with increasing parameter estimates in order to ensure parameter identifiability. The model proposed by NCK is a special case of a hidden Markov model.

Identifiability of model parameters is an important consideration for valid estimation and inference. Inferences cannot be definitively drawn if two or more explanations of the process are indistinguishable. Additionally, estimability of model parameters is crucial for valid inferences. The related concepts of identifiability and estimability are separately described and investigated in the context of misclassified binary longitudinal responses. We propose a model with both directions of misclassification that includes the model of NCK as a special case. The model consists of two distinct parts: the misclassification part specifies probabilities related to supplementary information and the unobservable true process is modelled as a continuous-time Markov chain with covariates. Section 2 describes the model and Section 3 examines parameter identifiability. A simple restriction permits parameter identifiability; however, parameter estimability is not assured. Parameter estimability is defined and discussed in Section 4 in conjunction with the sampling interval and model specification. A parasitic infection data set demonstrates methodology.

## 2 Model

We first present the notation and model in the absence of covariate information (Section 2.1). Section 2.2 gives the likelihood function when transition rates depend on covariates. For certain misclassification probabilities, the model does not permit estimation. These cases are outlined in Section 2.3.

### 2.1 Misclassification and the True Process

Suppose the two–state process is a homogeneous Markov process and there are $n_i$ observations at discrete times $t_{i1} < \cdots < t_{in_i}$ for subject $i$, $i = 1, \ldots, I$. Let $\Delta t_{ij} = t_{ij} - t_{i,j-1}$, $j = 2, \ldots, n_i$, be the inter-observation times for subject $i$. The observed state, $O_{ij}$, takes the values 0 or 1 depending on the state determined by the observed,

possibly misclassified, response for subject $i$ at observation $j$. The state of the unobservable true process at time $t$ is denoted by $\xi_i(t)$ for subject $i$. At observation time $t_{ij}$, the state of the unobservable true process is labelled $\xi_{ij} = \xi_i(t_{ij})$.

The observed state is misclassified if it differs from the true state. Suppose there is some supplementary information, termed misclassification predictors, available at each observation time, which may help clarify the relationship between the observed and true process. The misclassification predictors could come from an auxiliary series such as clinical symptoms or other clinical measurements (eg. heart rate, blood pressure) that are collected at the same discrete observation times. For notational simplicity, we assume that only one misclassification predictor is available, taking value $C_{ij}$ for subject $i$ at time $t_{ij}$. We denote the history as $\mathcal{H}_i^{(j)} = (O_{i1}, C_{i1}, \ldots, O_{ij}, C_{ij})$ and $\mathcal{H}_i^{(0)}$ is defined as the empty set.

The two types of misclassification probabilities are $\text{pr}(O_{ij} = 1 \,|\, \xi_{ij} = 0, C_{ij}) = v_{01}(C_{ij})$ and $\text{pr}(O_{ij} = 0 \,|\, \xi_{ij} = 1, C_{ij}) = v_{10}(C_{ij})$ with the 'proper' classification probabilities defined as $v_{00}(C_{ij}) = 1 - v_{01}(C_{ij})$ and $v_{11}(C_{ij}) = 1 - v_{10}(C_{ij})$. We assume a logistic link between the misclassification probabilities and predictors,

$$\begin{aligned} v_{01}(C_{ij}) &= \frac{e^{\alpha_0 + \alpha_1 C_{ij}}}{1 + e^{\alpha_0 + \alpha_1 C_{ij}}} \\ v_{10}(C_{ij}) &= \frac{e^{\alpha_0^* + \alpha_1^* C_{ij}}}{1 + e^{\alpha_0^* + \alpha_1^* C_{ij}}} \end{aligned} \tag{1}$$

with misclassification probability parameters $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)$ and $\boldsymbol{\alpha}^* = (\alpha_0^*, \alpha_1^*)$. The misclassification probabilities depend on the misclassification predictor in exactly the same way for all subjects at all observation times.

Suppose $\rho > 0$ and $\eta > 0$ ($\rho \neq \eta$) are the rates of transition from true states 0 to 1 and 1 to 0, respectively. The Markov transition probabilities for the true process are

$$\begin{aligned} P_{01}(\Delta t_{ij}) &= \text{pr}(\xi_{ij} = 1 \,|\, \xi_{i,j-1} = 0) = \frac{\rho}{\rho + \eta} \left\{1 - e^{-\Delta t_{ij}(\rho + \eta)}\right\} \\ P_{10}(\Delta t_{ij}) &= \text{pr}(\xi_{ij} = 0 \,|\, \xi_{i,j-1} = 1) = \frac{\eta}{\rho + \eta} \left\{1 - e^{-\Delta t_{ij}(\rho + \eta)}\right\} \end{aligned}$$

with $P_{00}(\Delta t_{ij}) = 1 - P_{01}(\Delta t_{ij})$ and $P_{11}(\Delta t_{ij}) = 1 - P_{10}(\Delta t_{ij})$ for $j = 2, \ldots, n_i$, $i = 1, \ldots, I$. Let $\pi_1 = \text{pr}(\xi_{ij} = 1) = \rho/(\rho + \eta)$ denote the steady state probability and $\pi_0 = 1 - \pi_1$.

With misclassification, the observed state no longer has the Markov property. The probability of an observed state depends on the last observed state, on earlier states through $\gamma_{ij} = \mathrm{pr}(\xi_{ij} = 1 \,|\, \mathcal{H}_i^{(j)})$, and on the transition probabilities $P_{ab}(\Delta t_{ij})$, $a, b \in \{0, 1\}$. The probability of observing a 1 conditional on the past observed responses and misclassification predictors is

$$\begin{aligned}\mathrm{pr}(O_{ij} = 1 \,|\, \mathcal{H}_i^{(j-1)}, C_{ij}) = \\ v_{11}(C_{ij})\ \{(1 - \gamma_{i,j-1})\, P_{01}(\Delta t_{ij}) + \gamma_{i,j-1}\, P_{11}(\Delta t_{ij})\} \\ +\ v_{01}(C_{ij})\ \{(1 - \gamma_{i,j-1})\, P_{00}(\Delta t_{ij}) + \gamma_{i,j-1}\, P_{10}(\Delta t_{ij})\}\,.\end{aligned}$$

Calculating $\gamma_{ij}$ explicitly is difficult, whereas using Bayes' rule gives the recursive form

$$\begin{aligned}\gamma_{ij} &= \frac{\mathrm{pr}(O_{ij} \,|\, \xi_{ij} = 1, \mathcal{H}_i^{(j-1)}, C_{ij})\mathrm{pr}(\xi_{ij} = 1 \,|\, \mathcal{H}_i^{(j-1)})}{\sum_{b=0}^{1}\mathrm{pr}(O_{ij} \,|\, \xi_{ij} = b, \mathcal{H}_i^{(j-1)}, C_{ij})\mathrm{pr}(\xi_{ij} = b \,|\, \mathcal{H}_i^{(j-1)})} \\ &= \begin{cases} \frac{v_{10}(C_{ij})\,\delta_{ij}}{\delta_{ij}\{v_{10}(C_{ij}) - v_{00}(C_{ij})\} + v_{00}(C_{ij})} & \text{if } O_{ij} = 0 \\ \frac{v_{11}(C_{ij})\,\delta_{ij}}{\delta_{ij}\{v_{11}(C_{ij}) - v_{01}(C_{ij})\} + v_{01}(C_{ij})} & \text{if } O_{ij} = 1 \end{cases}\end{aligned}$$

where $\delta_{ij} = \gamma_{i,j-1}\{1 - P_{01}(\Delta t_{ij}) - P_{10}(\Delta t_{ij})\} + P_{01}(\Delta t_{ij})$ for $j = 2, \ldots, n_i$, and $\delta_{i1} = \pi_1$. The model proposed by NCK was based on constant transition rates and one non-zero misclassification probability, $v_{10} > 0$. Without covariates and misclassification predictors, our model reduces to the NCK model if $\alpha_0$ is set to $-\infty$.

## 2.2 Incorporating Covariates

Including covariate information is straightforward. The transition rates of the last section become $\rho(\mathbf{x}_i) = \exp(\mathbf{x}_i\boldsymbol{\beta}_\rho)$ and $\eta(\mathbf{x}_i) = \exp(\mathbf{x}_i\boldsymbol{\beta}_\eta)$ where $\boldsymbol{\beta}_\rho$ and $\boldsymbol{\beta}_\eta$ are vectors of regression parameters and $\mathbf{x}_i$ is a vector of baseline covariates for subject $i$. The formulas for $P_{ab}(\Delta t_{ij})$, $\pi_1$, and $\gamma_{ij}$ all now depend on $\mathbf{x}_i$, $a, b \in \{0, 1\}$. The likelihood function is written as

$$L(\boldsymbol{\Theta}) = \prod_{i=1}^{I} \mathrm{pr}(O_{i1} \,|\, C_{i1}, \mathbf{x}_i; \boldsymbol{\Theta}) \prod_{j=2}^{n_i} \mathrm{pr}(O_{ij} \,|\, \mathcal{H}_i^{(j-1)}, C_{ij}, \mathbf{x}_i; \boldsymbol{\Theta})$$

where $\boldsymbol{\Theta} = (\boldsymbol{\alpha}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}_\rho, \boldsymbol{\beta}_\eta)$.

### 2.3 Excluded Cases

Information cannot be gained about the transition rates for certain values of the misclassification probabilities. If $v_{01}(C_{ij})+v_{10}(C_{ij}) = 1$, then $v_{00}(C_{ij}) = v_{10}(C_{ij})$ and $v_{11}(C_{ij}) = v_{01}(C_{ij})$. The probability of the observed state is completely independent of the unobservable true state and the likelihood is flat with respect to the regression parameters. We will not consider such cases in the remaining sections.

## 3 Parameter Identifiability

Unknown parameters in a model are nonidentifiable if more than one set of parameter values gives the same distribution function for the observation. Section 3.1 identifies exactly two sets of parameter values that imply the same distribution. A parameter restriction given in Section 3.2 provides an easy way to distinguish between the two sets. The restriction is applied to the NCK data set. Section 3.3 identifies the situations where the restriction is not needed.

### 3.1 Two Sets of Parameter Values Imply the Same Distribution

Suppose both transition rates depend on the same set of covariates and both misclassification probabilities depend on the same set of misclassification predictors. The distribution for the observed data, and hence the likelihood function, will be the same for two distinct sets of parameter values $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$ if

$$\begin{aligned} \boldsymbol{\Theta}_1 &= (\boldsymbol{\alpha}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}_\rho, \boldsymbol{\beta}_\eta) \\ \boldsymbol{\Theta}_2 &= (-\boldsymbol{\alpha}^*, -\boldsymbol{\alpha}, \boldsymbol{\beta}_\eta, \boldsymbol{\beta}_\rho). \end{aligned} \tag{2}$$

These sets of parameter values highlight the model symmetry. If the state labels are interchanged, then the transition probabilities switch and misclassification probabilities become proper classification probabilities under the new state labels. The permutation of state labels as a source of nonidentifiability is typical of hidden Markov models (Ito, Amari and Kobayashi, 1992, MacDonald and Zucchini, 1997). In particular, $L(\boldsymbol{\Theta}_1) = L(\boldsymbol{\Theta}_2)$ and $\widehat{\boldsymbol{\Theta}}_1$ and $\widehat{\boldsymbol{\Theta}}_2$ will both be maximum likelihood estimates for the data. In fact, these two sets are the only distinct sets that will lead to the same distribution. The full details of the proof follow in the Appendix.

Examining features of the relationships between the misclassification predictors and the misclassification probabilities as well as the covariate and transition probability relationships easily shows (2). Under the two sets, the misclassification probabilities are

$$\begin{aligned} v_{01}(C_{ij};\boldsymbol{\Theta}_1) &= v_{11}(C_{ij};\boldsymbol{\Theta}_2) \\ v_{00}(C_{ij};\boldsymbol{\Theta}_1) &= v_{10}(C_{ij};\boldsymbol{\Theta}_2). \end{aligned} \tag{3}$$

The misclassification probabilities under $\boldsymbol{\Theta}_1$ become correct classification probabilities under $\boldsymbol{\Theta}_2$. The transition probabilities have the relationship

$$\begin{aligned} P_{01}(t,\mathbf{x}_i;\boldsymbol{\Theta}_1) &= P_{10}(t,\mathbf{x}_i;\boldsymbol{\Theta}_2) \\ P_{10}(t,\mathbf{x}_i;\boldsymbol{\Theta}_1) &= P_{01}(t,\mathbf{x}_i;\boldsymbol{\Theta}_2) \end{aligned} \tag{4}$$

and $\pi_1(\mathbf{x}_i;\boldsymbol{\Theta}_1) = 1 - \pi_1(\mathbf{x}_i;\boldsymbol{\Theta}_2)$. Additionally, it follows from (3) and (4), that $\gamma_{ij}(\mathbf{x}_i;\boldsymbol{\Theta}_1) = 1 - \gamma_{ij}(\mathbf{x}_i;\boldsymbol{\Theta}_2)$ and

$$\mathrm{pr}(O_{ij}|\mathcal{H}_i^{(j-1)}, C_{ij};\boldsymbol{\Theta}_1) = \mathrm{pr}(O_{ij}|\mathcal{H}_i^{(j-1)}, C_{ij};\boldsymbol{\Theta}_2)$$

for all $j$.

### 3.2 Parameter Restrictions for Identifiability

One of the two sets in (2) can be eliminated if the misclassification probabilities are restricted to the interval $[0, 0{\cdot}5)$. If $0 \leq v_{01}(C_{ij};\boldsymbol{\Theta}_1) < 0{\cdot}5$ and $0 \leq v_{10}(C_{ij};\boldsymbol{\Theta}_1) < 0{\cdot}5$, then $0{\cdot}5 \leq v_{01}(C_{ij};\boldsymbol{\Theta}_2) \leq 1$ and $0{\cdot}5 \leq v_{10}(C_{ij};\boldsymbol{\Theta}_2) \leq 1$. Hence, the misclassification probabilities under parameter set $\boldsymbol{\Theta}_2$ are not admissible. This restriction is common in error-in-variables problems where the imperfect explanatory variable is dichotomous (Bekker, Van Montfort and Mooijaart, 1991) and seems to be a reasonable assumption. The restriction forces $\alpha_0 + \alpha_1 C_{ij} < 0$ and $\alpha_0^* + \alpha_1^* C_{ij} < 0$. This is a weak assumption if the observed process has some history of successful use in accessing the hidden process.

Table 1 shows the results when the constraint is applied to the parasitic infection data provided in NCK. This data set contains weekly diagnostic tests for the presence ($O_{ij} = 1$) or absence ($O_{ij} = 0$) of *Giardia lamblia* in $I = 58$ Kenyan children. Each child had between 10 and 44 weekly observations (median=14), $\Delta t_{ij} = 7$ days, $j = 2, \ldots, n_i$, $i = 1, \ldots, 58$. The diagnostic test was based on direct stool smears. The states, uninfected ($\xi_{ij} = 0$) and infected ($\xi_{ij} = 1$)

by the parasite, are not directly observable and are first assumed to follow a continuous-time Markov process. Jackknife standard errors are provided where subjects are dropped from the data set one at a time (see page 154 of Wolter, 1985).

By taking the reciprocal of the transition rates, the average time spent in the uninfected and infected states are estimated to be 67 and 56 days, respectively. Both misclassification probabilities are estimated to be about 10% and both are significantly different than zero. Note that including both types of misclassification is provided here for illustration only: $v_{01} = 0$ based on the NCK's description that parasites cannot be detected in the stool smear if the subject is uninfected.

Table 1: Parasitic infection data under a restricted model.

| | Estimates (se) | | Estimates |
|---|---|---|---|
| $\beta_{\rho 0}$ | -4·169 (0·306) | $\rho$ | 0·015 |
| $\beta_{\eta 0}$ | -4·013 (0·300) | $\eta$ | 0·018 |
| $\alpha_0$ | -2·071 (0·253) | $v_{01}$ | 0·112 |
| $\alpha_0^*$ | -2·202 (0·327) | $v_{10}$ | 0·100 |
| | log-likelihood: | | -545·547 |

All calculations are performed by a C program (Rosychuk, 1999) using the linear algebra package developed by Stuber (1997) and a direction set (Powell's) method provided in Press et al. (1992) for derivative-free maximization. The algorithm is started at several different starting conditions and is stopped when the log-likelihood function fails to increase by more than $10^{-6}$ on one iteration.

### 3.3 When Parameter Restrictions are Unnecessary

The key to parameter nonidentifiability was the relationships identified in (3) and (4). These relationships do not hold if one of the misclassification probabilities is zero, if the misclassification probabilities depend on different sets of misclassification predictors or if the transition rates depend on different sets of covariates. Hence, for these cases parameter restrictions will not be required for parameter identifiability.

# 4 Data Considerations for Parameter Estimability

If the observed data does not provide enough information to distinguish parameter values under the model, then the likelihood function will be close to flat as the parameter values vary. We say unknown parameters are estimable if in the region of largest likelihood values, the likelihood function is not flat in any direction of the allowable parameter values.

The importance of the sampling interval is illustrated by a simulation study in Section 4.1. In particular, it is seen that with sufficiently frequent sampling all of the parameters are estimable, with data generated from the assumed model. In Section 4.2, a simulated data set provides an example which suggests how problems with parameter estimability may signal problems with the model. (Of course, parameter estimability does not guarantee correct model specification.) The last section focuses on the situations where estimability is likely.

## 4.1 Sampling Interval

If the sampling interval is too long relative to the sojourn times of the true states, and even if the misclassification probabilities are constrained for parameter identifiability, not all parameters are estimable.

If the sampling times are far apart with respect to the state transition rates, observations $2, \ldots, n_i$ are like observations from Bernoulli trials when stationarity is assumed. When covariate and misclassification predictor variables are not available to supplement the data, the probability the observed state is 1 becomes $\pi_1 v_{11} + \pi_0 v_{01}$. Since the number of subjects and observations are fixed, only one degree of freedom is available. Hence, the function $\pi = \pi_1 v_{11} + \pi_0 v_{01}$ is estimable but individually $\pi_1$, $v_{01}$ and $v_{10}$ are not. Individual parameters are nonestimable even when one of the misclassification probabilities is zero. When covariates and misclassification predictors are available, the situation is similar.

A small simulation study, without covariates and misclassification predictors, was performed to investigate the behaviour of estimates with different state sojourn times. We generated data with the same number of subjects and visit times as the NCK data set. The state sojourn times were generated from exponential distributions with means

of 4, 7, 14, 21 or 42 days and misclassification probabilities were set as either 0·02 or 0·10. These sojourn times corresponded to about 0·57, 1, 2, 3 and 6 times the length of the seven days between observations. The data sets were simulated in two separate parts: exponential sojourn times were randomly generated to form the subject histories and the true states were then misclassified according to a random number generator.

Selected results for $v_{01}$ and $v_{10}$ appear in Table 2. These results give the average maximum likelihood estimates for 20 simulated data sets per simulation setting combination with standard errors in parentheses. The constraint proposed in Section 3.1 is used here.

Table 2: Average estimated mean times (EMTs) and misclassification probability estimates for simulations of 20 full realizations generated with specified mean times (MTs) and misclassification probabilities. Standard errors for the average are in parentheses.

| Setting | | | | Estimate(se) | | | |
|---|---|---|---|---|---|---|---|
| MT0 | MT1 | $v_{01}$ | $v_{10}$ | EMT0 | EMT1 | $\hat{v}_{01}$ | $\hat{v}_{10}$ |
| 7 | 7 | 0.02 | 0.02 | 14.450 (2.757) | 14.563 (3.235) | 0.111 (0.030) | 0.102 (0.031) |
| 7 | 7 | 0.02 | 0.10 | 14.361 (2.836) | 12.895 (4.999) | 0.124 (0.029) | 0.080 (0.031) |
| 7 | 7 | 0.10 | 0.10 | 81.399(44.235) | 41.748(16.276) | 0.241 (0.038) | 0.127 (0.040) |
| 7 | 14 | 0.02 | 0.02 | 8.639 (0.639) | 18.908 (1.945) | 0.059 (0.024) | 0.043 (0.014) |
| 7 | 14 | 0.02 | 0.10 | 9.815 (1.322) | 15.209 (1.944) | 0.100 (0.031) | 0.085 (0.025) |
| 7 | 14 | 0.10 | 0.10 | 13.303 (2.136) | 26.859 (4.972) | 0.176 (0.039) | 0.110 (0.025) |
| 7 | 21 | 0.02 | 0.02 | 8.543 (0.489) | 26.377 (1.926) | 0.056 (0.023) | 0.033 (0.008) |
| 7 | 21 | 0.02 | 0.10 | 9.870 (1.461) | 26.195 (3.861) | 0.093 (0.032) | 0.079 (0.017) |
| 7 | 21 | 0.10 | 0.10 | 10.433 (1.013) | 29.901 (5.534) | 0.176 (0.040) | 0.079 (0.019) |
| 7 | 42 | 0.02 | 0.02 | 8.954 (0.545) | 47.034 (3.846) | 0.118 (0.033) | 0.019 (0.005) |
| 7 | 42 | 0.02 | 0.10 | 8.530 (1.085) | 36.362 (7.122) | 0.124 (0.038) | 0.048 (0.013) |
| 7 | 42 | 0.10 | 0.10 | 12.821 (3.405) | 59.078(18.011) | 0.203 (0.045) | 0.069 (0.012) |
| 14 | 14 | 0.02 | 0.02 | 15.064 (0.574) | 15.042 (0.569) | 0.029 (0.009) | 0.027 (0.007) |
| 14 | 14 | 0.02 | 0.10 | 18.789 (1.383) | 17.143 (1.279) | 0.057 (0.014) | 0.092 (0.018) |
| 14 | 14 | 0.10 | 0.10 | 16.736 (1.370) | 15.305 (1.353) | 0.126 (0.018) | 0.080 (0.019) |
| 14 | 21 | 0.02 | 0.02 | 15.113 (0.523) | 21.586 (0.801) | 0.034 (0.010) | 0.022 (0.006) |
| 14 | 21 | 0.02 | 0.10 | 15.348 (0.760) | 20.996 (1.428) | 0.049 (0.012) | 0.079 (0.014) |
| 14 | 21 | 0.10 | 0.10 | 14.238 (0.842) | 19.686 (1.173) | 0.091 (0.016) | 0.066 (0.010) |
| 14 | 42 | 0.02 | 0.02 | 13.407 (0.826) | 39.170 (3.008) | 0.041 (0.025) | 0.018 (0.005) |
| 14 | 42 | 0.02 | 0.10 | 14.597 (0.677) | 40.435 (2.852) | 0.063 (0.017) | 0.077 (0.008) |
| 14 | 42 | 0.10 | 0.10 | 12.587 (0.751) | 42.007 (4.171) | 0.054 (0.015) | 0.086 (0.012) |
| 21 | 21 | 0.02 | 0.02 | 20.715 (0.614) | 21.724 (0.756) | 0.023 (0.005) | 0.022 (0.007) |
| 21 | 21 | 0.02 | 0.10 | 21.794 (1.061) | 21.104 (1.012) | 0.026 (0.008) | 0.090 (0.013) |
| 21 | 21 | 0.10 | 0.10 | 20.508 (1.044) | 21.215 (1.077) | 0.074 (0.010) | 0.109 (0.010) |
| 21 | 42 | 0.02 | 0.02 | 23.107 (0.967) | 44.967 (2.969) | 0.032 (0.007) | 0.021 (0.004) |
| 21 | 42 | 0.02 | 0.10 | 22.117 (0.879) | 43.664 (2.737) | 0.025 (0.005) | 0.097 (0.005) |
| 21 | 42 | 0.10 | 0.10 | 21.995 (1.304) | 44.930 (3.392) | 0.098 (0.013) | 0.091 (0.008) |
| 42 | 42 | 0.02 | 0.02 | 46.060 (1.424) | 43.559 (1.371) | 0.022 (0.004) | 0.027 (0.003) |
| 42 | 42 | 0.02 | 0.10 | 40.861 (1.464) | 40.767 (1.644) | 0.024 (0.003) | 0.103 (0.007) |
| 42 | 42 | 0.10 | 0.10 | 44.848 (2.139) | 43.964 (2.080) | 0.096 (0.007) | 0.105 (0.008) |

When one of the sojourn times has a mean of 4 days, the mean times and misclassification probabilities are poorly estimated. When the sampling interval is larger than the mean of the state sojourn time, the estimates are quite variable from one simulated data set to

another and maximum likelihood estimates can be difficult to obtain. For simulations when the mean times are at least 21 days in each state, both state EMTs and the misclassification probabilities are well estimated. When one of the mean times is 14 days and the other is at least 14 days, the misclassification probabilities are not necessarily well estimated even if the state EMTs are. Since all of the inter-observation times are seven days, the results suggest that collecting data a minimum of three times during the occupancy of each state seems to enable estimation of the state rates as well as estimation of small misclassification probabilities for the simulation settings considered here. However, if point estimates close to the true mean times are preferred, more frequent sampling is required.

## 4.2 Misspecified Models

Problems with parameter estimability can indicate model misspecification. A simulated data set with the same observation times as the PI data set provides illustration of the estimability problems which can arise when the model is misspecified.

We generated data in a similar manner as described in Section 4.1 with gamma sojourn times instead of exponential. To simulate heterogeneous subjects, three different sojourn time distributions were used to represent a missed important covariate. All subjects had the same scales for each distribution, 0.01, but the shapes were allowed to differ to provide mean times of 10, 14 or 18 days. The three distributions and subsets were: subjects $i$=1,3-17,19 had shapes 0.14 and 0.14, subjects 20-37 had shapes 0.14 and 0.18, and subjects 41-58 had shapes 0.14 and 0.10 for states 0 and 1, respectively. Additionally, we set all observed states to be 0 for subjects 38-40 and all observed states to be 1 for subjects 2 and 18. The simulation misclassification probabilities were $v_{01} = 0.1$ and $v_{10} = 0.2$. Clearly, the generation of gamma distributed sojourn times and the heterogeneity violate the Markov specification of the true process.

We consider the results from one simulated data set which was relatively difficult for estimation. Applying the restricted model described in Section 3.2 yields estimates $\hat{\rho} = 0.0022$, $\hat{\eta} = 0.0030$, $\hat{v}_{01} = 0.1276$ and $\hat{v}_{10} = 0.2211$ with a log-likelihood of -511.634. These results give estimated mean times of 453.9 and 334.4 days for states 0 and 1, respectively. Since half of the subjects are on study for 98 days or less, these results exceed the study times and suggest

that most subjects would not have transitioned during the study. Estimates such as these would indicate to an investigator a parameter estimability problem. Despite the relatively frequent sampling, the likelihood is indeed almost flat in one direction in the parameter space. See Figure 1. Additionally, the observed and expected transition counts could highlight a model misspecification. For this particular simulated data the observed (expected) counts are 434 (328.2), 118 (224.3), 117 (224.3) and 261 (153.2) for the 0 to 0, 0 to 1, 1 to 0 and 1 to 1 transitions, respectively.

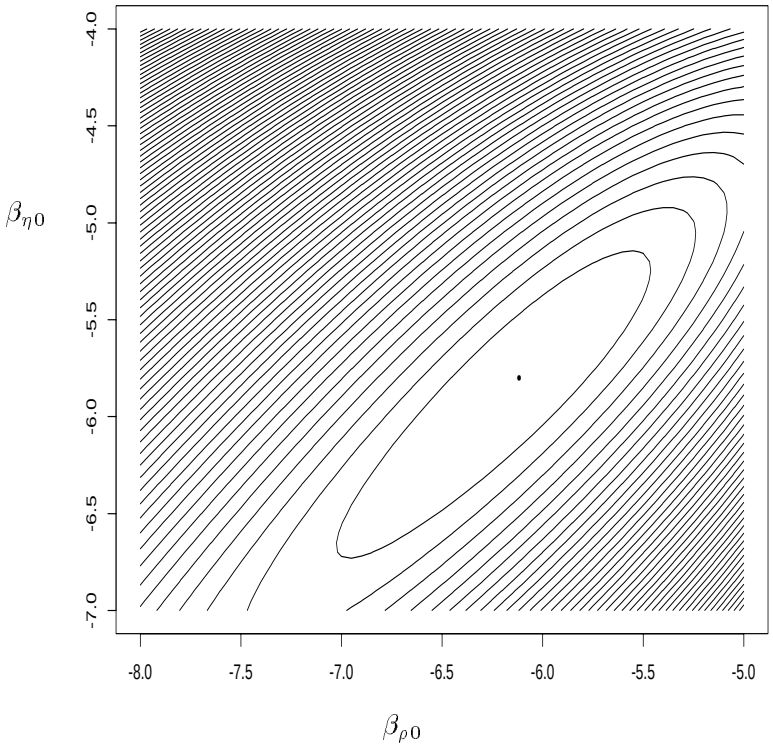


Figure 1: Log-likelihood contours for $\beta_{\rho 0}$ and $\beta_{\eta 0}$ when $v_{01} = 0.1276$ and $v_{10} = 0.2211$ for a simulated data set. Each contour represents a drop of 1 in the log-likelihood function from the MLE(*).

### 4.3 Properly Sampled and Specified Models

We shift our focus to situations when estimability is likely. Data reduction does not seem possible in the model proposed here. The transition counts are not sufficient statistics as in the case of Markov models, although we can consider these counts. The history and covariates will classify subjects into distinct groups and provided that these groups number at least one more than the number of parameters, estimability should not be a problem. *A fortiori*, if the transition counts together with the misclassification predictor pair counts and covariates themselves provide enough information for parameter estimability, then certainly the entire data set contains enough information to estimate all parameters.

Table 3: Expected counts for the misclassification predictor (MP) and observed process pairs under a stationarity assumption for $x = x_0$.

| | Observed Process | | | | |
|---|---|---|---|---|---|
| MP | 0-0 | 0-1 | 1-0 | 1-1 | Total |
| 0-0 | $N_{00}p_{11}$ | $N_{00}p_{12}$ | $N_{00}p_{12}$ | $N_{00}p_{14}$ | $N_{00}$ |
| 0-1 | $N_{01}p_{21}$ | $N_{01}p_{22}$ | $N_{01}p_{23}$ | $N_{01}p_{24}$ | $N_{01}$ |
| 1-0 | $N_{10}p_{21}$ | $N_{10}p_{23}$ | $N_{10}p_{22}$ | $N_{10}p_{24}$ | $N_{10}$ |
| 1-1 | $N_{11}p_{41}$ | $N_{11}p_{42}$ | $N_{11}p_{42}$ | $N_{11}p_{44}$ | $N_{11}$ |

Suppose we have only one covariate and a binary misclassification predictor. If we let $N_{ab}$ be the number of misclassification pairs where $C_{i,j-1} = a$, $C_{ij} = b$ and $x = x_0$, then the expected counts can be determined as in Table 3, $a, b \in \{0, 1\}$. Under a stationarity assumption, the probabilities of the observed pairs and misclassification pairs for a subject with covariate value $x = x_0$ can be easily calculated. The probabilities involve the cases when both observations are correctly classified, both observations are misclassified or only one of the observations is correctly classified and the other is misclassified. We denote these probabilities by $p_{de}$, $d, e = 1, 2, 3, 4$. Many of these probabilities are not distinct because

$$\pi_0(x_0)P_{01}(t, x_0) = \pi_1(x_0)P_{10}(t, x_0)$$

in the Markov model. Table 3 incorporates the identities among the $p_{de}$. Specifically, $p_{31} = p_{21}$, $p_{34} = p_{24}$, $p_{13} = p_{12}$, $p_{33} = p_{22}$ and $p_{43} = p_{42}$. Since $\sum_{e=1}^{4} p_{de} = 1$ for $d = 1, 2, 3, 4$, the number

of independent expected counts in the table is 7 and the number of (functionally independent) observed counts will be at least as large if all $N_{00}, N_{01}, N_{10}$ and $N_{11}$ are large enough. If the covariate has two levels, then the number of independent components of data available to estimate 8 parameters is 14. The parameters should be estimable in this case as well as other cases with more covariates and misclassification predictors.

Several steps can be taken to gain some insight if parameters are not estimable. Different parameterizations, such as $\pi$ in an earlier example, can be considered which may be estimable for a particular data set. Prior assumptions or ranges or prior distributions on some parameters may allow estimation of the remaining parameters. As Section 4.2 emphasized, attempting to fit an adequate model to describe the data is an important step in achieving estimability. Further, if calculating partial derivatives of the likelihood function is not prohibitive, a positive-definite Hessian matrix would support parameter estimability.

## 5 Discussion

Parameter identifiability and estimability issues have been investigated in a latent Markov model for possibly misclassified binary data. The transition rates and misclassification probabilities were allowed to differ for different covariate and misclassification predictor values, respectively. We proved that when the misclassification probabilities depend on the same set (possibly empty) of misclassification predictors and the transition probabilities depend on the same set of covariates, there are exactly two distinct sets of parameter values which yield the same distribution for the observations. These two sets arise from the symmetry of the model. If the labels for the states are interchanged, then the transition probabilities are interchanged and the misclassification probabilities become proper classification probabilities under the new state labels. Parameter identifiability, however, does not guarantee parameter estimability. We have also verified in a simulation example that all misclassification and transition probability parameters should be estimable as long as the model specification is correct, the sampling interval is frequent enough and simple range restrictions are reasonable.

We have considered identifiability as a property of the parameterization and estimability as a property of both the parameterization

and a particular data set. Establishing both parameter identifiability and estimability are crucial for drawing valid inferences.

## Acknowledgements

Research was supported by funding from the Natural Sciences and Engineering Research Council of Canada. Professor Rosychuk is an Alberta Heritage Population Health Investigator. The authors thank Professors J.F. Lawless and R.J. Cook, University of Waterloo, for helpful discussions and the anonymous referee for suggestions.



## References

Bekker, P. A. (1986), Comment on identification in the linear errors in variables model. Econometrica, **54**, 215–7.

Bekker, P. A., Van Montfort, K., and Mooijaart, A. (1991), Regression analysis with dichotomous regressors and misclassification. Statistica Neerlandica, **45**, 107–19.

Carroll, R. J. and Stefanski, L. A. (1990), Approximate quasi-likelihood estimation in models with surrogate predictors. Journal of the American Statistical Association, **85**, 652–63.

Cook, R. J., Ng, E. T. M., and Meade, M. O. (2000), Estimation of operating characteristics for dependent diagnostic tests based on latent Markov models. Biometrics, **56**, 1109–17.

Ito, J., Amari, S.-I., and Kobayashi, K. (1992), Identifiability of hidden Markov information sources and their minimum degrees of freedom. IEEE Transactions on Information Theory, **38**, 324–33.

MacDonald, I. L. and Zucchini, W. (1997), Hidden Markov and Other Models for Discrete-valued Time Series. New York: Chapman and Hall.

Nagelkerke, N. J. D., Chunge, R. N., and Kinoti, S. N. (1990), Estimation of parasitic infection dynamics when detectability is imperfect. Statistics in Medicine, **9**, 1211–9.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). Numerical Recipes in C: the art of scientific computing, Cambridge: Cambridge University Press, 2nd edition.

Rosychuk, R. J. (1999). Accounting for Misclassification in Binary Longitudinal Data, PhD thesis, University of Waterloo.

Rosychuk, R. J. and Thompson, M. E. (2002), Bias correction of two-state latent Markov process parameter estimates under misclassification. Statistics in Medicine, **22**, 2035–2055.

Stuber, J. L. (1997), The Virgil v0.7 sofware pakage. Sofware developed for M.Math. thesis. Available on-line from http://www.mathematicalsciences.com/virgil.

Waterloo Maple Inc. (1996), Maple V Release 4. Waterloo: Author.

Wolter, K. M. (1985), Introduction to Variance Estimation. New York: Springer-Verlag.

## Appendix: Exactly Two Distinct Sets of Parameter Values Imply the Same Distribution

By using characteristic functions, we can show that exactly the two distinct parameter sets given in (2) generate the same distribution for the data when the transition probabilities are not equal to 0 or 1.

Suppose that more than one set of parameter values gives the same distribution for the data. Let two distinct sets of parameter values, $\boldsymbol{\Theta}_1 = (\boldsymbol{\alpha}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}_\rho, \boldsymbol{\beta}_\eta)$ and $\boldsymbol{\Theta}_2 = (\widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{\alpha}}^*, \widetilde{\boldsymbol{\beta}}_\rho, \widetilde{\boldsymbol{\beta}}_\eta)$, generate the same distribution. The characteristic function for one subject with $n$ observations is

$$\varphi_n(s_n;\boldsymbol{\Theta}_1) = \sum_{\mathcal{K}_n} e^{\iota s_1 O_1 + \cdots + \iota s_n O_n} \times \operatorname{pr}(O_1|C_1, \mathbf{x};\boldsymbol{\Theta}_1) \prod_{j=2}^{n} \operatorname{pr}(O_j|\mathcal{H}_i^{(j-1)}, C_{ij}, \mathbf{x};\boldsymbol{\Theta}_1)$$

where $\mathcal{K}_n$ is the set of all possible binary sequences of length $n$, $\iota = (-1)^{\frac{1}{2}}$ and $s_n = (s_1, \ldots, s_n)$. If $\varphi_n(s_n;\boldsymbol{\Theta}_1) = \varphi_n(s_n;\boldsymbol{\Theta}_2)$, then $L(\boldsymbol{\Theta}_1) = L(\boldsymbol{\Theta}_2)$ for each possible sequence of the observed states.

Consider the simple case when $v_{01}(C_{ij}) = \exp(\alpha_0)/\{1+\exp(\alpha_0)\} = v_{01}$ and $v_{10}(C_{ij}) = \exp(\alpha_0^*)/\{1+\exp(\alpha_0^*)\} = v_{10}$ and only the first two observations are examined. A symbolic manipulator such as Maple (Waterloo Maple Inc., 1996) can be used to solve for $\rho(\mathbf{x})$ and $\eta(\mathbf{x})$ in $\varphi_2(s_2;\boldsymbol{\Theta}_1) = \varphi_2(s_2;\boldsymbol{\Theta}_2)$ as a function of the unknowns $v_{01}$, $v_{10}$, $\widetilde{v}_{01}$, $\widetilde{v}_{10}$, $\widetilde{\rho}(\mathbf{x})$, and $\widetilde{\eta}(\mathbf{x})$. The solutions are

$$\rho(\mathbf{x}) = \left[\log\left\{\frac{u_1\,\widetilde{\eta}(\mathbf{x}) + u_3\,\widetilde{\rho}(\mathbf{x})}{\widetilde{\rho}(\mathbf{x})(u_3+u_4-u_5)}\right\} + \log\left\{\frac{u_2\,\widetilde{\rho}(\mathbf{x}) + u_4\,\widetilde{\eta}(\mathbf{x})}{\widetilde{\eta}(\mathbf{x})(u_3+u_4-u_5)}\right\}\right] \times \frac{u_1\,\widetilde{\eta}(\mathbf{x}) + u_3\,\widetilde{\rho}(\mathbf{x})}{(t_2-t_1)\,\{\widetilde{\rho}(\mathbf{x}) + \widetilde{\eta}(\mathbf{x})\}\,u_5} + \frac{u_1\,\widetilde{\eta}(\mathbf{x}) + u_3\,\widetilde{\rho}(\mathbf{x})}{u_5} \tag{5}$$

and

$$\eta(\mathbf{x}) = \frac{u_4\,\widetilde{\eta}(\mathbf{x}) + u_2\,\widetilde{\rho}(\mathbf{x})}{u_1\,\widetilde{\eta}(\mathbf{x}) + u_3\,\widetilde{\rho}(\mathbf{x})} \times \rho(\mathbf{x}). \tag{6}$$

where $u_1 = v_{01} - \widetilde{v}_{01}$, $u_2 = v_{10} - \widetilde{v}_{10}$, $u_3 = \widetilde{v}_{10} + v_{01} - 1$, $u_4 = \widetilde{v}_{01} + v_{10} - 1$ and $u_5 = v_{01} + v_{10} - 1$.

Solutions of $\varphi_n(s_n;\boldsymbol{\Theta}_1) = \varphi_n(s_n;\boldsymbol{\Theta}_2)$ are also solutions of $\varphi_j(s_j;\boldsymbol{\Theta}_1) = \varphi_j(s_j;\boldsymbol{\Theta}_2)$ for any $j \in \{1, \ldots, n-1\}$. The only distinct parameter

sets satisfying $\varphi_3(s_3;\boldsymbol{\Theta}_1) = \varphi_3(s_3;\boldsymbol{\Theta}_2)$ are

$$\begin{aligned} v_{01} &= 1 - \widetilde{v}_{10} \quad \rho(\mathbf{x}) = \widetilde{\eta}(\mathbf{x}) \\ v_{10} &= 1 - \widetilde{v}_{01} \quad \eta(\mathbf{x}) = \widetilde{\rho}(\mathbf{x}) \end{aligned} \tag{7}$$

which can be found in an iterative manner by substituting (5) and (6) into $\varphi_3(s_3;\boldsymbol{\Theta}_1) = \varphi_3(s_3;\boldsymbol{\Theta}_2)$, solving for $v_{01}$ and $v_{10}$, and then substituting these functions back into (5) and (6). Since (7) must be true for a subject with any value of the covariates, then $\boldsymbol{\Theta}_2 = (-\boldsymbol{\alpha}^*, -\boldsymbol{\alpha}, \boldsymbol{\beta}_\eta, \boldsymbol{\beta}_\rho)$ which is exactly the solution given in (2).

Now consider the case of non-constant misclassification probabilities given in (1). For notational simplicity, assume that only one misclassification predictor is available and has value $C_j$ at time $t_j$, $j = 1, \ldots, n$. Under the two sets of parameter values, $\varphi_1(s_1;\boldsymbol{\Theta}_1) = \varphi_1(s_1;\boldsymbol{\Theta}_2)$ if and only if

$$\pi_1(\mathbf{x}) = \frac{u_3(C_1)\,\tilde{\pi}_1(\mathbf{x}) + u_1(C_1)\,\{1 - \tilde{\pi}_1(\mathbf{x})\}}{u_2(C_1)\,\tilde{\pi}_1(\mathbf{x}) + u_4(C_1)\,\{1 - \tilde{\pi}_1(\mathbf{x})\}} \times \{1 - \pi_1(\mathbf{x})\} \tag{8}$$

where the $u$'s are the same as defined above except they now depend on $C_1$.

Since the left-hand side of (8) does not depend on $C_1$, the right-hand side must also be independent of $C_1$. In particular, it must be the case that

$$\frac{u_3(C_1)\,\tilde{\pi}_1(\mathbf{x}) + u_1(C_1)\,\{1 - \tilde{\pi}_1(\mathbf{x})\}}{u_2(C_1)\,\tilde{\pi}_1(\mathbf{x}) + u_4(C_1)\,\{1 - \tilde{\pi}_1(\mathbf{x})\}} = w(\mathbf{x}) \tag{9}$$

where $w(\mathbf{x})$ does not depend on $C_1$. Suppose the misclassification predictor has at least two levels, where $c$ and $d$ are two of these levels. Since (9) is independent of the value of the misclassification predictor,

$$\frac{u_3(c)\,\tilde{\pi}_1(\mathbf{x}) + u_1(c)\,\{1 - \tilde{\pi}_1(\mathbf{x})\}}{u_2(c)\,\tilde{\pi}_1(\mathbf{x}) + u_4(c)\,\{1 - \tilde{\pi}_1(\mathbf{x})\}} = \frac{u_3(d)\,\tilde{\pi}_1(\mathbf{x}) + u_1(d)\,\{1 - \tilde{\pi}_1(\mathbf{x})\}}{u_2(d)\,\tilde{\pi}_1(\mathbf{x}) + u_4(d)\,\{1 - \tilde{\pi}_1(\mathbf{x})\}}.$$

Provided that $\mathbf{x}$ is not a constant, the terms involving the powers of $\tilde{\pi}_1(\mathbf{x})$ can be equated to give the two equations

$$\begin{aligned} &\{1 - \widetilde{v}_{01}(c) - \widetilde{v}_{10}(c)\}\,\{1 - v_{01}(d) - v_{10}(d)\} \\ &\qquad = \{1 - \widetilde{v}_{01}(d) - \widetilde{v}_{10}(d)\}\,\{1 - v_{01}(c) - v_{10}(c)\} \end{aligned} \tag{10}$$

$$\begin{aligned} &\{1 - v_{10}(d) - \widetilde{v}_{01}(d)\}\,\{\widetilde{v}_{01}(c) - v_{01}(c)\} \\ &\qquad = \{1 - v_{10}(c) - \widetilde{v}_{01}(c)\}\,\{\widetilde{v}_{01}(d) - v_{01}(d)\}. \end{aligned} \tag{11}$$

Certainly, these equations are satisfied with the distinct solution given in (2). We wish to consider what other values of the parameters yield equality. Solving (11) for $v_{01}(c)$ gives

$$\frac{\{1-\widetilde{v}_{01}(c)-v_{10}(c)\}v_{01}(d)+\{1-v_{10}(d)\}\widetilde{v}_{01}(c)-\{1-v_{10}(c)\}\widetilde{v}_{01}(d)}{1-v_{10}(d)-\widetilde{v}_{01}(d)} \tag{12}$$

in terms of the other unknowns. Substituting (12) for $v_{01}(c)$ in (10) and solving for $v_{01}(d)$ gives

$$v_{01}(d) = 1 - v_{10}(d)$$

using Maple. This solution implies $v_{01}(C_1) + v_{10}(C_1) = 1$, which is not an admissible case as mentioned in Section 2.3. Thus, the only distinct solutions are given in (2).