The International Journal of Biostatistics

Volume 6, *Issue* 1 2010 *Article* 23

A Stochastic EM Type Algorithm for Parameter Estimation in Models with Continuous Outcomes, under Complex Ascertainment

Maria Grünewald, *Stockholm University*
Keith Humphreys, *Karolinska Institutet*
Ola Hössjer, *Stockholm University*

Recommended Citation:
Grünewald, Maria; Humphreys, Keith; and Hössjer, Ola (2010) "A Stochastic EM Type Algorithm for Parameter Estimation in Models with Continuous Outcomes, under Complex Ascertainment," *The International Journal of Biostatistics*: Vol. 6: Iss. 1, Article 23.
DOI: 10.2202/1557-4679.1222
Available at: http://www.bepress.com/ijb/vol6/iss1/23

A Stochastic EM Type Algorithm for Parameter Estimation in Models with Continuous Outcomes, under Complex Ascertainment

Maria Grünewald, Keith Humphreys, and Ola Hössjer

Abstract

Outcome-dependent sampling probabilities can be used to increase efficiency in observational studies. For continuous outcomes, appropriate consideration of sampling design in estimating parameters of interest is often computationally cumbersome. In this article, we suggest a Stochastic EM type algorithm for estimation when ascertainment probabilities are known or estimable. The computational complexity of the likelihood is avoided by filling in missing data so that an approximation of the full data likelihood can be used. The method is not restricted to any specific distribution of the data and can be used for a broad range of statistical models.

KEYWORDS: ascertainment, stochastic EM algorithm, missing data, outcome-dependent sampling, genetic epidemiology

Author Notes: We are very grateful to Jonathan Prince for providing the AD data set and for valuable discussions. Keith Humphreys work was supported by the Swedish Research Council, grant number 523-2006-972, Maria Grünewald's work was supported by the Swedish Foundation for Strategic Research (SSF), grant number A3 02:129, and Ola Hössjer's work was supported by the Swedish Research Council, grant number 621-2005-2810.

1 Introduction

1.1 Complex ascertainment - background

Most standard statistical tools for analyzing data from observational studies assume that simple random sampling is used. Outcome dependent sampling may however increase study efficiency. The case-control design (Breslow, 1982), for example, has been widely used in epidemiology. An attractive feature of the design is that unbiased estimates of relative risks can be obtained by performing statistical analysis on the data using a logistic regression model, as if the data were from a prospective study. More complex sampling designs may further increase efficiency. In the two-stage case-control design some covariate information is recorded on all subjects included in a study (Stage 1) whilst other covariate information, e.g. more expensive covariates, is gathered only on a subset of samples (Stage 2); the probability that the subject is included in Stage 2 is dependent on Stage 1 covariates. There is a large literature dealing with how to analyze outcome dependent, and two-stage, samples when the outcome is categorical, using a pseudo or semi-parametric likelihood (Breslow and Cain, 1988, Breslow and Holubkov, 1997, Breslow and Chatterjee, 1999, Chen, 2003). In general, there is less written about how to deal with continuous outcomes under complex sampling designs, although some of the above mentioned literature does touch on the topic.

Outcome dependent sampling based on continuous outcomes is common in genetic epidemiology. In an ongoing study (www.biobanks.se/cardiovascular.htm) at the second author's institute individuals in the upper and lower tertiles of cholesterol distributions are selected from a cohort study of 60 year old men in Stockholm, for genotyping. This particular study has a *two-stage cohort* design and hence the study base is clearly defined. For such designs outcome variables Y are known for the entire cohort sample – unbiased estimation of regression parameters is possible and computationally straightforward via application of the EM algorithm. Often study bases are instead ill-defined, e.g. hospital-based studies; it is these study designs which we focus on in this article. One example is the genetic association study of type II diabetes described by Gu, Abulaiti, Ostenson, Humphreys, Wahlestedt, Brookes, and Efendic (2004) where sampling probabilities are directly dependent on continuous outcomes, which define the diabetes phenotype. Parameter estimation in regression models with continuous outcomes measured in such samples, i.e. obtained under outcome dependent sampling, will be biased unless the ascertainment scheme is accounted for. Another example of a study in which a continuous outcome is studied under a complex sampling scheme is provided by Prince, Zetterberg, Andreasen, Marcusson, and Blennow (2004), where association between variants in the ApoE gene and cerebrospinal fluid levels of Aβ42 (the 42-

amino acid fragment of β-amyloid) is studied in Alzheimer's patients and healthy controls. An analysis of the Alzheimer's disease data set is reported in Section 3.3.

Various inference procedures for complex ascertainment schemes have been proposed. We start by reviewing some of these.

1.2 Full likelihood approach

Although not fitting fully into the classical framework of missing data problems (Little and Rubin, 1987), ascertainment can still be viewed as a missing data problem. In missing data problems data are partitioned into observed data, $\mathbf{Z}^{obs}$, and missing data $\mathbf{Z}^{mis}$, and there is typically a well-defined set of subjects for which some variables have missing values, while there is partially complete information, $\mathbf{Z}^{obs}$, on remaining variables. In our setting all variables can be viewed as belonging to $\mathbf{Z}^{mis}$ when a subject is not ascertained. In more detail, we let $\mathbf{Z}^{com} = (Z_1, ..., Z_{n^{com}})$ denote the complete data set before ascertainment, where $Z_j = (X_j, Y_j)$ contains explanatory variables (X_j) and response variables (Y_j) of individual j. We assume that Z_j are independent and identically distributed, parameterized by θ and that either Z_j are completely observed ($A_j = 1$) or not observed at all ($A_j = 0$), where A_j is the ascertainment or sampling indicator of individual j. Individuals are sampled independently, with sampling probabilities not depending on data on other individuals, so that

$$P(A_j = 1|\mathbf{Z}^{com}, \theta) = P(A_j = 1|Z_j, \theta) \tag{1}$$

for all j. In all examples, we also assume that $P(A_j = 1|X_j, Y_j, \theta) = P(A_j = 1|Y_j)$, i.e. the probability of ascertainment is independent of θ and X_j given observed data for individual j. However, this restriction is not needed to define the method.

Put $\mathbf{A}^{com} = (A_1, \ldots, A_{n^{com}})$. The likelihood of complete data,

$$\begin{aligned} L^{com}(\theta, \mathbf{Z}^{com}, \mathbf{A}^{com}) =& P(\mathbf{Z}^{com}, \mathbf{A}^{com}|\theta, n^{com}) \\ =& \prod_{j=1}^{n^{com}} P(Z_j|\theta)P(A_j|Z_j) \propto \prod_{j=1}^{n^{com}} P(Z_j|\theta) \end{aligned} \tag{2}$$

is not known, since not all Z_j are observed. The full likelihood of observed data is obtained by integrating (2) over the missing variables n^{com} and Z^{mis}, weighted by their joint probability distribution. This means in particular that we must put a 'prior' distribution $Q_{n^{com}} = P(n^{com})$ on n^{com}, assumed not to depend on θ. Without loss of generality, we assume that the first n^{obs} individuals are ascertained and decompose $\mathbf{Z}^{com} = (\mathbf{Z}^{obs}, \mathbf{Z}^{mis})$ into observed data $\mathbf{Z}^{obs} = (Z_1, \ldots, Z_{n^{obs}})$ and missing

DOI: 10.2202/1557-4679.1222

data $\mathbf{Z}^{mis} = (Z_{n^{obs}+1}, \ldots, Z_{n^{com}})$. Then $\mathbf{A}^{com}$ becomes a known function of $\mathbf{Z}^{obs}$ and $\mathbf{Z}^{mis}$ and can be dropped in the notation. The full likelihood can be written

$$\begin{aligned} L(\theta; \mathbf{Z}^{obs}) &= P(\mathbf{Z}^{obs}|\theta) \\ &= \sum_{m=0}^{\infty} Q_{n^{obs}+m} \binom{n^{obs}+m}{m} \int_{n^{mis}=m} P(\mathbf{Z}^{obs}, \mathbf{Z}^{mis}|\theta, n^{mis}) d\mathbf{Z}^{mis} \\ &= \sum_{m=0}^{\infty} Q_{n^{obs}+m} \binom{n^{obs}+m}{m} (1-P_\theta)^m \prod_{j=1}^{n^{obs}} P(A_j = 1|Z_j) P(Z_j|\theta), \end{aligned} \tag{3}$$

where $n^{mis} = n^{com} - n^{obs}$ is the number of missing observations, i.e. the dimensionality of $\mathbf{Z}^{mis}$ and

$$P_\theta = P(A_j = 1|\theta) = \int_{Y_j} \int_{X_j} P(A_j = 1|Y_j) P(X_j, Y_j|\theta) dX_j dY_j. \tag{4}$$

A particular feature of the ascertainment problem is that n^{mis} is unknown, making (3) an intractable sum of terms.

Indeed, the intractability of (3) is in some sense explained by the fact that data are 'not missing at random' (NMAR, Little and Rubin, 2002), and sampling probabilities cannot be determined from $\mathbf{Z}^{obs}$ alone.

1.3 Conditioning on ascertainment

A second general approach for correcting for ascertainment is to base inference on observed data conditional on ascertainment, giving a conditional likelihood

$$\begin{aligned} L^{cond}(\theta; \mathbf{Z}^{obs}) &= \prod_{j=1}^{n^{obs}} P(Z_j|A_j = 1, \theta) = \prod_{j=1}^{n^{obs}} \frac{P(A_j = 1|Y_j) P(Z_j|\theta)}{P_\theta} \\ &\propto \prod_{j=1}^{n^{obs}} \frac{P(Z_j|\theta)}{P_\theta} = \frac{L^{naive}(\theta; \mathbf{Z}^{obs})}{P_\theta^{n^{obs}}}, \end{aligned} \tag{5}$$

where

$$L^{naive}(\theta; \mathbf{Z}^{obs}) = \prod_{j=1}^{n^{obs}} P(Z_j|\theta)$$

is the naive likelihood, not taking ascertainment into account, or equivalently, assuming $P(A_j|Z_j) \equiv 1$. However, this form of the likelihood also makes likelihood-based estimation computationally difficult. The computational problem arises when

(4) is intractable. For continuous Y some examples of such settings are: if X is continuous or a mixture of discrete and continuous variables, or if ascertainment probability is a continuous function of Y, which could be the case in size biased sampling, see Patil (2002). When analytical solutions are not available, methods for numerical integration, such as importance sampling, can be used. This approach is investigated here as a comparison to the SEM type algorithm described in Subsection 2.1.

An attractive approach to estimation for study designs where samples are drawn with different probabilities in different regions of the space of a continuous outcome, Y, is described by Zhou, Chen, Rissanen, Korrick, Hu, Salonen, and Longnecker (2007). They describe a semi-parametric empirical likelihood approach to analyze data that consist of both a simple random sample and supplement samples from strata that are presumed highly informative based on their values of Y. Features of the approach are that no parametric assumptions are required for covariates and that ascertainment probabilities are not required to be known or estimated. Often it is advantageous not to make parametric assumptions for covariates, although in genetics it can be advantageous (Chen and Chatterjee, 2007). Comparison with the approach by Zhou et al. (2007) is included in one of the examples in Section 4.

1.4 Retrospective likelihood approach

A third alternative to (3) and (5) is to use the retrospective likelihood $P(X|Y,A)$, utilizing the fact that ascertainment probabilities cancel out of the likelihood when A is independent of $X|Y$, leaving $P(X|Y,A) = P(X|Y)$. However, Y is typically not ancillary. This implies loss of information when conditioning on Y, and often the set of parameters describing the relationship between X and Y is not identifiable (Liang, 1983). Some, but not all of the parameters may however be identifiable. See Chen (2003) for a discussion of parameter identification for the general odds ratio function.

1.5 Summary of the paper

When inference is based on (3), it is useful to consider algorithms used in missing data problems, such as the Expectation Maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977) and its extensions. Wacholder and Weinberg (1994) used the EM algorithm to obtain Maximum Likelihood estimates in case-control studies with complex ascertainment. If calculating the expected complete data likelihood in the EM algorithm requires computationally demanding numerical integration,

DOI: 10.2202/1557-4679.1222

one way to side-step the problem is to simulate the missing data, and use the value of the observed sample mean instead of the calculated expectation. In this spirit we describe a novel approach for parameter estimation for data with ascertainment on one or more variables, when sampling probabilities are known/estimable. Ascertainment is allowed to depend on explanatory and/or outcome variables, which may be continuous and/or categorical. The algorithm is similar to a Stochastic EM (SEM) algorithm (Celeux and Diebolt, 1985). This approach is general in the sense that it is not restricted to a particular design. The SEM algorithm has been shown to be useful in a wide range of missing data problems such as time-to-event data with censoring data sets (Ip, 1994) and haplotype estimation (Tregouet, Escolano, Tiret, Mallet, and Golmard, 2004). The basic idea of the method is to regenerate the missing data. Combined with the observed data, an artificial random sample from the targeted population is then formed. Parameter estimates are obtained using using the Maximum Likelihood approach for the artificial random sample. The effect of randomness of the artificial sample on the model parameter estimates is reduced by averaging over estimates from repeated generations. The SEM algorithm is known to be more robust to poorly specified starting values than the deterministic EM algorithm (Gilks, Richardson, and Speigelhalter, 1996), which is a highly attractive feature in our setting. For further reading on the SEM algorithm see Gilks et al. (1996) and McLachlan and Krishnan (1997).

We will first, in Section 2.1, present the SEM type algorithm for use in the ascertainment problem as described above. Two other approaches, a data augmentation method due to Clayton (2003), and a method based on importance sampling are presented in Section 2.2 for comparison. Some examples are presented in Section 3. Analysis of the example data sets are presented in Section 4, evaluating the SEM type algorithm and comparing it with other methods. The results are discussed in Section 5.

2 Methods

2.1 A SEM type algorithm for the full likelihood approach

2.1.1 The SEM algorithm for incomplete data

Suppose interest is in estimating θ in a data set where some data are missing. If the incomplete data can be augmented to resemble complete data, an approximation of the full likelihood (3) can be used. The essential idea is to inflate the observed sample, using simulated observations.

The resulting algorithm is iterative, and can be summarized as follows:

1. Select a starting parameter value $\hat{\theta}_0 = \theta^*$. Put $i = 1$.
2. (Simulation-step)
   Simulate $N = 1$ set of missing data $\mathbf{Z}_i^{mis} = (Z_{i(n^{obs}+1)}, \ldots, Z_{in_i^{com}})$ $\sim \mathbf{Z}^{mis}|\mathbf{Z}^{obs}, \hat{\theta}_{i-1}$. Put $\mathbf{Z}_i^{com} = (\mathbf{Z}^{obs}, \mathbf{Z}_i^{mis})$ and compute the likelihood

$$
\begin{aligned}
&L^{appr}(\theta; \mathbf{Z}_i^{com}) \\
&= Q_{n_i^{com}} \binom{n_i^{com}}{n^{obs}} \prod_{j=1}^{n^{obs}} P(Z_j|\theta) P(A_j = 1|Z_j) \cdot \prod_{j=n^{obs}+1}^{n_i^{com}} P(Z_{ij}|\theta) P(A_{ij} = 0|Z_{ij}) \\
&\propto \prod_{j=1}^{n^{obs}} P(Z_j|\theta) \cdot \prod_{j=n^{obs}+1}^{n_i^{com}} P(Z_{ij}|\theta) \qquad (6)
\end{aligned}
$$

   which can be viewed as a Monte Carlo approximation of the full likelihood (3) of the observed data set, using a single imputed sample, and A_{ij} are the ascertainment indicators of Z_i^{com}.
3. (Maximization-step)
   Obtain new parameter estimates $\hat{\theta}_i = \arg\max_\theta L^{appr}(\theta; \mathbf{Z}_i^{com})$.
4. $i \to i+1$. If $i \leq B+I$, go to Step 2, otherwise compute

$$\tilde{\theta} = \sum_{i=B+1}^{B+I} \hat{\theta}_i / I,$$

   where B is the burn-in time and I the number of iterations after burn-in.

The required size of the chain, I, is determined both by the specific problem at hand and by how much extra variability is accepted. Gilks et al. (1996) suggest, in the setting of Markov chain Monte Carlo, that parallel chains are run, so that comparison can be made. For large I estimates from a converged SEM algorithm will be similar to estimates from a converged EM algorithm.

As mentioned above, the SEM algorithm can be viewed as an iterative single imputation method within a likelihood framework. By using $N = 1$ missing data sets per iteration, the ascertainment probabilities enter into the Simulation-step likelihood as a multiplicative constant, and hence can be dropped in the Maximization-step. As a result, regression and covariate distribution parameters can be estimated separately in the Maximization-step, as for a simple random sample.

2.1.2 Applying a SEM type algorithm to the ascertainment problem

We now discuss in more detail how to generate missing data in the simulation step of the SEM-algorithm. The non-ascertained component is considered missing and

DOI: 10.2202/1557-4679.1222

is imputed as $\mathbf{Z}_i^{mis}$ in Iteration i of the algorithm. It thus remains to specify how to simulate from $P(\mathbf{Z}^{mis}|\mathbf{Z}^{obs},\theta)$ for any parameter vector θ. Normally when the SEM algorithm is used to fill in missing data there is a fixed sample size and data are filled in for those observations where data are missing. Here we assume that the sample size, n^{com}, of the representative data, is not known, as it would be in a two-stage cohort design.

It follows from (1) that n^{mis} and $\mathbf{Z}^{obs}$ are conditionally independent given n^{obs}, and also that $\mathbf{Z}^{obs}$ and $\mathbf{Z}^{mis}$ are conditionally independent given n^{mis}. Hence

$$\begin{aligned} P(\mathbf{Z}^{mis}|\mathbf{Z}^{obs},\theta) &= P(n^{mis}|\mathbf{Z}^{obs},\theta)P(\mathbf{Z}^{mis}|n^{mis},\mathbf{Z}^{obs},\theta) \\ &= P(n^{mis}|n^{obs},\theta)\cdot P(\mathbf{Z}^{mis}|n^{mis},\theta) \\ &= P(n^{mis}|n^{obs},\theta)\cdot \prod_{j=n^{obs}+1}^{n^{com}} P(Z_j|A_j=0,\theta). \end{aligned} \tag{7}$$

Since the last product of (7) is obtained from the parametric model, it remains only to specify $P(n^{mis}|n^{obs},\theta)$. Let P_θ and Q_n be as in (3). It follows from Bayes' Rule that

$$\begin{aligned} P(n^{mis}=m|n^{obs}=k,\theta) &= P(n^{com}=k+m|n^{obs}=k,\theta) \\ &\propto P(n^{obs}=k|n^{com}=k+m,\theta)Q_{m+k} \\ &= \binom{m+k}{k} P_\theta^k(1-P_\theta)^m Q_{m+k} \end{aligned} \tag{8}$$

for $m=0,1,2,\ldots$ where the proportionality constant does not depend on m. In particular, with the improper 'prior' $Q_n = 1/n$ we obtain

$$n^{mis}=m|n^{obs}=k,\theta \sim \text{NegBin}(k,P_\theta). \tag{9}$$

Now (7)-(9) naturally give rise to a rejection algorithm for generating $\mathbf{Z}^{mis}$:

Simulate: Simulate data Z_j^{mis}, A_j^{mis}, $j=1,2,\ldots$ independently from $P(A,Z|\theta) = P(Z|\theta)P(A|Y)$ and stop when $|\{j; A_j^{mis}=1\}| = n^{obs}$.

Reject: Throw away the observations Z_j^{mis} with $A_j=1$ and keep those Z_j^{mis} with $A_j=0$.

Alternatively, given any Q_m, we may use a standard simulation technique for discrete random variables to first generate n^{mis} from (8), and then sample Z_j^{mis}, A_j^{mis}, $j=$

$1,2,\ldots$ until $|\{j;A_j = 0\}| = n^{mis}$. However, the rejection algorithm based on (9) is particularly appealing, since we obtain, in each simulation, an ascertained data set of correct size n^{obs}. In the rest of the paper, we will use (9).

2.1.3 Variance calculation

An approximation of the variance of $\tilde{\theta}$ can, according to Gilks et al. (1996), be computed by utilizing the property that the observed data likelihood in the EM algorithm can be specified in terms of the complete data likelihood (Louis, 1982), but replacing the theoretical mean and variance with bootstrap estimates (Efron, 1992). The bootstrap estimates are obtained as follows: Fill in the missing data with simulated data K times, using $\tilde{\theta}$, to obtain pseudo complete data sets $\mathbf{Z}_1^{com}, \mathbf{Z}_2^{com} \ldots \mathbf{Z}_K^{com}$. The observed information is

$$-l''(\theta, \mathbf{Z}^{obs}) = E_\theta[-l''(\theta, \mathbf{Z}^{com})] - Cov_\theta[l'(\theta, \mathbf{Z}^{com})], \tag{10}$$

where $l = logL$ is the log likelihood and the expectation and covariance are calculated over the K pseudo samples. The covariance matrix is then obtained by taking the inverse of the information matrix as usual. The variance is scaled with respect to n^{obs}.

2.2 Importance sampling and data augmentation

We now summarize two alternative strategies to obtain parameter estimates in data with non-random ascertainment. These approaches, in common with the SEM type algorithm, are simulation based and use Maximum Likelihood for estimation. Both approaches yield parameter estimates that can be viewed as Monte Carlo approximations of the Maximum Likelihood estimates obtained from the conditional likelihood (5).

2.2.1 Importance sampling

As mentioned above the difficulty in calculating the likelihood of the ascertained data lies in the integration of (5). Importance sampling (Hammersley and Handscomb, 1964) is a Monte Carlo method used for numerical integration. The basic idea is to sample from one distribution to obtain the expectation of another. This is advantageous for sampling efficiently but also when drawing samples from the target distribution is difficult. In general terms, for a random variable X which has density $f_1(x)$, the expectation of some function g of X can be written as

DOI: 10.2202/1557-4679.1222

$$\mu = E_{f_1}[g(X)] = \int g(x) f_1 dx = \int \frac{f_1}{f_2} g(x) f_2 dx = E_{f_2}[\frac{f_1}{f_2} g(X)]$$

for $f_2 > 0$ whenever the support of f_2 includes that of $g f_1 > 0$. This means that samples can be drawn from f_2 to obtain the expectation of $g(x)$ under f_1. We can apply the importance sampling technique to approximate (5). One way to implement importance sampling in this context is to draw observations from a distribution which has the same parametric form as the target distribution $P(Z|\theta)$, but in the place of θ, use naïve guesses of the values of θ, which we call θ^*, in analogy with the starting values for the SEM type algorithm. In this case $P(A = 1|\theta)$ is estimated by noting that

$$P(A = 1|\theta) = \int P(A = 1|Z) P(Z|\theta) dZ = \int [P(A = 1|Z) \frac{P(Z|\theta)}{P(Z|\theta^*)}] P(Z|\theta^*) dZ.$$

If we draw $\dot{N}$ observations from $P(Z|\theta^*)$ which we denote as $Z_1^*, \ldots, Z_{\dot{N}}^*$, we can estimate $P(A = 1|\theta)$ by

$$\hat{P}(A = 1|\theta) = \frac{1}{\dot{N}} \sum_{j=1}^{\dot{N}} P(A = 1|Z_j^*) \frac{P(Z_j^*|\theta)}{P(Z_j^*|\theta^*)}. \tag{11}$$

As a consequence, the contribution of individual i to the logarithm of the likelihood (5) is, up to a constant,

$$\log(P(Z_i|\theta)) - \log(P(A = 1|\theta)),$$

and can be approximated by replacing $P(A = 1|\theta)$ by (11), thereby obtaining

$$\log(P(Z_i|\theta)) - \log(\frac{1}{\dot{N}} \sum_{j=1}^{\dot{N}} P(A = 1|Z_j^*) \frac{P(Z_j^*|\theta)}{P(Z_j^*|\theta^*)}). \tag{12}$$

Since the approximation of the likelihood is expressed in terms of θ an approximation of the information matrix can be computed as minus the second derivative of the log likelihood as usual.

2.2.2 Data augmentation

Clayton (2003) derives an ascertainment corrected likelihood by using an analogy to the conditional likelihood for matched case-control data. The idea behind this approach is to simulate a number of ascertained *pseudo-observations* $Z_{i1}, \ldots, Z_{i\ddot{N}}$

for each real observation Z_i and use these in combination with the real data to build the likelihood. As in the importance sampling method the true parameter values θ are unknown and are substituted by guesses, θ^*. This means that $\{Z_{ij}\}_{j=1}^{\ddot{N}}$ are drawn from $P(Z_i|\theta^*, A_i = 1)$.

Given the pseudo-observations the log likelihood contribution of individual i can, up to a constant, be written as

$$\log(P(Z_i|\theta)) - \log(\sum_{j=1}^{\ddot{N}+1} \frac{P(Z_{ij}|\theta)}{P(Z_{ij}|\theta^*)}) \tag{13}$$

where $Z_{i,\ddot{N}+1} = Z_i$. It is easy to see that (13) equals the log conditional likelihood contribution of individual i. The reason is that the ascertainment probabilities then cancel out. Clayton shows that the derivative of (13) with respect to θ yields a score function which can be interpreted as a Monte Carlo approximation of the score function obtained from (5).

Since an expression for the likelihood is available, parameter estimates can be obtained using Maximum Likelihood. Variances of these estimates are obtained as usual by calculating the information matrix from the likelihood. The likelihood (13) is similar to the likelihood approximated with the importance sampler, (12), especially when ascertainment probabilities are $0/1$. The essential differences are that

- Data are drawn under ascertainment in (13), using the data augmentation method, while they were drawn from the population distribution in (12), using the importance sampler.
- The sum in the second term is over the pseudo-observations only in (12) while the real observation are also included in (13).
- In (13) a separate correction term for ascertainment is calculated for each real observation while in (12) $P(A = 1)$ is calculated only once.

The last of these differences means that while $\dot{N}$ pseudo-observations are produced in the importance sampler, for a sample size of n^{obs} real observations, $\ddot{N} \times n^{obs}$ pseudo-observations are produced in the data augmentation method.

2.3 Comparison of full and conditional likelihood

In Subsection 2.1 we introduced a new computational approximation of the full likelihood and in Subsection 2.2 we reviewed to approximations of the conditional

DOI: 10.2202/1557-4679.1222

likelihood. It is then of interest to compare the two likelihoods. It follows from (3) that

$$L(\theta;\mathbf{Z}^{obs}) \propto L^{naive}(\theta;\mathbf{Z}^{obs}) \sum_{m=0}^{\infty} Q_{n^{obs}+m} \binom{n^{obs}+m}{m} (1-P_\theta)^m.$$

Note that this holds since $P(A_j = 1|Z_j)$ is assumed known, and therefore is independent of θ. However, for the particular choice $Q_n = 1/n$ adopted in Subsection 2.2, it is possible to show (using e.g. the probability distribution of a negative binomial random variable) that

$$\sum_{m=0}^{\infty} Q_{n^{obs}+m} \binom{n^{obs}+m}{m} (1-P_\theta)^m = \frac{1}{n^{obs} P_\theta^{n^{obs}}}.$$

From (5) we obtain the conclusion

$$L(\theta) \propto L^{cond}(\theta). \tag{14}$$

That is, when $Q_n = 1/n$, the full and conditional likelihoods are equivalent. Thus our SEM type algorithm can also be viewed as a computational approximation of the conditional likelihood (1.5). However, the equivalence (14) does not hold for general 'priors' Q_n.

3 Examples

To illustrate the performance of the methods described above we will look at two simulated data examples, and one real data example. The simulated data examples are included to allow comparison of the results with true answers. Sensitivity to poorly specified starting values and ascertainment probabilities is investigated. The first simulation example is based on a univariate continuous outcome and the second simulation is based on a more complex example with a multivariate outcome. For this example we found that our method provides valid parameter estimates while the importance sampler and the data augmentation method fail when θ^* is poorly specified. Both examples are based on a single explanatory variable X.

The simulations are inspired by genetic epidemiology, where non-random ascertainment is widely used for the reason that genetic data have traditionally been more expensive to collect than response variable measurements. In particular the outcomes are thought of as traits representing the metabolic syndrome (Agardh, Ahlbom, Andersson, Efendic, Grill, Hallqvist, Norman, and Ostenson, 2003). The metabolic syndrome comprises many health related outcomes that can affect each other in complex ways. In our simulation studies we represent only simplified models of the metabolic syndrome, using outcome variables only to represent BMI and plasma glucose level.

In the examples ascertainment probabilities are assumed known. In reality, these quantities will usually have to be either estimated from data, or inferred from external sources, adding an extra source of uncertainty that has not been taken into account here. In the second real data example ascertainment probabilities are calculated from age stratified prevalences of Alzheimer's disease published in Fratiglioni, Grut, Forsell, Viitanen, Grafstrom, Holmen, Ericsson, Backman, Ahlbom, and Winblad (1991).

3.1 Simulation Model *i*

$$X(\text{Genotype}) \rightarrow Y(\text{BMI}) \rightarrow A \text{ (Ascertainment)}.$$

Figure 1: Data structure in simulation Model i.

For the first simulation model we use a single categorical covariate, X. The model is based on a genetics example where $X = (0\ 1\ 2)$ represents the genotypes (AA, Aa, aa) of a single nucleotide polymorphism (SNP) with alleles A and a and a minor allele frequency of $\exp(\beta_{0X})/(1+\exp(\beta_{0X})) \approx 0.2$ ($\beta_{0X} = -1.4$), so that genotypes AA, Aa and aa have approximate population frequencies 0.64, 0.32 and 0.04. The distribution of the univariate outcome, conditional on $X = x$ is Gaussian with mean $\beta_{0Y} + \beta_{XY} \times x$ and variance σ_Y^2. We use values $\beta_{0Y} = 24, \beta_{XY} = 4$ and $\sigma_Y = \sqrt{2}$, chosen so that Y loosely represents BMI. Individuals with a BMI of 30 or more are defined as *obese* (as according to the WHO definition). About 10 percent of the Swedish population in the ages of 25-64 have such a BMI according to the WHO MONICA project (Tolonen, Kari, and Ruokokoski, 2000). Ascertainment probabilities are dependent on outcome/phenotype values: $P(A|y \geq 30) = 1$, $P(A|y < 30) = 0.067$, giving approximately equal numbers of obese and non-obese subjects. Sampling is continued until the required total sample size is obtained (we based simulations on sample sizes of 300 and 3000). This sampling procedure is similar to the one used by Gu et al. (2004). The difference is that in our simulation subgroup sample sizes are not fixed, whereas in Gu et al. (2004) they are. In simulating data we generate samples with random subgroup sample sizes, since this corresponds directly to the way the data are analyzed. The asymptotic equivalence of estimators, whether the subgroup sample sizes are regarded as fixed or random, has been discussed by Breslow, Robins, and Wellner (2000) in the case-control setting. The simulation model can be represented graphically as depicted in Figure 1.

DOI: 10.2202/1557-4679.1222

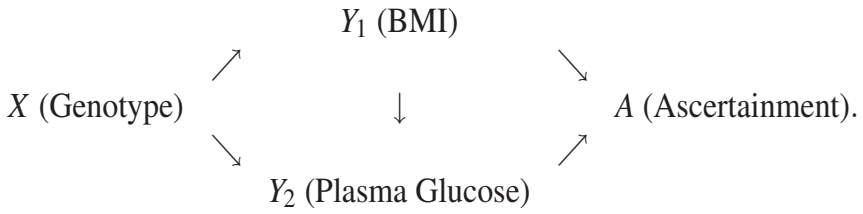

Figure 2: Data structure in simulation Model *ii*.

In this example, due to the simplicity of the model, evaluation of the integral (4) would actually be straightforward.

3.2 Simulation Model *ii*

In the second simulation we again assume a single categorical covariate, X, representing a SNP genotype. Instead of a single outcome as in Model *i*, we here use two, Y_1 and Y_2, considered to represent BMI and plasma glucose level, respectively. Obesity, measured in terms of BMI, is a *co-morbid* disease of plasma glucose level; BMI is dependent on genotype and, in turn, affects plasma glucose level. The genotype is assumed to have an additive effect on both outcomes, and Y_1 has an additive effect on Y_2. Given $X = x$, Y_1 has distribution $N(\beta_{0Y_1} + \beta_{XY_1} \times x, \sigma_{Y_1})$ and Y_2, given $X = x$ and $Y_1 = y_1$ has distribution $N(\beta_{0Y_2} + \beta_{XY_2} \times x + \beta_{Y_1Y_2} \times y_1, \sigma_{Y_2})$. Parameter values are chosen to represent outcomes accordingly; $\beta_{0Y_1} = 24$, $\beta_{XY_1} = 4$, $\sigma_{Y_1} = \sqrt{2}$, $\beta_{0Y_2} = 3$, $\beta_{XY_2} = 1$, $\beta_{Y_1Y_2} = 1/15$ and $\sigma_{Y_2} = 0.5$. The ascertainment probability is dependent upon both outcomes, as specified in Table 1. Model *ii* can be illustrated graphically as in Figure 2.

	$Y_1 < 30$	$Y_1 \geq 30$
$Y_2 < 7.8$	0.1	0.3
$Y_2 \geq 7.8$	0.3	1

Table 1: Ascertainment probabilities in Model *ii*.

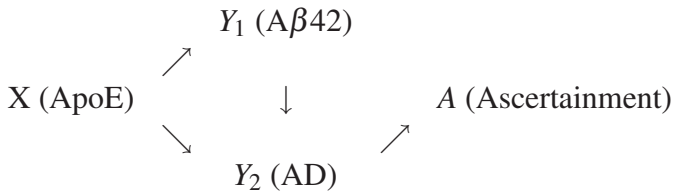

Figure 3: Data structure in Alzheimer's disease data example.

3.3 Alzheimer's disease data

Prince et al. (2004) examine the relationship between the ApoE gene, levels of Aβ42 in cerebrospinal fluid (CSF), and Alzheimer's disease (AD). Allele $\varepsilon 4$ of the ApoE gene is a well-documented risk factor for AD. Several studies have also found reduced levels of Aβ42 in CSF in AD patients. The relationship between ApoE and Aβ42 is however less documented. Prince et al. (2004) investigated the relationship in 563 AD patients and 118 healthy controls separately, and reported a statistically significant association between ApoE variants and Aβ42 levels in both groups. Under an assumption of rare disease, regressing Aβ42 levels on ApoE genotypes in controls provides approximately unbiased estimates of the effect of ApoE genotypes on Aβ42 levels in the population. Since, however, there are considerably more cases than controls there will be precision gain from incorporating all subjects in the regression analysis. To do so requires appropriate handling of the ascertainment scheme. Levels of Aβ42 in cerebrospinal fluid are thought to play a key role in mediating neurodegeneration in Alzheimer's disease and hence the causal relationship between ApoE variants, Aβ42 levels, AD and ascertainment is likely to be as depicted in Figure 3, where $Y_2 = 1$ for AD patients and $Y_2 = 0$ for healthy controls. We apply the SEM type algorithm to obtain ascertainment corrected estimates of the effect of ApoE genotypes on Aβ42 levels in the population, using both controls and cases from the Prince et al. (2004) study. The number of $\varepsilon 4$ alleles is here assumed to have an additive effect on both outcomes, and Y_1 has an additive effect on Y_2. Given $X = x$, Y_1 has distribution $N(\beta_{0Y_1} + \beta_{XY_1} \times x, \sigma_{Y_1})$ and Y_2, given $X = x$ and $Y_1 = y_1$, has distribution $Bin(1, p)$ where

$$p = \frac{\exp(\beta_{0Y_2} + \beta_{XY_2} \times x + \beta_{Y_1Y_2} \times y_1)}{1 + \exp(\beta_{0Y_2} + \beta_{XY_2} \times x + \beta_{Y_1Y_2} \times y_1)}.$$

Ascertainment probabilities are inferred from age specific prevalences of AD (Fratiglioni et al., 1991) to obtain AD expected prevalence in a population with an appropriate age distribution. Due to sparse information on prevalences for young

DOI: 10.2202/1557-4679.1222

subjects our analysis is restricted to subjects 75 years old or older, even though cases as young as 52 years old are available in the Prince et al. (2004) data set. We note that an inverse-probability- of-sampling weighted regression approach has recently been proposed for parameter estimation for the model described above, in the context of a genome-wide association scan for breast density, analysed as a secondary trait from breast cancer case-control samples (Monsees, Tamimi, and Kraft, 2009).

4 Results

All calculations were performed using the software R (The R Development Core Team, 2001). In the analysis below $\ddot{N} = 50$ is used in the data augmentation method and $\dot{N} = 30000$ is used in the importance sampler. Clayton (2003) points out that the information loss in the data augmentation method appear to be of the order $\ddot{N}/(\ddot{N}+1)$. Grünewald (2004) investigates the choice of $\ddot{N}$ and $\dot{N}$ in a simulation example similar to Model *i* described below, and conclude that the statement by Clayton (2003) holds, while the importance sampler needs a larger simulated data size.

4.1 Results for Model *i*

4.1.1 Parameter estimates and variance estimates when $\theta^* = \theta$

Estimates from the SEM type algorithm, the importance sampler and the data augmentation method using $\theta^* = \theta$ are presented in Table 2, as well as naïve estimates, calculated by optimizing the likelihood of the data without ascertainment correction. Estimates using the method by Zhou et al. (2007) and inverse probability-weighted (IPW) estimates are also presented in Table 2. The Zhou et al. (2007) estimates and IPW estimates were calculated using code provided by Zhou et al. (2007), which was modified slightly to fit Model *i*. The Zhou et al. (2007) estimates did not need specification of θ^*. Standard errors of the means of the estimates are presented in parentheses. For the SEM type algorithm we calculated the variance estimate of Gilks et al. (1996) based on (10) for each of the 1000 simulations, using $K = 5000$. The means of these standard errors across the 1000 simulations are presented in Table 3. For comparison standard errors based on observed variability in simulations are also presented. The standard errors by Gilks et al. (1996) appear to estimate this variability well. The standard errors calculated using (10) were

	True values	Naive estimates	Importance sampling	Data augmentation	SEM type algorithm	Weighted estimates	Zhou ODS estimates
$\hat{\beta}_{0X}$	-1.4	-0.107 (0.0028)	-1.394 (0.0023)	-1.394 (0.0023)	-1.393 (0.0022)	-	-
$\hat{\beta}_{0Y}$	24	24.426 (0.0044)	24.006 (0.0039)	24.004 (0.0039)	24.006 (0.0039)	24.000 (0.0044)	24.002 (0.0041)
$\hat{\beta}_{XY}$	4	4.141 (0.0031)	3.997 (0.0035)	3.998 (0.0034)	3.998 (0.0034)	4.004 (0.0052)	3.998 (0.0034)
$\hat{\sigma}_Y$	$\sqrt{(2)}$ ≈ 1.41	1.591 (0.0017)	1.411 (0.0018)	1.412 (0.0018)	1.410 (0.0017)	-	-

Table 2: Model *i*. Comparison of estimates when $\theta^* = \theta$. Results based on 1000 simulations with $n^{obs} = 300$, $\dot{N} = 30000$, $\ddot{N} = 50$ and $I = 2000$. Standard errors are reported in parentheses.

also used to construct 95% confidence intervals around the estimates obtained using the SEM type algorithm. The empirical coverage probabilities, based on 1000 simulations, were 0.96, 0.95, 0.95 and 0.95 for β_{0X}, β_{0Y}, β_{XY} and σ_Y respectively.

	Standard errors based on observed variability in simulations	Standard errors calculated using method in Gilks et al. (1996)
$\hat{\beta}_{0X}$	0.0022	0.0023
$\hat{\beta}_{0Y}$	0.0039	0.0039
$\hat{\beta}_{XY}$	0.0034	0.0035
$\hat{\sigma}_Y$	0.0017	0.0017

Table 3: Model *i*. Comparison of standard errors calculated using method in Gilks et al. (1996) and standard errors reflecting observed variation between simulations. $\theta^* = \theta$. Standard errors are calculated for mean estimates, based on 1000 simulations with $n^{obs} = 300$, $I = 2000$ and $K = 5000$.

The variability of the estimates appear to be similar for all methods except the IPW estimates, for which standard errors were larger. It is worth noting that the Gilks et al. (1996) method of calculating standard errors does not take into account the chain length of the SEM, so it is advisable to run a long chain to avoid variability that is unaccounted for. The chain length, I, in Model *i* was 2000.

Except naïve estimates, all methods compared provide estimates which are appropriately corrected for ascertainment. The bias in the naïve estimates was not very large in this specific example.

DOI: 10.2202/1557-4679.1222

4.1.2 Parameter estimation when θ^* is poorly specified

To investigate the behavior of the simulation based methods under poorly specified θ^* simulations were run for $\beta^*_{XY} = 0$ and 2, while remaining starting values were specified as their true parameter value counterparts. We first used sample sizes of 300. The running time of the SEM type algorithm was longer when θ^* was misspecified, to allow for convergence. As for Markov Chain Monte Carlo simulations, an appropriate burn in period has to be identified. When the algorithm has converged to a distribution around the parameter estimates, the standard errors of the estimates after burn in are the same as for correctly specified starting values. In our simulations the SEM type algorithm always converged and gave the same parameter estimates for poorly specified θ^* as for correctly specified θ^*, presented in Table 2.

When the data augmentation method was run with poorly specified θ^* parameter estimates were unbiased but had large standard errors, as can be seen in Table 4. As Clayton (2003) suggests, running a moderate amount of iterations of the data augmentation method improves the performance when θ^* is poorly specified. That is, the standard error estimates become smaller, approaching values that would be obtained if true/population parameter values were used as starting values.

In Table 4 parameter estimates and standard errors of mean estimates for the importance sampler are presented. The importance sampler yields incorrect parameter estimates. For example, when $\beta^*_{XY} = 2$ the mean estimate of $\hat{\beta}_{0X}$ was -1.164, with a standard error of 0.0091; for a true value of -1.4. The highly deviant results when $\beta^*_{XY} = 0$ were mainly driven by the results in four of the 1000 simulations. The inflation of the standard errors appears to be more pronounced in the importance sampler than in the data augmentation method. Since the importance sampler estimator is claimed to be unbiased it may seem surprising that the parameter estimates in the example are biased. A condition for the importance sampler is that the sampling distribution f_2 should be positive whenever $gf_1 > 0$. This condition is fulfilled in the simulations above, but when θ^* is misspecified f_2 may be so small in some regions where gf_1 is large, that no observations are actually sampled. The performance of the method may be improved by a better choice of sampling distribution, for example by using a mixture distribution (Hesterberg, 1995), or by iterating the choice of θ^*.

	True	$\beta^*_{XY} = \beta_{XY} = 4$		$\beta^*_{XY} = 2$		$\beta^*_{XY} = 0$	
		DA	IS	DA	IS	DA	IS
$\hat{\beta}_{0X}$	-1.4	1.394	-1.394	-1.386	-1.164	-1.320	4.15
		(0.0023)	(0.0023)	(0.0028)	(0.0091)	(0.0102)	(0.1222)
$\hat{\beta}_{0Y}$	24	24.004	24.006	24.001	23.838	24.002	36877.65
		(0.0039)	(0.0039)	(0.0043)	(0.0072)	(0.0046)	(29518.79)
$\hat{\beta}_{XY}$	4	3.998	3.997	4.008	4.316	4.079	-52721.59
		(0.0034)	(0.0035)	(0.0046)	(0.0122)	(0.0132)	(36968.86)
$\hat{\sigma}_Y$	$\sqrt{(}2) \approx 1.41$	1.412	1.411	1.411	1.41	1.409	41888.25
		(0.0018)	(0.0018)	(0.0021)	(0.0051)	(0.0029)	(25869.27)

Table 4: Model *i*. Data augmentation method (DA) and importance sampling (IS) with β^*_{XY} misspecified, and the remaining parameters at ideal starting values. Results based on 1000 simulations with $n^{obs} = 300$, $\dot{N} = 50$ and $\dot{N} = 30000$. Standard errors for mean estimates are reported in parentheses.

4.1.3 Incorrect specification of ascertainment probabilities

A simulation study was performed to investigate the effect of incorrect specification of ascertainment probabilities in the SEM type algorithm. Parameter estimates were calculated for different incorrect values $P^{assumed}(A|y < 30)$. For these simulations we used a sample size of 3000. The algorithm was run for 500 steps after a burn in of 50 steps. Results are summarized in Figure 4. The dashed lines represent naive estimates, $P^{assumed}(A|y < 30) = 1$, and the solid line is drawn at the correct parameter value. In this specific example estimates differed the most from true parameter values at $P^{assumed}(A|y < 30) = 1$.

4.1.4 Incorrect specification of error term distribution

In Model *i* we assume the error terms of Y to be Gaussian. There are multiple ways in which the error terms may be misspecified. We performed a simulation study investigating one type of misspecification, where the true error terms is a mixture of two Gaussian distributions with different variances. X was generated as in Model *i* and $Y|X$ was generated as follows: conditional on $X = x$ and s, Y has mean $\beta_{0Y} + \beta_{XY} \times x$ and variance $\sigma^2|s = (\sqrt{(}2) \times s)^2$ where s is stochastic with distribution $P(s = 1) = 0.5$ and $P(s = 6) = 0.5$. We simulated data with $\beta_{XY} = (1,2,3,4,5,6,7)$. We used ascertainment probabilities $P(A|y < 30) = 0.2$ and $P(A|y \geq 30) = 1$, and

DOI: 10.2202/1557-4679.1222

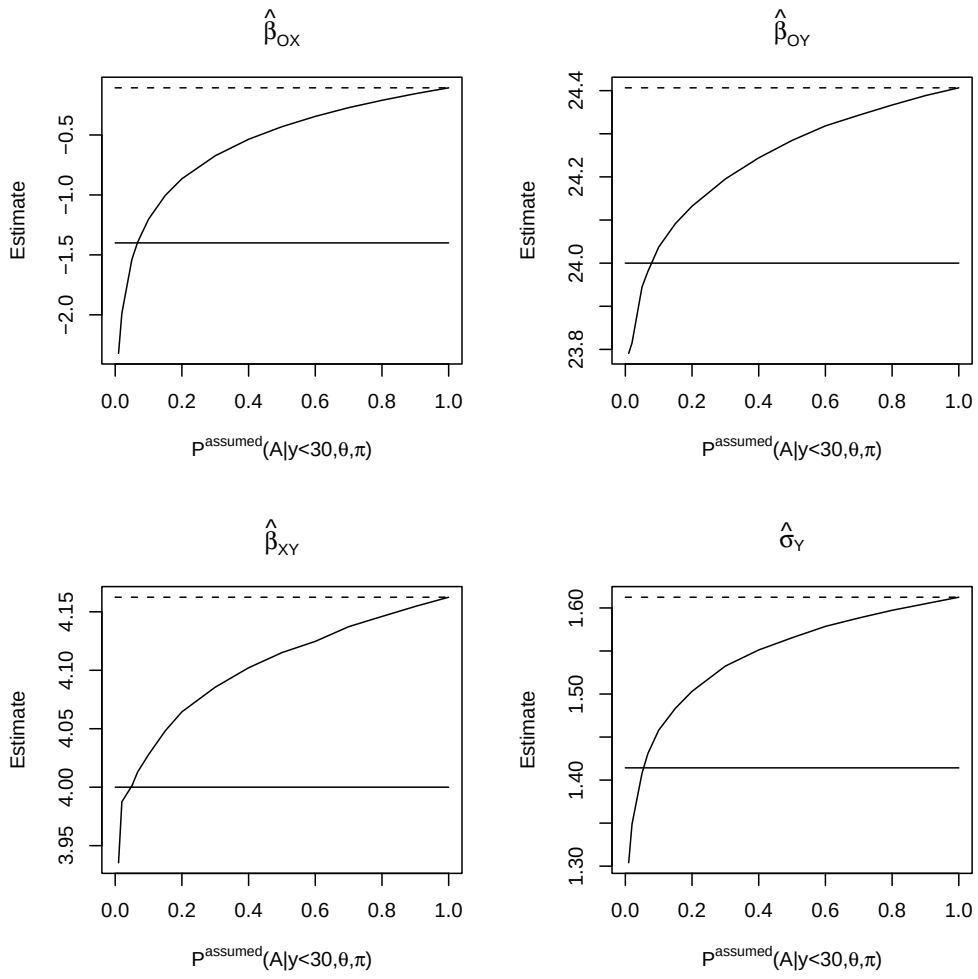

Figure 4: Model *i*. Effect of misspecification of ascertainment probability $P^{assumed}(A|y<30)$ when $P(A|y<30)=0.067$ in the SEM type algorithm. The naïve estimates are represented by dashed lines and the correct estimates are represented by solid lines. Estimates are based on a simulation with $n^{obs}=3000$.

a sample size $n^{obs} = 10.000$. Parameter estimates were obtained using the SEM type algorithm, with chain length $I = 2000$, and using the naive approach. For comparison, parameter estimates were also obtained for data generated under the assumed $Y|X$ Gaussian distribution, using $P(s = 1) = 1$ and $P(s = 6) = 0$. Results are presented in Figure 5. The estimates from the SEM type algorithm were biased when the error term distribution was misspecified. For most values of β_{XY} the size of this bias was smaller than the bias of the corresponding naive estimates, but somewhat larger than the bias of the naive estimates with the distribution for the error term distribution correctly specified.

4.2 Results for Model *ii*

When using $\theta^* = \theta$ for Model *ii* the importance sampler, the data augmentation method and the SEM type algorithm all gave reasonable estimates. The method by Zhou et al. (2007) was not used for Model *ii*, since it does not allow a multivariate outcome.

The methods were also used with poorly specified θ^*. The value of θ^* was ($\beta^*_{0X} = 0$, $\beta^*_{0Y_1} = \beta_{0Y_1}$, $\beta^*_{XY_1} = 0$, $\sigma^*_{Y_1} = \sigma_{Y_1}$, $\beta^*_{0Y_2} = \beta_{0Y_2}$, $\beta^*_{XY_2} = 0$, $\beta^*_{Y_1Y_2} = 0$, $\sigma^*_{Y_2} = \sigma_{Y_2}$). This value was chosen to investigate the robustness of the methods under extreme misspecification of θ^* in a complex model. As can be seen from Table 5 under this value of θ^* neither the data augmentation method nor the importance sampler obtain adequate parameter estimates. To investigate whether iterating the data augmentation method compensates for poorly specified θ^* a few exploratory runs were made. However, even after several iterations, the estimates behaved erratically, and we did not observe convergence towards true parameter values.

The SEM type algorithm did converge to appropriate parameter estimates but took longer to converge than it did in Model *i*. A run of the SEM type algorithm on a single data set is shown in Figure 6. Since the algorithm is run on a dataset, which in itself contains some uncertainty, convergence will be seen towards the estimated parameter values corrected for ascertainment, rather than towards the true/population parameter values.

4.3 Results for Alzheimer's disease data

The results of the analysis of the Alzheimer's disease data are presented in Table 6. We included 392 subjects in our analysis, of which 39 were controls and 353 were cases. From fitting a standard linear regression model, to controls only, we obtained an estimate of -164.99 (SE=52.81) for the (additive) effect of the ApoE $\varepsilon 4$ allele on A$\beta 42$ level. By including all 392 subjects and applying our SEM type algorithm,

DOI: 10.2202/1557-4679.1222

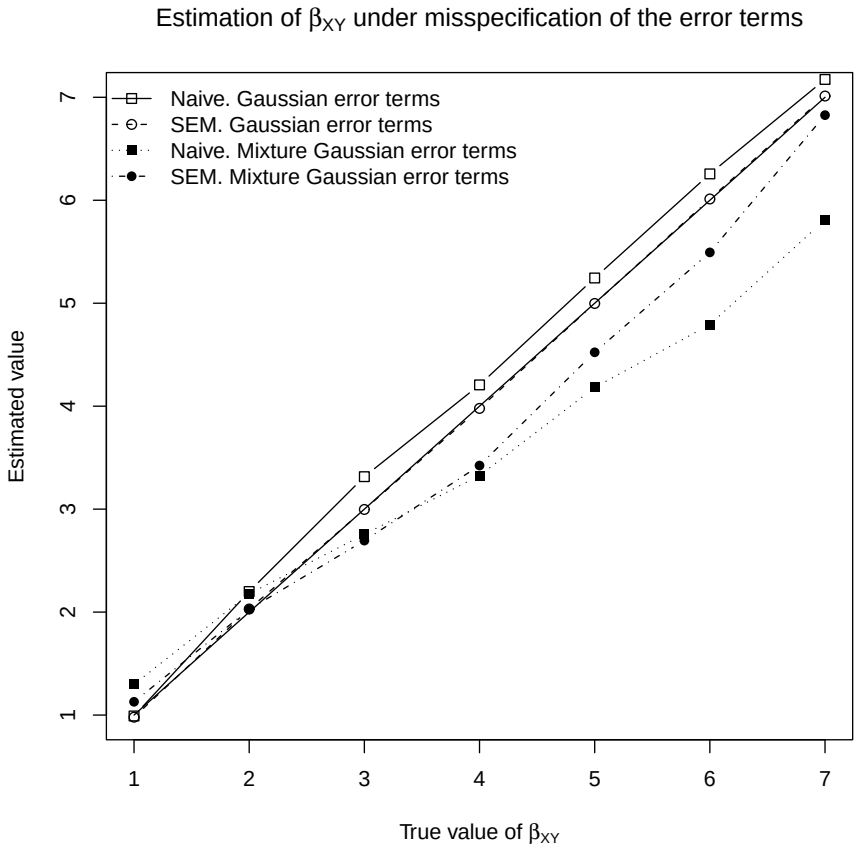

Figure 5: Model *i*. Estimation of β_{XY} under misspecification of the distribution of error terms in the SEM type algorithm, as described in Section 4.1.4. Estimates assume Gaussian error terms and are based on a simulation with $P(A|y < 30) = 0.2$, $n^{obs} = 10.000$ and $I = 2000$. Correct parameter values are represented by the 45° solid line.

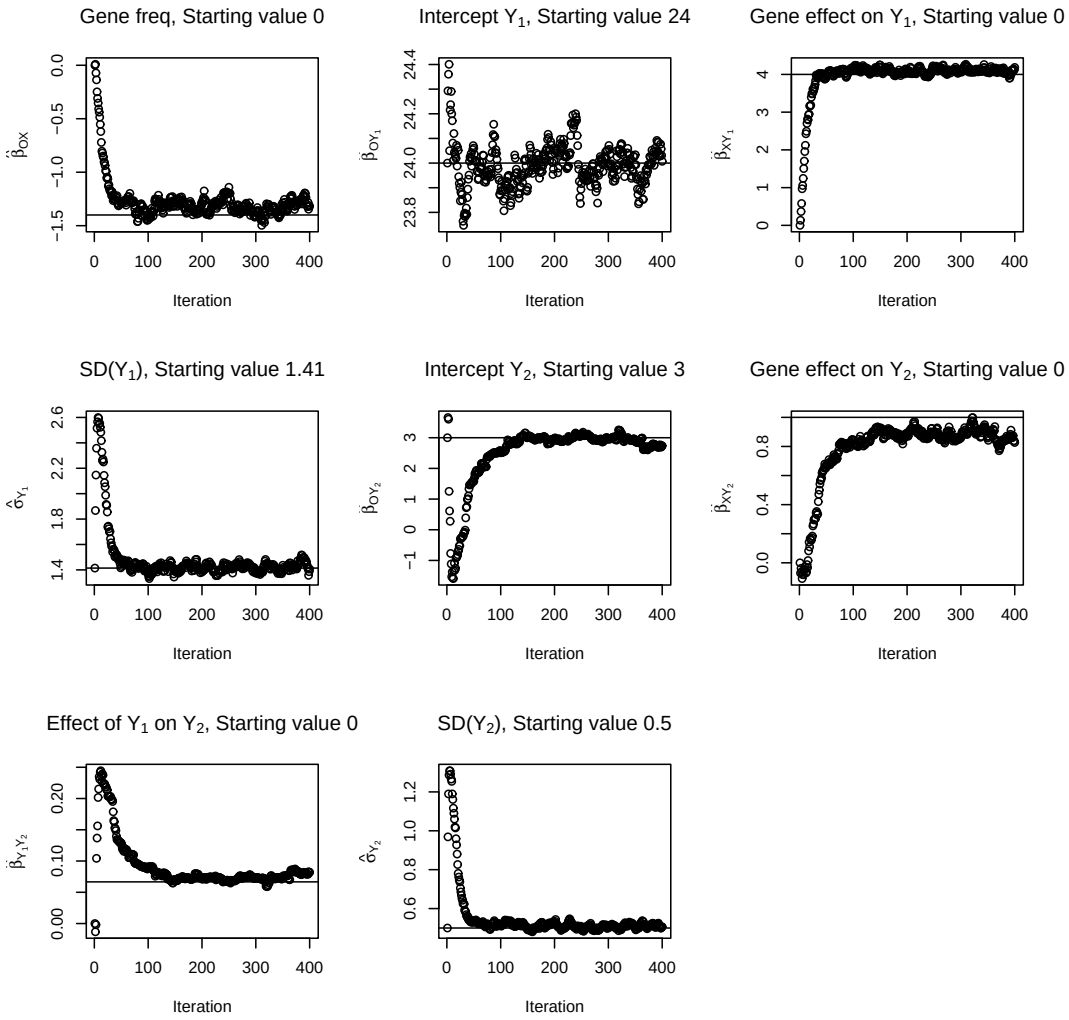

Figure 6: Model *ii*. The first 400 iterations in the SEM type algorithm for misspecified θ^*. True parameter values as solid lines. $n^{obs} = 300$.

DOI: 10.2202/1557-4679.1222

	True θ	θ^*	DA	IS
$\hat{\beta}_{0X}$	-1.4	0	-0.520	-1.040
			(0.036)	(0.034)
$\hat{\beta}_{0Y_1}$	24	24	23.660	24.427
			(0.036)	(0.028)
$\hat{\beta}_{XY_1}$	4	0	0.354	0.608
			(0.015)	(0.030)
$\hat{\sigma}_{Y_1}$	$\sqrt{(}2) \approx 1.41$	$\sqrt{(}2)$	1.744	1.668
			(0.005)	(0.013)
$\hat{\beta}_{0Y_2}$	3	3	4.828	4.982
			(0.008)	(0.016)
$\hat{\beta}_{XY_2}$	1	0	0.591	0.648
			(0.006)	(0.012)
$\hat{\beta}_{Y_1Y_2}$	$1/15 \approx 0.067$	0	0.062	0.015
			(0.001)	(0.001)
$\hat{\sigma}_{Y_2}$	0.5	0.5	0.071	0.004
			(0.003)	(0.001)

Table 5: Model *ii*. The data augmentation method (DA) and importance sampling (IS) under misspecified θ^*. Results based on 1000 simulations with $n^{obs} = 300$, $\dot{N} = 30000$ and $\ddot{N} = 50$. Standard errors for mean estimates are reported in parentheses.

using an ascertainment probability of $P(A = 1|Y_2 = 0) = 0.039$, we obtained an estimate of -106.56 (SE=13.98) for the (additive) effect of the ApoE $\varepsilon 4$ allele on A$\beta 42$ level (p-value=2.5×10^{-4} based on a normal approximation). As can be seen from Table 6 we were however not able to establish significant evidence of a direct effect of ApoE $\varepsilon 4$ on the risk of AD. This may be a result of restricting the age span to 75+ year olds, so that young AD cases, who may be highly informative about β_{XY_2}, were excluded from the analysis.

5 Conclusions

In this paper we have presented an algorithm that can be used to correct for ascertainment. The computational complexity of the likelihood under ascertainment is avoided by filling in missing data so that the full data likelihood can be used. An advantage of the method is that it is not restricted to any specific statistical model -some of the traditional methods to correct for ascertainment handle only specific sampling schemes/statistical models. Also, the complexity of the ascertainment

	Estimate	SE
β_{0X}	-0.931	0.1044
β_{0Y_1}	737.904	11.7508
β_{XY_1}	-106.565	13.9833
σ_{Y_1}	184.335	10.5104
β_{0Y_2}	5.045	0.5567
β_{XY_2}	-0.110	0.2056
$\beta_{Y_1Y_2}$	-0.011	0.0009

Table 6: Alzheimer's data example: Parameter estimates, and standard errors of estimates.

scheme hardly affects the complexity of the calculations, since the ascertainment probabilities are used only when simulating data, and not in the likelihood.

In order to illustrate some key points about performance of the algorithm we have kept the simulation examples fairly simple in terms of ascertainment schemes. The real data analysis was based on a model structure which has recently received attention in Genetic Epidemiology, for analyzing secondary traits from breast cancer case-control samples (Monsees et al., 2009). For other diseases (e.g. type II diabetes) it is desirable to study multivariate phenotypes (e.g. fasting insulin, fasting glucose, BMI) on which sample selection probabilities can be dependent. More complex ascertainment schemes may slow down the algorithm somewhat since the simulation of data under ascertainment is likely to involve a larger number of operations. The maximization step will however not be affected, as mentioned above. One possible complication with complex ascertainment schemes is that they may be harder to infer from external data than more simple schemes.

For well specified starting values the SEM type algorithm, as well as the two other simulation based methods investigated, perform well. For poorly specified starting values the SEM type algorithm seems to perform better than the other methods, when only a single iteration is used, both with regards to bias and to variability of the estimates.

The SEM algorithm may be sensitive to distributional assumptions on the response and explanatory variables, especially if the ascertainment probability is low, so that a large proportion of data are filled in. In the sensitivity analysis of specification of error term distribution in Section 4.1.4, the SEM estimates were biased, although less biased than the corresponding naive estimates. The outcomes in the examples are assumed to be normally distributed given genotype scores, but since real data often do not follow standard distributions, nonparametric extensions of the method would be of interest. The Aβ42 variable in the AD data set may benefit from a non-normal distributional assumption, or from a log-transformation. For categorical X, such as the genetic variables in our examples, specification of

DOI: 10.2202/1557-4679.1222

the distribution is unproblematic, while for continuous variables specification may be more difficult.

Whilst the distribution of the pseudo-complete data set can be checked using a standard QQ-plot (or similar), this may be misleading since the combined data are a mixture of data from the population distribution and data simulated according to the distributional assumptions. Deviations from normality, for example, are easier to detect if only the ascertained data are plotted, using an adjusted version of the QQ-plot. Some preliminary results regarding this adjustment exist, but are not presented here. If the ascertainment scheme is extreme, for example if only the tails of the distribution are sampled, so that $P(A = 1|Y = y) = 0$ for some y, distributional assumptions cannot be verified in such regions, and an analysis of sensitivity (to the distributional assumptions) is worthwhile. Also, the extremes of a distribution may accumulate outliers, which may distort the observed distribution. In such data one might consider to categorize the outcome to allow analysis using logistic regression, since not much information is lost by categorizing when the observed range of the continuous outcome is small. Note however that if the sampled data have a high proportion of outliers, the efficiency of the ascertainment scheme may also be reduced (Allison, Heo, Schork, Wong, and Elston, 1998).

The algorithms described in Section 2 require specification of sampling probabilities given the data. These probabilities are often not known and approximations may have to be made using, for example, registry data or prior knowledge about disease occurrence. For ill-defined study designs sensitivity analysis may be informative, with ascertainment probabilities as sensitivity parameters. A sensitivity analysis for the SEM type algorithm was performed on simulated data in Section 4.1.3. This analysis indicates that misspecification of ascertainment probabilities may indeed bias the estimates. A small simulation study, not included here, indicates that the sensitivity of the data augmentation method to misspecification of ascertainment probabilities is similar to what is shown for the SEM method. An alternative strategy for the SEM type algorithm is to maximize L^{appr} for each of a number of different sampling probability functions. Then the maximum of these maxima yields the final parameter estimate. This approach is more robust towards misspecification of sampling probabilities, when identifiable. A disadvantage though is that the variance calculations of Subsection 2.1.3 are complicated.

The estimator of Zhou et al. (2007) is attractive since it does not require specification of any covariate distribution nor of any sampling probabilities. It would be interesting to compare the two approaches for more simulated and real data sets.

An interpretation of the equivalence of the conditional and full likelihood when $Q_n = 1/n$ is that this choice of Q_n corresponds to no prior information about n^{comp}. When additional information about n^{comp} is available the full likelihood will

differ from the conditional one. An extreme case, $Q_n = 1_{n=n^{comp}}$, occurs when n^{comp} is known. In Grünewald and Hössjer (2010) the efficiencies of the ML-estimators based on these two versions of the likelihood are compared.

Our choice of sampling distribution for the importance sampler used here did not perform well when θ^* was poorly specified. Although some general recommendations for improving the sampling distribution can be made we believe that the choice will also depend on the specific analysis to be performed. This may prove problematic in real applications, where competence in such choices may not be available.

The SEM type algorithm was slower to run than the other methods discussed. The speed of the algorithm may be a problem if ascertainment probability is low for some portion of data, since large sets of data will then have to be filled in. An alternative to sampling the whole set of missing data is to simulate only a portion, n^{obs}/q, where q is a fixed number, of the data and weigh up the likelihood contribution of the simulated data. Data can for example be simulated as above until n^{obs}/q observations from Z_j^{mis} with $A_j^{mis} = 1$ are obtained. Too small values of n^{obs}/q will however cause too large variability in parameter estimates. Ripatti, Larsen, and Palmgren (2002) suggest a rule for increasing the number of samples in a Monte Carlo EM (Wei and Tanner, 1990) algorithm when approaching convergence. The basic idea of altering the number of samples when approaching the estimate could be used also in our setting. If the size of the missing data is small it is possible to reduce the variability per simulation step by choosing $N > 1$, giving an algorithm similar to the Monte Carlo EM, although this complicates the maximization step. Even if the running time can be shortened somewhat by sampling techniques as this, due to the distributional assumptions discussed above, the SEM type algorithm is best suited for data where sampling probabilities are not extremely small.

Other sampling strategies, e.g. two-stage designs, where some information is retained on all individuals, may in some cases be handled by a slightly modified version of the algorithm. This sampling scheme is on the other hand more similar to the classical missing data setting, with a known number of observations missing at random (MAR). Then methods such as multiple imputation (Little and Rubin, 1987) may be useful.

Another possibility is to consider Bayesian inference, letting θ be random with a prior distribution $P(\theta)$. It turns out that our SEM approach can be modified to a Bayesian inference with small modifications. If $\{(\mathbf{Z}_i^{mis}, \theta_i)\}_{i=1}^{B+I}$ are simulated from the posterior $\mathbf{Z}^{mis}, \theta | \mathbf{Z}^{obs}$ we could use a blockwise Metropolis-Hasting algorithm and alternate between updating $\mathbf{Z}^{mis}$ and θ. The resulting algorithm is very similar to the SEM algorithm. The Simulation-step can still be used for updating

DOI: 10.2202/1557-4679.1222

$\mathbf{Z}^{mis}$, whereas the Maximization-step has to be modified. Rather than maximizing a likelihood, the updated θ_i is drawn from the blockwise posterior distribution $\theta|\mathbf{Z}^{obs},\mathbf{Z}^{mis}$. A related iterative Bayesian procedure is data augmentation (Tanner and Wong, 1987, Tanner, 1991), where $N > 1$ missing data sets are imputed within each iteration.

References

Agardh, E., A. Ahlbom, T. Andersson, S. Efendic, V. Grill, J. Hallqvist, A. Norman, and C. Ostenson (2003): "Work stress and low sense of coherence is associated with type 2 diabetes in middle-aged Swedish women." *Diabetes Care*, 26, 719–24.

Allison, D., M. Heo, N. Schork, S. Wong, and R. Elston (1998): "Extreme selection strategies in gene mapping studies of oligogenic quantitative traits do not always increase power." *Human Hered.*, 48, 97–107.

Breslow, N. (1982): "Design and analysis of case-control studies," *Annual Review of Public Health*, 3, 29–5.

Breslow, N. and K. Cain (1988): "Logistic regression for two-stage case-control data," *Biometrika*, 75, 11–20.

Breslow, N., J. M. Robins, and J. Wellner (2000): "On the semi-parametric efficiency of logistic regression under case-control sampling," *Bernoulli*, 6, 447–455.

Breslow, N. E. and N. Chatterjee (1999): "Design and analysis of two-phase studies with binary outcome applied to wilms tumour prognosis," *Applied Statistics*, 48, 457–468.

Breslow, N. E. and R. Holubkov (1997): "Maximum likelihood estimation of logistic regression parameters under two- phase, outcome-dependent sampling," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59, 447–461.

Celeux, G. and J. Diebolt (1985): "The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem," *Computational Statistics [Formerly: Computational Statistics Quarterly]*, 2, 73–82.

Chen, H. Y. (2003): "A note on the prospective analysis of outcome-dependent samples," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65, Part 2, 575–584.

Chen, J. and N. Chatterjee (2007): "Exploiting hardy-weinberg equilibrium for efficient screening of single SNP associations from case-control studies," *Human Heredity*, 63, 196–204.

Clayton, D. (2003): "Conditional likelihood inference under complex ascertainment using data augmentation." *Biometrika*, 90, 976–981.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977): "Maximum likelihood from incomplete data via the EM algorithm (with discussion)," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39, 1–37.

Efron, B. (1992): "Missing data, imputation and the bootstrap," Technical report, Division of Biostatistics, Stanford University.

Fratiglioni, L., M. Grut, Y. Forsell, M. Viitanen, M. Grafstrom, K. Holmen, K. Ericsson, L. Backman, A. Ahlbom, and B. Winblad (1991): "Prevalence of Alzheimer's disease and other dementias in an elderly urban population: Relationship with age, sex, and education," *Neurology*, 41, 1886–1892.

Gilks, W. R., S. Richardson, and D. J. Speigelhalter (1996): *Markov Chain Monte Carlo in practice*, London. pp 15, 259-273: Chapman & Hall, first edition.

Grünewald, M. and O. Hössjer (2010): "Efficient ascertainment schemes for maximum likelihood estimation," *Journal of Statistical Planning and Inference*, 140, 2078 – 2088.

Grünewald, M. (2004): *Genetic association studies with complex ascertainment*, Licenciate thesis 2004:5, Stockholm University, Department of Mathematics, Stockhom University, 10691 Stockholm, Sweden.

Gu, H., A. Abulaiti, C. Ostenson, K. Humphreys, C. Wahlestedt, A. Brookes, and S. Efendic (2004): "Single nucleotide polymorphisms in the proximal promoter region of the adiponectin (APM1) gene are associated with type 2 diabetes in Swedish caucasians." *Diabetes*, 53, Suppl 1:S31–5.

Hammersley, J. M. and D. C. Handscomb (1964): *Monte Carlo methods*, London: Methuen.

Hesterberg, T. (1995): "Weighted average importance sampling and defensive mixture distributions," *Technometrics*, 37, 185–194.

Ip, E. H. S. (1994): "A stochastic EM estimator in the presense of missing data -theory and applications." *Technical report, Department of Statistics, Stanford University*.

Liang, K.-Y. (1983): "On information and ancillarity in the presence of a nuisance parameter," *Biometrika*, 70, 607–612.

Little, R. J. A. and D. Rubin (1987): *Statistical analysis with missing data*, Hoboken, N.J.: John Wiley & Sons.

Little, R. J. A. and D. Rubin (2002): *Statistical analysis with missing data*, Wiley series in probability and statistics, New York; Chichester: John Wiley & Sons, Inc.

Louis, T. A. (1982): "Finding the observed information matrix when using the EM algorithm," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 44, 226–233.

DOI: 10.2202/1557-4679.1222

McLachlan, G. J. and T. Krishnan (1997): *The EM algorithm and extensions*, John Wiley & sons Inc, chapter 6.

Monsees, G., R. Tamimi, and P. Kraft (2009): "Genome-wide association scans for secondary traits using case-control samples," *Genetic Epidemiology*, 33, 717–728.

Patil, G. (2002): "Weighed distributions," *Encyclopedia of Environmetrics*, 4, 2369–2377.

Prince, J. A., H. Zetterberg, N. Andreasen, J. Marcusson, and K. Blennow (2004): "Apoe $\varepsilon 4$ allele is associated with reduced cerebrospinal fluid levels of a$\beta 42$," *Neurology*, 62, 2116–2118.

Ripatti, S., K. Larsen, and J. Palmgren (2002): "Maximum likelihood inference inference for multivariate frailty models using an automated monte carlo EM algorithm," *Lifetime Data Analysis*, 8, 349–360.

Tanner, M. (1991): *Tools for statistical inference. Observed data and data augmentation methods.*, Berlin: Springer.

Tanner, M. A. and W. H. Wong (1987): "The calculation of posterior distributions by data augmentation," *Journal of the American Statistical Association*, 82, 528–540.

The R Development Core Team (2001): "R," Version 1.4.0.

Tolonen, H., K. Kari, and E. Ruokokoski (2000): "Monica population survey data book," WWW-publications from the WHO MONICA Project "`http://www.ktl.fi/publications/monica/surveydb/title.htm`".

Tregouet, D., S. Escolano, L. Tiret, A. Mallet, and J. L. Golmard (2004): "A new algorithm for haplotype-based association analysis: the Stochastic-EM algorithm," *Annals of Human Genetics*, 68, 165–177.

Wacholder, S. and C. R. Weinberg (1994): "Flexible maximum likelihood methods for assessing joint effects in case-control studies with complex sampling," *Biometrics*, 50, 350–357.

Wei, G. C. G. and M. A. Tanner (1990): "A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms," *Journal of the American Statistical Association*, 85, 699–704.

Zhou, H., J. Chen, T. H. Rissanen, S. A. Korrick, H. Hu, J. T. Salonen, and M. P. Longnecker (2007): "Outcome-dependent sampling: An efficient sampling and inference procedure for studies with a continuous outcome," *Epidemiology*, 18, 461–468.