Two-part Models for Assessing Insurance Misrepresentation

Li-Chieh Chen¹, Jianxi Su¹*, and Michelle Xia²

¹Department of Statistics, Purdue University, West Lafayette, IN, 47906, United States.
²Division of Statistics, Northern Illinois University, Dekalb, IL, 60115, U.S.A.

February 26, 2019

Abstract

Modeling insurance claims is a classical actuarial problem. Actuaries analyze the claim distribution based on a given set of rating factors, yet some of which may be subject to misrepresentation. In this paper, we propose a novel class of two-part insurance loss models that can a) adjust for the misrepresentation risk in insurance underwriting; b) account for the zero-inflated feature of insurance data; c) allow for arbitrary numbers of risk factors in the analysis. The unobserved feature of misrepresentation is addressed through a latent factor on the misrepresentation status, shared by the regression models on the claim occurrence and severity. Taking advantage of the tractability of the proposed models, we derive explicit iterative formulas for the expectation maximization algorithm adopted in the estimation process. Explicit expressions are also obtained for the observed Fisher information matrix, enabling large-sample inference on the rating factors in a computationally convenient manner. The practical usefulness of the proposed models is demonstrated by two empirical applications based on real insurance data.

Keywords: Expectation maximization algorithm, loss modeling, misrepresentation fraud, mixture models, zero-inflation.

JEL classifications: C02, C46

*Corresponding author. E-mail: jianxi@purdue.edu; postal address: 150 N. University Street, West Lafayette, IN, 47906, United States.
1 Introduction

Fraud is a pervasive issue in the insurance industry. Reported by the Coalition Against Insurance Fraud, insurance costs due to fraudulent activities amount to more than $80 billion per year in the US, and the Federal Bureau of Investigation estimates the annual total cost of non-health insurance fraud to be more than $40 billion (FBI, 2011). Misrepresentation is a common type of insurance fraud where policy applicants make untrue statements on certain rating factors in order to lower the premium. Rating factors that are usually subject to misrepresentation include self-reported traits such as the smoking status and medical history in the context of health insurance, and the mileage and use of vehicle in the context of auto insurance. Without an adequate risk management process in place, misrepresentation activities can result in a tremendous amount of extra cost to insurance companies and lower-risk insureds who do not misrepresent. Misrepresentation risk is also of particular interest to insurance regulators who attempt to understand how the presence of fraudulent behaviors by a group of insured individuals might affect the welfare of the others (Gabaldón et al., 2014). Thereby, misrepresentation risk management is a crucially important topic in actuarial research.

At the heart of misrepresentation risk management is the quantitative analysis of misrepresentation prevalence within the insureds pool. There is a rich body of literature focusing on improving business strategies and policy designs for mitigating fraud risk, yet statistical or probabilistic models tailored for misrepresentation risk assessment seem to have gathered little attention. What complicates the quantitative modeling of misrepresentation risk, from a statistical standpoint, is the unobservable nature of fraudulent activities. Traditional statistical methods (e.g., discriminant and regression analysis) require the access to a sample frame containing the random variable of concern (i.e., the misrepresentation status under the policy level). In the practical instances, there are masses of claim data available to the insurance companies, but knowledge on whether misrepresentation actually occurred cannot be obtained until a formal investigation is undertaken on a policy. Validating misrepresentation can be an excessively costly process for insurance companies, sometimes even impossible to conduct from a practical perspective. Worse still, as many perpetrators of insurance frauds have a learning curve, the predictor variables for misrepresentation may vary substantially over time and depend heavily on the insurance policy designs. Hence, even if insurance companies may gather some historical data concerning misrepresentation activities, they may not be credible for the managerial needs. For measuring misrepresentation risk, mathematically sound yet easy-to-implement models relying only on the available ratemaking data, are more likely to be appreciated by insurance practitioners.

In order to overcome the obstacles noted above, a useful statistical ingredient for modeling misrepresentation risk is the application of mixture models under a hybrid structure (see, e.g., Blostein and Miljkovic, 2019; Lee and Lin, 2010; Miljkovic and Grün, 2016; Ratovomirija et al., 2017, for a wide range of recent applications of mixture models in actuarial mathematics). Specifically, we use latent factors to endogenize the probabilistic mixtures of two groups among the insurance applicants who denied the risk status, consisting of the applicants with and without misrepresentation. We
assume the claim distribution for each of the two groups follows a known parametric form. By fitting the models to the original unclassified ratemaking data, the estimates of the models’ parameters can provide statistical evidence for assessing the severity of misrepresentation risk. Another attractive merit for adopting the aforementioned mixtures approach is its compliance with the existing generalized linear model (GLM) ratemaking framework which has now evolved as an industry standard (see, Frees, 2009, for a comprehensive reference). The concerns about the misrepresentation risk can be conveniently embedded into the existing ratemaking framework through additional latent factors. The resulting models will retain the flexibility, tractability, and interpretability in the classical GLM ratemaking framework.

To the best of our knowledge, Xia and Gustafson (2016) was the first attempt in the literature of using mixture models to study insurance misrepresentation. In their paper, they used Bayesian methods to make inference on the parameters of their misrepresentation models. For insurance applications, frequentist inference is widely adopted and sometimes more favored by the industry. In order to facilitate industrial applications, more recently, Akakpo et al. (2018) proposed a frequentist alternative to implement the models of Xia and Gustafson (2016).

The earlier studies of both Xia and Gustafson (2016) and Akakpo et al. (2018) are based on a simplified setup which only includes one single risk factor that is subject to misrepresentation, without considering other risk factors. Misrepresentation models that allow for multiple risk factors are rarely found in the literature, except for a recent attempt by Xia et al. (2018). What is more, all the aforementioned misrepresentation models can only handle claim data with positive outcomes. It is well known that insurance claim data typically exhibit a semi-continuous pattern, meaning that there are a noticeable portion of zeros in conjunction with a positive continuous component. Data of this kind are also known as zero-inflated data in the field of statistical modeling. The zeros which correspond to the policyholders without making any claim over a given policy period (e.g., a month/quarter/year), play a critically important role in many insurance applications including policy design, ratemaking, and risk management. We naturally argue that these zero claim entries should not be neglected in the analysis of misrepresentation risk, since they may carry valuable information on misrepresentation. For instance, in the context of health insurance, a smoker who makes an untrue statement on the tobacco consumption status may have a different likelihood of making positive claims when compared with a non-smoker. Modeling the zero and positive claim records simultaneously under a unified framework can help to reveal additional information on misrepresentation. However, to the best of our knowledge, there is no existing misrepresentation model that can account for the zero-inflated feature of insurance data. This paper contributes to filling this gap.

In this paper, we aim to put forward a class of zero-inflated loss models adjusting for misrepresentation risk. The proposed models advance the ones by Xia and Gustafson (2016) and Akakpo et al. (2018) from two perspectives. First, we employ a two-part structure for assessing the misrepresentation risk. One part of the proposed models accounts for the misrepresentation information from the occurrence of positive claims, while the other part captures the misrepresentation signal from the severity of positive claims. Second, our misrepresentation models can incorporate an arbitrary number of risk factors in the loss occurrence and severity models. In so doing, the proposed models are able to better account for the
variability induced by multiple risk factors, leading to a more reliable assessment on misrepresentation. We will provide a
detailed documentation about the proposed models’ implementation based on the frequentist approach (with the Bayesian
implementation also briefly summarized in Appendix B). Despite that the proposed models are considerably more complex
than the ones of Xia and Gustafson (2016) and Akakpo et al. (2018), we are still able to obtain explicit expressions for
the estimation algorithm and the observed Fisher information matrix for the purpose of frequentist inference.

In addition to the aforementioned methodological contributions, our paper makes major empirical contributions to
the existing misrepresentation literature. Due to the scarce availability of insurance data in the academia, previous
misrepresentation risk models were only illustrated using the Medical Expenditure Panel Survey (MEPS) data where the
participants had relatively weak financial incentives for misrepresentation. In this article, we implement the proposed
models on a set of real insurance data coming from a group health insurance product. A distinct feature that makes our
data suitable for studying misrepresentation models is that the rating factor on tobacco consumption is acquired through
a self-certification process. Owing to the financial incentive towards the insureds and the lack of medical screening by the
insurance provider, the smoking status can be subject to a high level of misrepresentation risk.

The rest of this article proceeds as follows. Section 2 lays the theoretical groundwork for establishing the two-part
loss models adjusting for misrepresentation. Statistical estimation and inference for the proposed models will be discussed
in Section 3. By means of simulations, we examine the performance of the proposed estimation procedure in Section
4. In Section 5, we present the application of our proposed models on two sets of real insurance data for assessing
misrepresentation risk. Section 6 concludes the paper. All the technical proofs are relegated to Appendix A in order to
simplify the layout.

2 The proposed model

In this section, we propose a class of ratemaking models which can (a) adjust for misrepresentation risk; (b) incorporate an
arbitrary number of risk/ratemaking factors in the loss analysis; and (c) account for the zero-inflated feature of insurance
data. Toward this aim, we begin with the notion for describing the risk factor that is subject to misrepresentation. Based
on the application problems of interest, we confine ourselves to the case where the potentially misrepresented factor is
binary, denoted by $V \in \{0, 1\}$. Due to the potential financial incentives, insufficient screenings and/or ineffective oversight,
the observed variable $V^* \in \{0, 1\}$ to be reported by policyholders, may or may not have the same value as the true variable
$V$. For example, in the health insurance application, one may think of $V$ as the true smoking status of an insured and
$V^*$ as the self-reported smoking status. Without loss of generality, assume $V = 0$ represents the negative risk status
associated with a lower insurance premium.

In the insurance context, misrepresentation usually occurs in a single direction which benefits the insureds. In other
words, it is impossible for a insured to misrepresent if it will cause his/her premium to rise. This unidirectional feature
implies $P[V^* = 1|V = 0] = 0$. We say misrepresentation occurs if the event \( \{V^* = 0|V = 1\} \) occurs. The associated misrepresentation probability is defined as $p := P[V^* = 0|V = 1] \in [0,1]$.

Let $Y$ be a response variable that represents the aggregate claim/loss under the policy level over a given policy period. Furthermore, assume there are $k \in \mathbb{N}$ correctly measured rating factors $X = (X_1, \ldots, X_k) \in \mathbb{R}^k$ that are predictive of the loss outcome $Y$. In classical insurance ratemaking, actuaries are typically interested in modeling the conditional distribution of $(Y|V, X)$ via the GLM. As we mentioned earlier in the introduction, the current paper focuses on the aggregate loss variable $Y$ that possesses a zero-inflated loss distribution.

In analyzing zero-inflated data, one common approach is the Tweedie GLM in which the claim response variable $Y$ is interpreted as a compound Poisson summand of independent gamma random variables (e.g., Gordon and Jørgensen, 2002; Jørgensen and de Souza, 1994). The associated Tweedie distribution has a positive probability mass at zero. Another strand of the related literature focuses on the use of so-called two-part models (Frees, 2009). The essence of two-part models is to decompose the claim response variable into two components: frequency and severity. The claim frequency/occurrence part indicates whether or not a positive claim has occurred, typically modeled by a binomial regression. Given that a claim has occurred, another regression is specified on the claim amount to model the severity. Compared to the Tweedie GLM approach, the two-part models feature the flexibility in specifying different structures for modeling the claim frequency and severity, as well as the separability in the estimation process (i.e., parameters under the frequency and severity parts can be calibrated separately; see Frees, 2009, for more details). Therefore, it is more convenient to incorporate the misrepresentation effects based on the two-part models. For this reason, in our paper, we will choose the two-part model approach for handling zero-inflated data.

In order to facilitate our subsequent discussion, some elementary notations are defined herein. First, we assume the sets of explanatory variables predictive of the claim occurrence and severity are not necessarily identical. For $\mathcal{F}, \mathcal{S} \subseteq \{1, \ldots, n\}$, let $X^{\square} = \{X_i : i \in \square\}$ where ‘\(\square\)’ can be one of $\{\mathcal{F}, \mathcal{S}\}$, denote the sets of rating factors that influence the claim frequency and severity respectively. Second, we conventionally denote constant matrices by boldface capital letters (e.g., $A$), and constant row vectors by boldface lower case letters (e.g., $a$). The notation ‘$|a|$’ refers to the cardinality of vector $a$. An $n(\in \mathbb{N})$-dimensional column vector of 1 is denoted by $1_n$. For any two vectors $v_1$ and $v_2$, the vector concatenation notation is $v_1 \parallel v_2 := [v_1 \ v_2]$.

We are now in the position to specify the two-part models adjusting for misrepresentation risk. We suggest using a logistic regression to model the occurrence of positive insurance claims. Given that a claim has occurred, we assume the claim size to follow the lognormal distribution. We openly admit that the aforementioned assumptions seem to be rather ad. hoc. at the first glance, but they are indeed very common model choices in the insurance literature, particularly for the analysis of healthcare expenditures (see, e.g., Akakpo et al., 2018; Diehr et al., 1999; Duan et al., 1983; Zhou and Tu, 2000, and the references therein). Extending the proposed framework to the more general exponential family of distributions is feasible and will be addressed in future research.
We next formulate our proposed setup mathematically. Let \( G = I_{(Y > 0)} \) be an indicator variable describing whether or not a positive claim has occurred and 

\[
\pi_{V, X^F} := \Pr(G = 1 | V, X^F) = \mathbb{E}[G | V, X^F].
\]

We assume

\[
(G | V = v, X^F = x^F) \sim \text{Bernoulli}(\pi_{v, x^F}), \quad \logit(\pi_{v, x^F}) = \beta_0 + \sum_{i \in F} \beta_i x_i + \beta_{k+1} v,
\]

where \( v \in \{0, 1\} \), \( x^F \in \mathbb{R}^{|F|} \), and \( \beta = (\beta_0, \{\beta_i\}_{i \in F}, \beta_{k+1}) \) contains the regression parameters for the claim occurrence model. Recall that a random variable \( X \in \mathbb{R}_+ \) is said to have a lognormal distribution with location parameter \( \mu \in \mathbb{R} \) and shape parameter \( \sigma \in \mathbb{R}_+ \), succinctly \( X \sim \text{LN}(\mu, \sigma^2) \), if the associated probability density function (PDF) is

\[
g(x; \mu, \sigma) = \frac{1}{x \sigma \sqrt{2\pi}} \exp \left[-\frac{(\log x - \mu)^2}{2\sigma^2}\right], \quad x > 0.
\]

Given that a positive claim has occurred, we assume

\[
(Y | G = 1, V = v, X^S = x^S) \sim \text{LN}(\mu_{v, x^S}, \sigma^2), \quad \mu_{v, x^S} = \alpha_0 + \sum_{i \in S} \alpha_i x_i + \alpha_{k+1} v,
\]

where \( v \in \{0, 1\} \), \( x^S \in \mathbb{R}^{|S|} \), and \( \alpha = (\alpha_0, \{\alpha_i\}_{i \in S}, \alpha_{k+1}) \) contains the regression parameters for the claim severity model.

The following assertion states the conditional distribution of the claim variable \( Y \) given the true variable \( V \) and the correctly measured variables \( X \). The proof is straightforward and is hence omitted.

**Proposition 1.** Suppose \( Y \) follows a two-part model admitting representations (1) through (3), then the conditional PDF of \( Y \) given \( V = v \in \{0, 1\} \) and \( X = x \in \mathbb{R}^k \) is given by

\[
f_{Y|V=v,X=x}(y; \alpha, \beta, \sigma) = I_{(y > 0)} \pi_{v, x^F} g(y; \mu_{v, x^S}) + I_{(y = 0)} (1 - \pi_{v, x^F}), \quad y \geq 0,
\]

where \( \alpha \) and \( \beta \) contain the regression parameters for the claim occurrence and size distributions respectively, \( \sigma \) is the claim size shape parameter, and \( g(\cdot) \) denotes the lognormal PDF given in (2).

It is noteworthy that in PDF (4), \( Y \) and \( X \) are observable while the true status variable \( V \) is unobservable. Instead, what can be observed is the reported variable \( V^* \) which may or may not coincide with \( V \). In misrepresentation modeling and insurance ratemaking, the condition distribution of \( Y \) given \( (V^*, X) \) plays a more important role in understanding and adjusting for the impact from misrepresentation. Henceforth, for notational convenience, we suppress the condition in PDF (4) and simply write

\[
f_v(y; \alpha, \beta, \sigma) := f_{Y|V=v,X=x}(y; \alpha, \beta, \sigma), \quad y \geq 0, \quad v \in \{0, 1\}.
\]
Recall that we have assumed misrepresentation to be unidirectional (i.e., \( \Pr[V^* = 1|V = 0] = 0 \)). If \( V^* = 1 \), then

\[
f_{Y|V^* = 1, X=x}(y; \alpha, \beta, \sigma) = f_1(y; \alpha, \beta, \sigma), \ y \geq 0.
\]

(6)

If \( V^* = 0 \), then misrepresentation may happen, and \( V \) and \( V^* \) may not be identical. Let \( \lambda = \Pr[V = 1|V^* = 0] \), we have

\[
f_{Y|V^* = 0, X=x}(y; \alpha, \beta, \sigma, \lambda) = (1 - \lambda)f_0(y; \alpha, \beta, \sigma) + \lambda f_1(y; \alpha, \beta, \sigma), \ y \geq 0,
\]

(7)

where \( f_v(\cdot) \) for \( v \in \{0, 1\} \) are defined as per Equation (5). It is important to point out that the parameter \( \lambda \) is different from the misrepresentation probability \( p = \Pr[V^* = 0|V = 1] \). Yet, they are closely related by \( \lambda = \frac{\frac{p}{\Pr[V = 1]} - 1}{\frac{1}{1 - p} - 1} \). When \( \Pr[V = 1] \) is fixed (e.g., the population smoking rate in the health insurance context), \( \lambda \) is an increasing transform of the misrepresentation probability \( p \). Therefore, the value of \( \lambda \) estimated from regular ratemaking data can be viewed as a measure of misrepresentation risk. In order to distinguish between \( \lambda \) and \( p \), we term \( \lambda \) the prevalence of misrepresentation throughout the rest of this paper.

Clearly, PDF (7) admits a probability structure of a two-component mixture. For the application of mixture models, one of the major concerns is the models’ identifiability. The issue is particularly critical for the proposed misrepresentation models since each individual mixture component also contains a two-part semi-continuous regression structure. In order to ensure valid inference on the proposed model’s parameters (i.e., \( \alpha, \beta, \sigma, \lambda \)), the model’s identifiability needs to be verified. Generally speaking, PDF (7) is not identifiable. To see this, note that \( f_0 \) and \( f_1 \) in Equation (5) are different only up to an intercept term in the regression model. It is straightforward to check that

\[
f_{Y|V^* = 0, X=x}(y; \alpha, \beta, \sigma, \lambda) = f_{Y|V^* = 0, X=x}(y; \alpha^*, \beta^*, \sigma, (1 - \lambda)) \text{ for all } y \geq 0 \text{ where } \alpha^* = (\alpha_0 + \alpha_{k+1}, \{\alpha_i\}_{i \in S}, -\alpha_{k+1}) \text{ and } \beta^* = (\beta_0 + \beta_{k+1}, \{\beta_i\}_{i \in F}, -\beta_{k+1}).
\]

Due to this exchangeability issue, it is impossible to distinguish the component that corresponds to the misrepresented cases (i.e., \( f_1 \)) solely based on PDF (7). However, taking advantage of the unidirectional nature of misrepresentation in our setup, component distribution \( f_1 \) can be identified separately from the subgroup of observations with \( V^* = 1 \) that has PDF (6). This unidirectional feature indeed resolves the exchangeability issue in PDF (7) and naturally makes our proposed misrepresentation models identifiable. This implies all the parameters in the proposed models can be estimated consistently from the observed data, without resorting to the true status variable or requiring any additional assumption on misrepresentation.

### 3 Model estimation and inference

The misrepresentation problem presented in Section 2 can be viewed as a special case of unidirectional misclassification (e.g., Gustafson, 2014). In the study of misclassification, the Bayesian approach is commonly used in the literature. This is particularly true when the underlying models are not identifiable, which makes the prior knowledge on misrepresentation
necessary for proper inference.

Owing to the model’s identifiability, the proposed two-part models can be implemented using either Bayesian or frequentist methods. We deliberately place our special focus on the frequentist approach in order to be better in line with the existing literature on GLM ratemaking. Compared with the Bayesian approach, the frequentist method enjoys the merit of mathematical tractability and practical simplicity. For instance, in a Bayesian analysis, an improper prior may yield an improper posterior distribution, even when the likelihood is identifiable. The implementation of mixture models in a Bayesian framework requires posterior simulations via Markov chain Monte Carlo (MCMC) methods. The MCMC implementation with a set of weakly identifiable parameters may result in poor inference due to slow convergence and short effective chain lengths. The slow convergence problem becomes more pronounced when there is a posterior correlation between successive samples. From the perspective of required computation resources, the Bayesian implementation needs to sample a latent misrepresentation indicator for each observation in every iteration of the MCMC. Even under the simplest setup having only one risk factor and no two-part model structure, the computational time required for Bayesian MCMC can be 1000 times longer than that required by frequentist inference (e.g., Akakpo et al., 2018). Thus, the frequentist method may be more suitable for insurance companies that usually deal with large volume data. Nevertheless, if the risk analyst or company’s senior management aims to insert a subjective assessment on the severity and uncertainty of misrepresentation, the Bayesian methods become immediately useful owing to its ability to incorporate external information through priors. For the sake of completeness while keeping the focus of this current paper, we relegate the discussion on a Bayesian implementation of the proposed two-part models to Appendix B.

Consider a random sample of size \( n \in \mathbb{N} \) from the observed variables \((Y, V^*, X)\), denoted respectively by \( y = (y_1, \ldots, y_n) \in \mathbb{R}^n_0, v^* = (v^*_1, \ldots, v^*_n) \in \{0, 1\}^n \), and by \( x = (x_1, \ldots, x_n) \in \mathbb{R}^k \times \mathbb{R}^n \) with \( x_i = (x_{i1}, \ldots, x_{ik})' \in \mathbb{R}^k; i = 1, \ldots, n \). Shorthand the parameter vector by \( \Psi = (\alpha, \beta, \sigma, \lambda) \). The partial likelihood function underlying the two-part misrepresentation model (6) and (7) can be obtained via

\[
L(\Psi) := L(\Psi; y, v^*, x) = \prod_{i=1}^n \left\{ v^*_i f_1(y_i; \Psi) + (1 - v^*_i) \left[ (1 - \lambda) f_0(y_i; \Psi) + \lambda f_1(y_i; \Psi) \right] \right\}
\]

\[
= \prod_{i=1}^n \left\{ v^*_i f_1(y_i; \Psi) + (1 - v^*_i) \left[ \sum_{j=0}^{1} [(1 - j)(1 - \lambda) + j\lambda] f_j(y_i; \Psi) \right] \right\},
\]

where \( f_0 \) and \( f_1 \) are formulated according to Equation (5) with the parameters \((\alpha, \beta, \sigma)\) in \( \Psi \). The corresponding partial log likelihood function for the two-part model is given by

\[
l(\Psi) := l(\Psi; y, v^*, x) = \sum_{i=1}^n v^*_i \log f_1(y_i; \Psi) + \sum_{i=1}^n (1 - v^*_i) \log \left\{ \sum_{j=0}^{1} [(1 - j)(1 - \lambda) + j\lambda] f_j(y_i; \Psi) \right\}.
\]

(8)

Unfortunately, due to the additive structure in the latter logarithm function of Equation (8), no closed-form expression
can be directly obtained for the maximum likelihood estimator (MLE) of $\Psi$. To circumvent the difficulty, we resort to the expectation maximization (EM) algorithm which has evolved as a prevailing method for solving the MLE of mixture (regression) models. The EM algorithm is a natural choice for implementing our two-part misrepresentation models which are a hybrid of mixture regression and single component regression based on semi-continuous outcomes. As we shall see in a moment, prominent advantages of the EM algorithm for the proposed two-part misrepresentation models include a) explicit formulas for iterative updates of parameters (except for $\beta$); b) numerically stable algorithm with increasing likelihood from iteration to iteration; c) no requirement for large computational capacity or storage. We refer the readers to McLachlan and Krishnan (2008); McLachlan and Peel (2000) for more detailed comparisons between the EM algorithm and other numerical methods.

Deriving the MLE of $\Psi$ would be much more straightforward if the true status $V$ is observable, or equivalently if the misrepresentation status is known. In reality, collecting information about the true status variable $V$ can be excessively costly for insurance companies and sometimes even impossible from a practical standpoint. In the EM algorithm, we introduce a set of latent variables to endogenize the occurrence of misrepresentation so that the complete-data likelihood can be used to estimate $\Psi$. Specifically, let $z = (z_1, \ldots, z_n) \in \{0,1\}^n$ be a vector of latent variables, with $z_i = 1$, $i = 1, \ldots, n$, indicates misrepresentation occurs in the $i$-th observation, and $z_i = 0$ otherwise. In this article, we assume the population misrepresentation prevalence $\lambda$ does not depend on the response variable $Y$, meaning the distribution of $z$ is not related to the parameters in $\Psi$ other than $\lambda$. This assumption is made because we mainly focus on assessing the presence of a significant level of misrepresentation in the population (see the adoption of the same model assumption in Akakpo et al., 2018; Xia and Gustafson, 2016). For the proposed two-part models, relaxing the assumption and embedding predictive modeling on misrepresentation at the policy level is possible and in fact is a very interesting research topic. We leave this extension for future work.

With the latent vector $z$ in place, we have readily obtained a vector of complete data $c = (y, x, v^*, z)$. For the proposed two-part models, the complete-data likelihood associated with $c$ can be expressed as

$$L_c(\Psi) := L(\Psi; y, x, v^*, z) = \prod_{i=1}^{n} \left\{ v_i^* f_1(y_i; \Psi) + (1 - v_i^*) \left[ ((1 - \lambda) f_0(y_i; \Psi))^{1-z_i} + (\lambda f_1(y_i; \Psi))^{z_i} \right] \right\}$$

$$= \prod_{i=1}^{n} \left\{ v_i^* f_1(y_i; \Psi) + (1 - v_i^*) \sum_{j=0}^{1} \left[ ((1 - j)(1 - \lambda) + j\lambda) f_j(y_i; \Psi) \right]^{(1-j)(1-z_i)+jz_i} \right\},$$

with the corresponding complete-data log likelihood function given by

$$l_c(\Psi) = \sum_{i=1}^{n} v_i^* \log f_1(y_i; \Psi) + \sum_{i=1}^{n} (1 - v_i^*) \left\{ \sum_{j=0}^{1} \left[ [(1 - j)(1 - z_i) + jz_i] \log \left[ ((1 - j)(1 - \lambda) + j\lambda) f_j(y_i; \Psi) \right] \right\}. \quad (9)$$

The EM algorithm for implementing the proposed two-part misrepresentation models consists of the expectation and
maximization steps in each iteration, which will be elaborated in detail next.

### 3.1 Expectation step

Recall that the EM algorithm is an iterative algorithm for computing MLE. For $s \in \mathbb{N}_0$, let $\Psi^{(s)} = (\alpha^{(s)}, \beta^{(s)}, \sigma^{(s)}, \lambda^{(s)})$ be the updated estimate of $\Psi$ in the $s$-th iteration. In the expectation step (a.k.a., E-step), the algorithm calculates the conditional expectation of complete-data likelihood (9) with respect to $z$, given $(y, x, v^*)$ and $\Psi^{(s)}$. The conditional mean $\eta^{(s)} = \mathbb{E}[z | y, x, v^*, \Psi^{(s)}]$ plays an important role in the E-step. For $i = 1, \ldots, n$, if $v_i^* = 0$, we can derive

$$
\eta_i^{(s)} = \mathbb{E}[z_i | y_i, x_i, v_i^* = 0, \Psi^{(s)}] = P[z_i = 1 | y_i, x_i, v_i^* = 0, \Psi^{(s)}] = \frac{\lambda^{(s)} f_1(y_i; \Psi^{(s)})}{\sum_{j=0}^{1} [(1-j)(1-\lambda) + j\lambda^{(s)}] f_j(y_i; \Psi^{(s)})},
$$

where (1) holds due to the independence assumption between $z$ and $Y$ stated earlier. Note that $\eta^{(s)} = (\eta_1^{(s)}, \ldots, \eta_n^{(s)})$ are the posterior probabilities on the occurrence of misrepresentation based on the updated parameter values in the $s$-th iteration. Because misrepresentation may only occur in one direction, if $v_i^* = 1$, then $\eta_i^{(s)} = 0$.

The conditional expectation of the complete-data log likelihood function is referred to as the Q-function in the related literature, and can be computed for the proposed two-part misrepresentation models via

$$
Q \left( \Psi | \Psi^{(s)} \right) = \mathbb{E} \left[ l_c(\Psi) | \Psi^{(s)} \right] = \sum_{i=1}^{n} v_i^* \log f_1(y_i; \Psi) + \sum_{i=1}^{n} (1 - v_i^*) \left\{ \sum_{j=0}^{1} [(1-j)(1-\eta_j^{(s)}) + j\eta_j^{(s)}] \log [(1-j)(1-\lambda) + j\lambda] f_j(y_i; \Psi) \right\}.
$$

The Q-function is a critical input of the maximization step for updating the parameters in the EM algorithm.

### 3.2 Maximization step

Given the parameter values from the $s$-th iteration, the maximization step (a.k.a., M-step) aims to update the values of $\Psi$ by maximizing Q-function (10). Owing to the linearity of Equation (10), the optimization problem can be solved separately for different parameters. The next assertion spells out the iterative formulas for updating the parameters $(\alpha^{(s+1)}, \sigma^{(s+1)}, \lambda^{(s+1)})$, which are all explicit. At the outset, let us define

$$
B = \begin{pmatrix}
I_{(y_1 > 0)} \cdots I_{(y_1 > 0)x_1} & \cdots & v_1^* + (1 - v_1^*)\eta_1^{(s)} \\
I_{(y_2 > 0)} \cdots I_{(y_2 > 0)x_2} & \cdots & v_2^* + (1 - v_2^*)\eta_2^{(s)} \\
\vdots & \vdots & \vdots \\
I_{(y_n > 0)} \cdots I_{(y_n > 0)x_n} & \cdots & v_n^* + (1 - v_n^*)\eta_n^{(s)}
\end{pmatrix},
$$

$$
R = \begin{pmatrix}
v_1^* + (1 - v_1^*)\eta_1^{(s)} \\
v_2^* + (1 - v_2^*)\eta_2^{(s)} \\
\vdots \\
v_n^* + (1 - v_n^*)\eta_n^{(s)}
\end{pmatrix},
$$

$$
T = \begin{pmatrix}
I_{(y_1 > 0)\log(y_1)} \\
I_{(y_2 > 0)\log(y_2)} \\
\vdots \\
I_{(y_n > 0)\log(y_n)}
\end{pmatrix},
$$

(11)
and
\[ E = \begin{pmatrix} 0 & 0 \\ 0 & r^T 1_n - r^T r \end{pmatrix}. \]

where \(1_n\) denotes an \(n\)-dimensional column vector of 1.

**Proposition 2.** In the \((s + 1)\)-th iteration of the EM algorithm, the maximization of \(Q\)-function (10) with respect to \((\alpha, \sigma, \lambda)\) has an explicit solution given by

\[
\hat{\lambda}^{(s+1)} = \left[ \sum_{i=1}^{n} (1 - v_i^*) \right]^{-1} \sum_{i=1}^{n} (1 - v_i^*) \eta^{(s)}_i, \quad \hat{\alpha}^{(s+1)} = (B^T B + E)^{-1} B^T t,
\]

and

\[
(\hat{r}^2)^{(s+1)} = \frac{\sum_{i=1}^{n} I_{y_i > 0} \left[ \left( \log(y_i) - \hat{\mu}^{(s+1)}_{1,x_i} \right)^2 \left( v_i^* + (1 - v_i^*) \eta^{(s)}_i \right) + (1 - v_i^*) \left( 1 - \eta^{(s)}_i \right) \left( \log(y_i) - \hat{\mu}^{(s+1)}_{0,x_i} \right)^2 \right]}{\sum_{i=1}^{n} I_{y_i > 0}},
\]

where

\[
\hat{\mu}^{(s+1)}_{v,x} = \hat{\alpha}_0 + \sum_{j \in S} \hat{\alpha}_j x_{ij} + \hat{\alpha}_{k+1} v \quad \text{for} \quad v \in \{0, 1\}, \ i = 1, \ldots, n.
\]

Unfortunately, in the same vein as logistic regression, there is no explicit form available for the iterative formulas of \(\beta\). In order to obtain the update for the estimate of \(\beta\), we have to use numerical algorithms such as the Newton-type methods available in the R function \(nlm()\). On the other hand, at the expense of a restriction on the occurrence probability of non-zero claims which is deterministic, i.e., \(\pi_{v,x} = \pi_v\) for \(v \in \{0, 1\}\), then all the parameters will have explicit iterative formulas. Simplified models of this kind can be useful for preliminary analysis of misrepresentation risk, or when the rating factor exposed to misrepresentation is the only significant variable in the claim occurrence model.

**Corollary 3.** In the two-part misrepresentation models, if \(\pi_{v,x} = \pi_v = \beta_0 + \beta_1 v\) for \(v \in \{0, 1\}\), then the iterative formula for updating \(\beta\) is given by

\[
\hat{\beta}_0^{(s+1)} = \frac{\sum_{i=1}^{n} I_{y_i > 0} (1 - v_i^*) \left( 1 - \eta^{(s)}_i \right)}{\sum_{i=1}^{n} (1 - v_i^*) \left( 1 - \eta^{(s)}_i \right)} \quad \text{and} \quad \hat{\beta}_1^{(s+1)} = \frac{\sum_{i=1}^{n} I_{y_i > 0} \left( v_i^* + (1 - v_i^*) \eta^{(s)}_i \right)}{\sum_{i=1}^{n} \left( v_i^* + (1 - v_i^*) \eta^{(s)}_i \right)} - \hat{\beta}_0^{(s+1)}.
\]

The iterative formulas for \((\alpha^{(s+1)}, \sigma^{(s+1)}, \lambda^{(s+1)})\) are same as the ones in Proposition 2.

Based on the updated estimates from the M-step, we can then move on to the E-step for iteration \((s + 1)\). The EM algorithm iterates through the E-step and M-step, and stops until the improvement in the partial log likelihood between two consecutive steps is below a pre-specified threshold.
3.3 Large sample inference based on the observed Fisher information

As hitherto, we have derived the iterative formulas in the EM algorithm for estimating the parameters of the proposed two-part misrepresentation models. Another statistical aspect that is important for us to explore is the associated large-sample inference which plays a critical role in many insurance applications such as assessing the risk effects of rating factors, confidence interval construction, and hypothesis testing.

It is common in practice to approximate the covariance matrix of the maximum likelihood estimators (MLEs) \( \hat{\Psi} \) using the inverse of the observed Fisher information matrix denoted \( I(\hat{\Psi}) \). In order to derive the observed Fisher information matrix, one way to proceed is directly computing the Hessian matrix for the partial log likelihood function (8). However, as we have noted earlier, log likelihood function (8) is intractable to deal with. An alternative way to tackle the problem is by using the complete-data likelihood function in aid of the following well-known result.

**Lemma 4 (Louis, 1982).** Let \( U \) and \( W \) be the vectors of observed and complete random data, respectively. Denote by \( I(\Psi; U = u) \) the observed Fisher information matrix and by \( I_c(\Psi; U = u) \) the expected Fisher information matrix based on the complete-data likelihood function. Moreover, let \( S_c(W = w; \Psi) \) be the gradient vector of the complete-data log likelihood function. It holds that

\[
I(\Psi; U = u) = I_c(\Psi; U = u) - \text{Cov}(S_c(W; \Psi)|U = u). \tag{12}
\]

In order to present the observed Fisher information matrix for the proposed two-part misrepresentation models, let us introduce the following shorthand notations. Recall that “\( \cdot \parallel \cdot \)" is the vector concatenation operator. For \( i = 1, \ldots, n \), we define

\[
\begin{align*}
    d_{i1} &= -I_{(y_i > 0)}(1 - v_i^*)\alpha_{k+1}\sigma^{-2} (1\|x_{i1}^S); \\
    d_{i2} &= I_{(y_i > 0)}(1 - v_i^*)(\log(y_i) - \mu_{1,x_i^S})\sigma^{-2}; \\
    d_{i3} &= -(\pi_{1,x_i^F} - \pi_{0,x_i^F})(1 - v_i^*) (1\|x_{i1}^F); \\
    d_{i4} &= (I_{(y_i > 0)} - \pi_{1,x_i^F})(1 - v_i^*); \\
    d_{i5} &= -\frac{1}{2}I_{(y_i > 0)}(1 - v_i^*)\left[ (\log(y_i) - \mu_{0,x_i^S})^2 - (\log(y_i) - \mu_{1,x_i^S})^2 \right]/\sigma^4; \\
    d_{i6} &= (1 - v_i^*)\left[ \lambda(1 - \lambda) \right]^{-1}.
\end{align*}
\]

Furthermore, let \( b_i \) be the \( i \)-th row of \( B \) defined as per (11), \( c_i = \left[ (\log(y_i) - \mu_{1,x_i^S}) + \alpha_{k+1}(1 - v_i^*)(1 - \eta_i) \right] \), and
\[ q_i = \left[ (\log(y_i) - \mu_{1,x_i^q}) + \alpha_{k+1}(1-v_i^*)(1-\eta_i) \right], \quad i = 1, \ldots, n. \] Then we set

\[
A_{\alpha,i} = I(y_i>0)\sigma^{-2} b_i^T b_i; \quad A_{\beta,i} = \left( c_11_{|x_i^q|+1} \right) \left\{ (1 - \pi_{1,x_i^q})\pi_{1,x_i^q} \left[ v_i^* + (1-v_i^*)\eta_i \right] \right\} (1 ||x_i^q||1)(1 ||x_i^q||1); \\
a_{(\alpha,\sigma^2),i} = I(y_i>0)\sigma^{-4} \left\{ q_i1_{|S|+1} \left[ (\log(y_i) - \mu_{1,x_i^q})(v_i^* + (1-v_i^*)z_i) \right] \right\}; \\
a_{\sigma^2,i} = I(y_i>0)\sigma^{-6} \left[ (\log(y_i) - \mu_{1,x_i^q})^2(v_i^* + (1-v_i^*)\eta_i) + (\log(y_i) - \mu_{0,x_i^q})^2(1-v_i^*)(1-\eta_i) - 0.5\sigma^2 \right]; \\
a_{\lambda,i} = (1 - v_i^*) \left[ \lambda^{-2} - (1 - 2\lambda)[\lambda(1-\lambda)]^{-2}(1-\eta_i) \right].
\]

For \( i = 1, \ldots, n \), we let

\[
A_i = \begin{pmatrix}
A_{\alpha,i} & 0 & a_{(\alpha,\sigma^2),i} & 0 \\
0 & A_{\beta,i} & 0 & 0 \\
a_{(\alpha,\sigma^2),i} & 0 & a_{\sigma^2,i} & 0 \\
0 & 0 & 0 & a_{\lambda,i}
\end{pmatrix}.
\] (13)

The next assertion spells out the observed Fisher information matrix associated with partial log likelihood function (8).

**Proposition 5.** The observed Fisher information matrix associated with partial log likelihood function (8) has the form of representation:

\[
\mathcal{I}(\Psi) = \sum_{i=1}^{n} A_i - \sum_{i=1}^{n} (1 - \eta_i) d_i^T d_i,
\] (14)

where \( d_i = (d_{i1}, \ldots, d_{i6}) \) and \( \eta_i = \mathbb{E}[z_i \mid y_i, x_i, v_i^*, \Psi] \) for \( i = 1, \ldots, n. \)

**Remark 1.** Note that the covariance term in formula (14) corresponds to the missing information matrix (see, Louis, 1982; McLachlan and Krishnan, 2008) which is equal to zero if \( v_i^* = 1 \). This observation is intuitive based on the unidirectional feature of our misrepresentation problem. When \( v_i^* = 1 \), there is no misrepresentation, implying no missing information.

In concluding the current section, we make an important note about the comparison between the explicit and numerical approaches for computing the observed Fisher information matrix. Different from the proposed explicit approach, the numerical approach uses finite difference methods to approximate the Hessian matrix of partial log likelihood function (8). A variety of finite difference methods including the Secant method, three-pint central difference formula, complex-step derivative approximation and so forth, have been implemented in different R functions. Nevertheless, we argue that when there are explicit expressions available, they are more preferable than numerical approximations due to the following reasons. First, as finite difference methods are based on the Taylor expansions of the partial log likelihood function,
these methods are subject to quantization and/or truncation errors. Such numerical errors can be notoriously hard and sometimes even impossible to assess. Second, in the implementation of finite difference methods, one often needs to select the appropriate step sizes. The selection is rather subjective, and identifying the optimal step size is not a simple task. Third, recall that the misrepresentation parameter $\lambda$ in our model is bounded between 0 and 1. Therefore, our estimation problem should be viewed as a constrained optimization, while many Hessian matrix approximation methods implemented in \( \mathbb{R} \) are based on unconstrained optimization. Potential error may occur due to this reason, especially when the value of $\lambda$ is close to the boundaries (i.e., 0 or 1).

The explicit formulas we have established in this section do not suffer from the aforementioned numerical issues. What is more, the observed Fisher information matrix in Proposition 5 only contains simple matrix operations that are convenient to compute. We have put together the proposed estimation and inference processes into an \( \mathbb{R} \) package\(^1\). Therefore, the proposed two-stage models are immediately useful for insurance practitioners who are interested in modeling misrepresentation risk. It is noteworthy that, although the current study is motivated by insurance applications, the proposed models may also be useful in other fields where unidirectional misclassification is concerned. Examples include social desirability errors in survey sampling, financial and accounting fraud under the government and corporate settings, and self-reported drug consumption in medical research.

## 4 Simulation

In this section, we use a simulation study to validate the estimation and inference procedures proposed in Section 3. We assume there are four available rating factors \((X_1, X_2, X_3, V)\), in which \(V\) is unobservable and subject to misrepresentation. The observed counterpart of \(V\) is \(V^*\). We generate observed samples in accordance with the following assumptions: \(X_1 \sim \text{Normal}(0, 1)\), \(X_2 \sim \text{Bernoulli}(0.6)\), and \(X_3 \sim \text{Uniform}(0, 5)\). For the binary factor \(V\) that is subject to misrepresentation, we assume $\mathbb{P}[V = 1] = 0.2$, and use $\mathbb{P}[V^* = 0|V = 1] = 0.4$ to generate the reported variable \(V^*\). The associated misrepresentation prevalence is $\lambda = 0.09$. To generate the response variable \(Y\), we suppose the claim severity and occurrence are associated with different sets of rating factors, i.e., \(S = \{1, 2, 3\}\) and \(F = \{1, 2\}\). Table 4 summaries the values of regression parameters we use for the simulation study.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>(\alpha_0)</th>
<th>(\alpha_1)</th>
<th>(\alpha_2)</th>
<th>(\alpha_3)</th>
<th>(\sigma^2)</th>
<th>(\beta_0)</th>
<th>(\beta_1)</th>
<th>(\beta_2)</th>
<th>(\beta_v)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario 1</td>
<td>2.5</td>
<td>1.8</td>
<td>1.6</td>
<td>1.3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1.8</td>
<td>1.4</td>
</tr>
<tr>
<td>Scenario 2</td>
<td>2.5</td>
<td>1.8</td>
<td>1.6</td>
<td>1.3</td>
<td>0.25</td>
<td>2</td>
<td>1.8</td>
<td>1.4</td>
<td>1.2</td>
</tr>
<tr>
<td>Scenario 3</td>
<td>2.5</td>
<td>1.8</td>
<td>1.6</td>
<td>1.3</td>
<td>0.04</td>
<td>2</td>
<td>1.8</td>
<td>1.4</td>
<td>1.2</td>
</tr>
</tbody>
</table>

Table 1: Parameter values assumed for the simulation. The three scenarios correspond to the cases when the effect size of \(V\) is small \((\alpha_c/\sigma = 1)\), medium \((\alpha_c/\sigma = 2)\) and large \((\alpha_c/\sigma = 5)\).

In the simulation study, we are going to start with a sample of 100 and increase the sample size by a multiplicative

\(^1\)Available for download via [http://www.stat.purdue.edu/~jianxi/research.html](http://www.stat.purdue.edu/~jianxi/research.html)
factor of 4. Within each sample size, 150 sets of simulations are conducted. We use the EM algorithm from Section 3 to estimate the parameters $\alpha$, $\beta$, $\sigma^2$ and $\lambda$. The observed Fisher information matrix derived in Section 3.3 will be used to approximate the standard errors of the estimates. Figures 1 - 3 depict the distribution of the estimates and the corresponding standard errors based on the simulated data. As we can observe, the estimates for all the parameters converge to the true values with the increasing sample size, and the standard errors decay to zero. The three scenarios demonstrate how varying effect size of $V$ influences the performance of the estimators on $\alpha$. We observe that with the same sample size, the MLEs have higher precision (i.e., lower standard errors) when the effect size is larger. The estimators for $\beta$ perform approximately the same across the three scenarios since $\sigma$ is only associated with the claim severity distribution. As the sample size increases, the standard errors decay with a speed of $\sqrt{1/N}$ approximately, which coincides with the large-sample theories underlying MLEs.

5 Real applications

In this section, we present two empirical studies to demonstrate the practical usefulness of the proposed two-part misrepresentation models. Both studies are based on real data obtained respectively from a large public institution’s group health insurance and the Medical Expenditure Panel Survey (MEPS).

5.1 Group health insurance application

5.1.1 Data description

The data that we are going to consider in the first application, come from the group health insurance offered in a large public institution located in the US. The group health insurance consists of three plan choices having low/medium/high deductible, respectively. The premium rate is based on the plan tier (i.e., the deductible level), plan type (employee only, employee and spouse, employee and children, and employee and family) and salary level. Additionally, a surcharge of $500 will be applied to smoking adults (this is about 25% of the average premium rate). What makes the data special for studying misrepresentation risk is that the tobacco consumption status is acquired through a self-reporting process. In order to manage the potential fraudulent behaviors, the institution notifies the employees that misrepresentation may lead to employment termination. However, no medical screening is conducted to verify the smoking status. Because of the financial incentive towards the insureds, it is natural to suspect that the self-reported smoking status is subject to a high level of misrepresentation risk. In fact, according to the year 2016 data, there is a relatively low smoking percentage of about 10% in the group insurance (see, Table 2), while the average smoking rate is 21% in the state where the institution is located. Thereby, we are interested in assessing the following hypothesis in order to examine whether the notice on potential employment termination is sufficient for deterring misrepresentation behaviors.
Hypothesis 1: There exists a significant level of misrepresentation effect on the self-reported smoking status among the adult insureds of the group health insurance.

The group health insurance data contain detailed policy eligibility information and medical claim records. By means of detailed records, we refer to experiences at the individual levels, including biographical characteristics (e.g., age, gender, marriage, employment status, smoking history and so forth), plan selection, insurance coverage period, and claim experience by year. Each individual, family, and claim can be identified with unique policy, family and claim identifiers, and thus can be conveniently linked together. Although multiple years of data are recorded, due to substantial changes in the policy design over time, we place our focus on the year 2016 data which is most recent and reliable. Since our major
Figure 2: Violin plots for the estimates of $\beta$ and the associated standard errors (from left to right are respectively scenarios 1, 2, 3).

Figure 3: Violin plots for the estimates of $\sigma, \lambda$ and the associated standard errors (from left to right are respectively scenarios 1, 2, 3).
interest is to study misrepresentation on the smoking status, we only consider adults with age spanned from 24 to 65 who are required to report their tobacco usage. We further separate the individuals into two groups: a younger group (aged from 24 to 55) and an older group (aged from 56 to 65). Within the younger group, we observe that the average medical expense is higher for smokers compared with non-smokers. This finding is intuitive and coincides with the majority of empirical evidence reported in the health care literature. However, interestingly, a reversed relationship is observed within the older age group of insureds, indicating a lower average medical expense for smokers. More surprisingly, although we do not intend to report in details herein, a similar pattern is found in the widely used MEPS data. We believe the observation is very interesting and consider it as a promising topic for further research. Due to the aforementioned controversy, we confine our current study to the younger group of insurers whose difference in average claim amount between smokers and non-smokers is consistent with the related literature.

Table 2 outlines the definitions and descriptive statistics of the key variables recorded in the eligibility file of the data. We consider the year 2016 data and exclude the older age group. In light of the summary statistics, we may conclude that a typical policyholder in the data under investigation is a non-smoking female employee aged at 41 having the medium deductible plan.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>The age of the policyholder</td>
<td>41.02</td>
<td>8.23</td>
</tr>
<tr>
<td>Gender</td>
<td>=1 if the policyholder is female, 0 if male</td>
<td>0.54</td>
<td>0.50</td>
</tr>
<tr>
<td>Tobacco</td>
<td>=1 if the policyholder is smoker, 0 otherwise</td>
<td>0.11</td>
<td>0.31</td>
</tr>
<tr>
<td>Relation</td>
<td>=1 if the policyholder is employee, 0 if spouse</td>
<td>0.71</td>
<td>0.45</td>
</tr>
<tr>
<td>Plan code</td>
<td>=1 if the plan’s deductible is low, 2 if medium, 3 if high</td>
<td>2.11</td>
<td>0.67</td>
</tr>
</tbody>
</table>

Table 2: Summary statistics for the eligibility characteristics.

In order to illustrate the differences in the policyholder characteristics among varying plan choices, we exhibit in Table 3 the descriptive statistics of key variables by plan. Some observations are (a) the medium deductible plan is the most popular plan choice, accounting for more than half of the employees; (b) male employees at a younger age tend to choose an insurance plan with a higher deductible; (c) smokers tend to purchase a lower deductible plan, probably due to worse health status (i.e., anti-selection); (d) employees having a larger family tend to select a higher deductible plan probably for avoiding the high premium of the low deductible plan.

Finally, Figure 4 depicts the distribution for the logarithm of the annual total positive claims by plan. We observe that the empirical densities of the log annual claims are roughly bell shaped across three plans.

5.1.2 Data analysis

We aim to assess the misrepresentation on smoking status for the group insurance plan. To this end, we compare two types of claim analysis in what follows. In the adjusted analysis, we consider the two-part ratemaking models proposed in Section
Table 3: Summary statistics for eligibility characteristics by plan.

<table>
<thead>
<tr>
<th>Plan code</th>
<th>Variable</th>
<th>1 (low deductible)</th>
<th>2 (medium deductible)</th>
<th>3 (high deductible)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>Standard Deviation</td>
<td>Mean</td>
</tr>
<tr>
<td>Age</td>
<td>40.97</td>
<td>8.54</td>
<td></td>
<td>41.32</td>
</tr>
<tr>
<td>Gender</td>
<td>0.57</td>
<td>0.50</td>
<td></td>
<td>0.56</td>
</tr>
<tr>
<td>Tobacco</td>
<td>0.13</td>
<td>0.34</td>
<td></td>
<td>0.11</td>
</tr>
<tr>
<td>Relation</td>
<td>0.79</td>
<td>0.41</td>
<td></td>
<td>0.68</td>
</tr>
<tr>
<td>Number of obs</td>
<td>1747</td>
<td></td>
<td></td>
<td>5396</td>
</tr>
<tr>
<td>Percentage</td>
<td>17.47%</td>
<td></td>
<td></td>
<td>53.96%</td>
</tr>
</tbody>
</table>

Figure 4: The histogram of the logarithm of annual (positive) claims with an empirical density estimate overlaid (from left to right are plans 1, 2 and 3).

2 which can account for the potential misrepresentation effect. In the unadjusted analysis, we neglect misrepresentation by fixing $\lambda = 0$, under which the proposed misrepresentation models reduce to a two-part lognormal regression model. We choose the individual’s annual total medical claim amount (after adjusting for insurance deductible, co-insurance, policy limit, etc.) as the response variable. Setting the original medical expenditure as the response variable yields very similar results for our analysis and thus will not be reported separately herein. We use the unadjusted model to choose significant risk factors for modeling the claim occurrence and severity. The age, gender and tobacco status are chosen for both the claim occurrence and severity models, whereas the relation variable is only included for the claim severity model. Consequentially, for the two types of analysis, we set

$$S = \{\text{age, gender, relation, tobacco}\} \quad \text{and} \quad F = \{\text{age, gender, tobacco}\}.$$ 

Based on the nature of the risk factors, the tobacco status is the only variable that may be subject to misrepresentation.

We use the EM algorithm proposed in Section 3 to compute the MLEs for the two-part misrepresentation model. For brevity, we focus on the medium deductible plan as it consists of more than a half of the observations and thus would provide the most credible results. The results for the other two plans are similar and thus will not be presented here.
The point and interval estimates for the relativity of the risk factors are presented in Figures 5 - 6. Our observations are as follows. First, tobacco consumption has a positive effect on the size of annual insurance claims but a negative effect on the odds of positive insurance claim occurrence. Gender and age variables have similar effects on both claim size and occurrence. Second, compared with the unadjusted analysis in which potential misrepresentation is neglected, acknowledging misrepresentation in the adjusted analysis yields a larger variability in the estimated risk effect on the smoking variable. However, the difference in the estimates and variability of the other rating variables seems to be minor between the adjusted and unadjusted models.

Figure 5: The point and 95% confidence interval estimates for the relativity of $\alpha$ for the group health insurance data. The number displayed beside each confidence interval is the interval length.

Regarding the misrepresentation prevalence $\lambda$, the point estimate is 0.126 and the 95% confidence interval estimate is $(-0.552, 0.803)$. The high standard error of $\hat{\lambda}$ implies weak statistical learning on the misrepresentation parameter from the data. This suggests that misrepresentation may be insignificant in the data. However, the interval estimate of $\lambda$ may not be sufficient for drawing a reliable conclusion about misrepresentation, since the null value lies on the boundary of the parameter space. In order to obtain a more reliable statistical assessment, we may conduct a likelihood ratio test (LRT) between the adjusted and unadjusted models. If the LRT suggests the adjusted model does not outperform the

Figure 6: The point and 95% confidence interval (with interval length displayed) estimates for the relativity of $\beta$ for the group health insurance data.
unadjusted one, then we may confirm that misrepresentation is not significant and thus Hypothesis 1 should be rejected.

Due to the mixture structure involved in the proposed two-part misrepresentation model, the LRT statistic no longer follows a chi-squared distribution asymptotically. In order to determine the appropriate critical value for the LRT, we conduct a parametric bootstrap to approximate the null distribution of the LRT statistic. Precisely, based on the estimated parameters from the unadjusted model, we simulate 500 datasets of the same sample size for the response variable (i.e., the annual insurance claims) according to the conditional distribution given by the null model (i.e., the unadjusted model). Then both the adjusted and unadjusted models are fitted to each set of the simulated data, and the corresponding value of the LRT statistic is computed. The $p$-value of the LRT statistic on the original data can be approximated based on the empirical distribution function of the LRT statistics calculated using the simulated data. For the group health insurance data of interest, the LRT statistic has an approximate $p$-value of 39%. The LRT result shows little statistical evidence in support of Hypothesis 1, and we conclude that there is no significant misrepresentation effect in the data. Therefore, the existing strategy of potential employment termination seems to be effective for deterring misrepresentation frauds for the group health plan under consideration.

### 5.2 MEPS application

In the second application, we consider the misrepresentation risk in the MEPS data. The MEPS data consists of a sequence of surveys on the medical expenditure and health insurance status of the US population. Since the MEPS data has been widely used in the actuarial literature, we do not intend to provide a detailed description of it herein. Instead, we refer interested readers to, e.g., Bernard and Banthin (2006); Frees (2009); Hua (2015); Kashihara and Carper (2009).

In this subsection, the objective of our study is to assess the misrepresentation on the self-reported uninsured status. The respondents in MEPS may misrepresent on the uninsured status due to the enactment of the Patient Protection and Affordable Care Act (PPACA) in 2014, under which individuals are mandated to purchase health insurance in order to avoid tax penalties. Studies in Akakpo et al. (2018); Xia and Gustafson (2018) showed that there is no statistical evidence of misrepresentation on uninsured status from the 2012 and 2013 data prior to the implementation of the PPACA. We are interested in the year 2014 data in which misrepresentation is more likely to occur. We consider the following hypothesis:

**Hypothesis 2:** There exists a significant level of misrepresentation on the self-reported uninsured status among the MEPS respondents.

For our analysis, we choose the net total medical expenditure (after insurance and various discounts/waivers) as the response variable. The PPACA mandates that insurance companies can only set health insurance premium according to age, location, tobacco usage, plan type (individual/family), and level of coverage. Hence, considering the restriction imposed by the PPACA, we only select the variables that are typically available for insurance companies for ratemaking purposes. The risk factors in our analysis are gender, age, tobacco use and uninsured status. A preliminary analysis
using the unadjusted model shows that both the claim severity and occurrence are significantly associated with the aforementioned factors. Thus, we set

\[ S = F = \{ \text{age, gender, tobacco, uninsured status} \}. \]

For the MEPS data, the study in Xia and Gustafson (2016) suggested that the misrepresentation on smoking status is not significant. Therefore, in our current analysis, the uninsured status is considered as the only potentially misrepresented factor.

Figures 7 - 8 display the point and interval estimates for the relativity of the risk factors. With an estimated relativity below 1, the uninsured status has a negative effect on both the claim severity and the odds of claim occurrence. Similar to the group health insurance example, acknowledging misrepresentation in the adjusted analysis yields larger variability in the estimated risk effect for the uninsured variable. In addition, the adjusted model gives a larger negative relative effect (a value further away from 1) for the uninsured status, revealing a larger impact from the adjustment of misrepresentation. The difference in both the estimates and variability is also quite noticeable for the other rating factors. This is different than the variability patterns in the previous example where misrepresentation is insignificant. The point estimate for the misrepresentation prevalence is \( \hat{\lambda} = 0.08 \) and the 95% confidence interval estimate is (0.002, 0.163). The observations noted above seem to support Hypothesis 2.

![Figure 7: The point and 95% confidence interval (with length displayed) estimates for the relativity of \( \alpha \) for the MEPS data.](image)

Similar to the analysis of the group health insurance data, the LRT is used to identify the statistical evidence of misrepresentation from the MEPS data. In this analysis, the parametric bootstrap method yields a p-value of about 2% for the LRT. Therefore, Hypothesis 2 should be accepted at the 95% significance level. There is a significant statistical evidence to support the presence of misrepresentation on the uninsured status of the MEPS respondents. Our empirical finding is consistent with the one from Akakpo et al. (2018) that is drawn without considering other risk factors or the zero-inflated feature of the data. With the aids of a more comprehensive misrepresentation risk modeling framework.
proposed in this current article, our conclusion is statistically more reliable than the one in loc. cit. The proposed two-part misrepresentation models are also more appropriate for insurance ratemaking applications where there are multiple rating factors and zero-inflation for the aggregate claim amount.

6 Conclusions and future work

In this article, we considered a class of two-part loss models adjusting for misrepresentation. Frequentist estimation based on the EM algorithm was proposed for the implementation of two-part misrepresentation models. Explicit formulas are derived for the E-step and M-step of the EM algorithm, as well as the observed Fisher information matrix for statistical inference purposes. Simulation results show that the EM algorithm provides reliable convergence performance for the proposed models. On the one hand, the proposed models can be used to assess misrepresentation risk. On the other hand, they contribute to an improved GLM loss modeling framework in which the impact of misrepresentation is accounted for.

Moving forward, there are a number of avenues to pursue in future research. First, in addition to lognormal severity distribution assumed for deriving explicit forms, we may consider alternative distributions widely used in the GLM modeling, including Gamma, Pareto, Weibull, etc. In these extended models, explicit expressions can no longer be guaranteed in the estimation and inference processes. We will seek stable and efficient numerical methods to tackle the problem. Second, in this paper, we assumed the prevalence of misrepresentation to be constant. Thereby, our models do not provide prediction of the misrepresentation probability at the policy level. In future research, we can incorporate latent binomial regression models on the misrepresentation prevalence. In so doing, the resulting models can help insurance companies to identify the policies that are most vulnerable to misrepresentation, based on the individual’s rating factors. Third, the current proposed models can handle only one potentially misrepresented risk factor. We are interested in extending the models to handle multiple misrepresented risk factors. The challenges for studying such an extension are the models’ identifiability and the dependencies among multiple misrepresented risk factors. There are many potential

Figure 8: The point and 95% confidence interval (with length displayed) estimates for the relativity of $\beta$ for the MEPS data.
applications for these generalized models. For instance, we can take into consideration of passive (i.e., second-hand) smoking effect in the analysis of misrepresentation on family smoking status. Another potential application is to study the misrepresentation on the smoking status and uninsured status jointly in the MEPS data.

References


A Technical proofs

Proof of Proposition 2. The derivation for all three formulas hinges on the partial derivative of Q-function (10). We begin with the iterative formula for λ. Setting

$$\frac{\partial}{\partial \lambda} Q \left( \Psi \mid \Psi^{(s)} \right) = \sum_{i=1}^{n} (1 - v_i^*) \left[ \eta_i^{(s)} - \frac{1}{1 - \lambda} + \eta_i^{(s)} \right] = 0$$

yields \( \hat{\lambda}^{(s+1)} = \sum_{i=1}^{n} (1 - v_i^*) \eta_i^{(s)} / \sum_{i=1}^{n} (1 - v_i^*) \).

Turning to the estimator for \( \alpha \), we first recall

$$\log(f_v(y)) = I_{(y > 0)} \log \left[ \pi_{v,x} g(y; \mu_{v,x}, \sigma) \right] + I_{(y = 0)} \log(1 - \pi_{v,x}), \ y \geq 0, \ v = 0,1.$$  

For expositional reasons, assume \( x_0 = 1 \). It is straightforward that, for \( v \in \{0,1\} \) and \( j \in \{0,S\}, \)

$$\frac{\partial \log(f_v(y))}{\partial \alpha_j} = I_{(y > 0)} \left( \frac{\log(y) - \mu_{v,x,S}}{\sigma^2} \right) \frac{\partial \mu_{v,x,S}}{\partial \alpha_j} = I_{(y > 0)} \left( \frac{\log(y) - \mu_{v,x,S}}{\sigma^2} \right) x_j,$$

and

$$\frac{\partial \log(f_v(y))}{\partial \alpha_{k+1}} = \begin{cases} 0, & v = 0 \\ I_{(y > 0)} \left( \frac{\log(y) - \mu_{v,x,S}}{\sigma^2} \right), & v = 1. \end{cases}$$

Simple algebraic operation yields the following partial derivative formula:

$$\frac{\partial}{\partial \alpha_j} Q \left( \Psi \mid \Psi^{(s)} \right) \propto \sum_{i=1}^{n} I_{(y_i > 0)} v_i^* (\log(y_i) - \mu_{1,x_i}) x_{ij} + \sum_{i=1}^{n} I_{(y_i > 0)} (1 - v_i^*) \left[ (1 - \eta_i^{(s)}) \log(y_i) - \mu_{0,x_i}) x_{ij} + \eta_i^{(s)} (\log(y_i) - \mu_{1,x_i}) x_{ij} \right]$$

$$= \sum_{i=1}^{n} I_{(y_i > 0)} \log(y_i) x_{ij} - \sum_{i=1}^{n} \sum_{h \in S} I_{(y_i > 0)} x_{ij} x_{ih} - \alpha_{k+1} \sum_{i=1}^{n} I_{(y_i > 0)} x_{ij} \left( v_i^* + (1 - v_i^*) \eta_i^{(s)} \right),$$
for \( j \in \{0, S\} \), and

\[
\frac{\partial}{\partial \alpha_{k+1}} Q \left( \Psi \mid \Psi^{(s)} \right) \propto \sum_{i=1}^{n} I_{(y_i > 0)} v_i^* (\log(y_i) - \mu_{1,x_i^S}) + \sum_{i=1}^{n} I_{(y_i > 0)} (1 - v_i^*) \left[ \eta_i^{(s)} (\log(y_i) - \mu_{1,x_i^S}) \right]
\]

\[
= \sum_{i=1}^{n} I_{(y_i > 0)} \log(y_i) \left( v_i^* + (1 - v_i^*) \eta_i^{(s)} \right) - \sum_{j \in \{0, S, k+1\}} \alpha_s \sum_{i=1}^{n} I_{(y_i > 0)} x_{ij} \left( v_i^* + (1 - v_i^*) \eta_i^{(s)} \right).
\]

Setting \( \frac{\partial}{\partial \alpha_{j}} Q \left( \Psi \mid \Psi^{(s)} \right) = 0 \) for \( j \in \{0, S, k+1\} \) forms a system of \((|S| + 2)\) linear equations with the same number of unknowns. The linear equation system can be solved via \((B^T B + E)^{-1} B^T \mathbf{t}\). This yields the estimator for \( \alpha \).

Finally, set \( \frac{\partial}{\partial \pi^*} Q \left( \Psi \mid \Psi^{(s)} \right) = 0 \). With some simple algebraic calculations, we obtain

\[
(\hat{\sigma}^2)^{(s+1)} = \frac{\sum_{i=1}^{n} \left[ v_i^* (\log(y_i) - \mu_{1,x_i^S})^2 + (1 - v_i^*) \left( 1 - \eta_i^{(s)} \right) (\log(y_i) - \mu_{0,x_i^S})^2 + \eta_i^{(s)} (\log(y_i) - \mu_{1,x_i^S})^2 \right] I_{(y_i > 0)}}{\sum_{i=1}^{n} I_{(y_i > 0)}}.
\]

This completes the proof of the proposition.

\[\square\]

**Proof of Corollary 3.** Under the simplified assumption \( \pi_{v,x} = \pi_v = \beta_0 + \beta_1 v \) for \( v = 0, 1 \), we first aim to find \( \hat{\pi}_0 \) and \( \hat{\pi}_1 \) that maximize Q-function (10). Note that

\[
\frac{\partial}{\partial \pi_v} \log(f_v(y)) = I_{(y>0)} \frac{1}{\pi_v} - (1 - I_{(y>0)}) \frac{1}{1 - \pi_v}, \quad v = 0, 1.
\]

Set

\[
\frac{\partial}{\partial \pi_0} Q \left( \Psi \mid \Psi^{(s)} \right) = \sum_{i=1}^{n} (1 - v_i^*) \left( 1 - \eta_i^{(s)} \right) \left( I_{(y_i > 0)} \frac{1}{\pi_0} - (1 - I_{(y_i > 0)}) \frac{1}{1 - \pi_0} \right) = 0,
\]

and

\[
\frac{\partial}{\partial \pi_1} Q \left( \Psi \mid \Psi^{(s)} \right) = \sum_{i=1}^{n} (v_i^* + (1 - v_i^*) \eta_i^{(s)}) \left( I_{(y_i > 0)} \frac{1}{\pi_1} - (1 - I_{(y_i > 0)}) \frac{1}{1 - \pi_1} \right) = 0.
\]

We get

\[
\hat{\pi}_0 = \frac{\sum_{i=1}^{n} I_{(y_i > 0)} (1 - v_i^*) \left( 1 - \eta_i^{(s)} \right)}{\sum_{i=1}^{n} (1 - v_i^*) \left( 1 - \eta_i^{(s)} \right)} \quad \text{and} \quad \hat{\pi}_1 = \frac{\sum_{i=1}^{n} I_{(y_i > 0)} \left( v_i^* + (1 - v_i^*) \eta_i^{(s)} \right)}{\sum_{i=1}^{n} \left( v_i^* + (1 - v_i^*) \eta_i^{(s)} \right)}.
\]

The iterative formula for \( \beta \) can be obtained via the one-to-one correspondence between \((\beta_0, \beta_1)\) and \((\pi_0, \pi_1)\). The proof is completed.

\[\square\]

**Proof of Proposition 5.** For ease of presentation, we only report the Fisher information matrix for the \( i \)-th sample, \( i = 1, \ldots, n \). The associated complete-data log likelihood function is denoted by \( l_{c,i} \). With an observed sample of size \( n \), we can compute the observed Fisher information matrix by summing the individual information matrices over the \( n \) observations.

First, we study the expected Fisher information matrix associated with the complete-data log likelihood function used
Fisher information matrix associated with the complete-data log likelihood function can be now constructed according to the EM algorithm. To this end, we need the expected second derivatives of the complete-data log likelihood function. Tedious yet manageable calculations yield

\[-E[\partial^2 l_{c,i}/\partial \alpha_j \partial \alpha_l] = I_{(y_i > 0)} \sigma^{-2} x_{ij} x_{ih}, \ j, h \in \{0, S\};\]

\[-E[\partial^2 l_{c,i}/\partial \alpha_j \partial \alpha_{k+1}] = I_{(y_i > 0)} \sigma^{-2} x_{ij} [v_i^*(1 - v_i^*) \eta_i], \ j \in \{0, S, k + 1\};\]

\[-E[\partial^2 l_{c,i}/\partial \beta_j \partial \beta_l] = x_{ij} x_{ih} \left[(1 - \pi_{1,x_i^p}) \pi_{1,x_i^p} [v_i^*(1 - v_i^*) \eta_i] + \left[(1 - \pi_{0,x_i^p}) \pi_{0,x_i^p}\right] (1 - v_i^*)(1 - \eta_i)\right], \ j, h \in \{0, F\};\]

\[-E[\partial^2 l_{c,i}/\partial \beta_j \partial \beta_{k+1}] = x_{ij} \left[(1 - \pi_{1,x_i^p}) \pi_{1,x_i^p} \right] [v_i^*(1 - v_i^*) \eta_i], j \in \{0, F, k + 1\};\]

\[-E[\partial^2 l_{c,i}/\partial \alpha_j \partial \sigma^2] = I_{(y_i > 0)} \sigma^{-4} x_{ij} \left[(\log(y_i) - \mu_{1,x_i^p}) + \alpha_{k+1}(1 - v_i^*)(1 - \eta_i)\right], j \in \{0, S\};\]

\[-E[\partial^2 l_{c,i}/\partial \alpha_{k+1} \partial \sigma^2] = I_{(y_i > 0)} \sigma^{-4} \left[(\log(y_i) - \mu_{1,x_i^p})(v_i^*(1 - v_i^*) \lambda_i)\right];\]

\[-E[\partial^2 l_{c,i}/\partial \beta_j \partial \sigma^2] = I_{(y_i > 0)} \sigma^{-6} \left[(\log(y_i) - \mu_{1,x_i^p})^2 (v_i^*(1 - v_i^*) \eta_i) + \left[(\log(y_i) - \mu_{0,x_i^p})^2 (1 - v_i^*)(1 - \eta_i) - 0.5 \sigma^2\right] ;\]

\[-E[\partial^2 l_{c,i}/\partial \lambda_1 \partial \lambda_l] = (1 - v_i^*) \left[\lambda^{-2} - (1 - 2\lambda) \lambda(1 - \lambda) \right]^{-2} (1 - \eta_i).\]

The expected second derivatives are equal to zero otherwise. By using the aforementioned derivative formulas, the Fisher information matrix associated with the complete-data log likelihood function can be now constructed according to Equation (13).

Next, we study the covariance matrix of the gradient vector of the complete data log likelihood function. We again set \(x_{i0} = 1, i = 1, \ldots, n\), for notational convenience. It holds that

\[\partial l_{c,i}/\partial \alpha_j = I_{(y_i > 0)} \sigma^{-2} x_{ij} \left[(\log(y_i) - \mu_{1,x_i^p}) + \alpha_{k+1}(1 - v_i^*)(1 - \eta_i)\right], j \in \{0, S\};\]

\[\partial l_{c,i}/\partial \alpha_{k+1} = I_{(y_i > 0)} \sigma^{-2} \left[(\log(y_i) - \mu_{1,x_i^p})(v_i^*(1 - v_i^*) \eta_i)\right];\]

\[\partial l_{c,i}/\partial \beta_j = x_{ij} \left[I_{(y_i > 0)} (1 - \pi_{1,x_i^p}) - I_{(y_i = 0)} \pi_{1,x_i^p}\right] + \pi_{1,x_i^p} - \pi_{0,x_i^p}\right] (1 - v_i^*)(1 - \eta_i), j \in \{0, F\};\]

\[\partial l_{c,i}/\partial \beta_{k+1} = \left[I_{(y_i > 0)} (1 - \pi_{1,x_i^p}) - I_{(y_i = 0)} \pi_{1,x_i^p}\right] (v_i^*(1 - v_i^*) \eta_i);\]

\[\partial l_{c,i}/\partial \sigma^2 = \frac{1}{2} I_{(y_i > 0)} \sigma^{-4} \left[(\log(y_i) - \mu_{1,x_i^p})^2 - \sigma^2 + \left[(\log(y_i) - \mu_{0,x_i^p})^2 - (\log(y_i) - \mu_{1,x_i^p})^2\right] (1 - v_i^*)(1 - \eta_i)\right];\]

\[\partial l_{c,i}/\partial \lambda = (1 - v_i^*) \left[\lambda^{-1} - (\lambda(1 - \lambda))^{-1} (1 - \eta_i)\right].\]

Note that all the partial derivatives reported above are linear in \(z_i\). Thereby, we have readily got

\[\text{Cov}(S_{c,i}(z_i; \Psi)|y_i, x_i, v_i^*) = \text{Var}(z_i) \ b_i^T b_i = \eta_i(1 - \eta_i) \ b_i^T b_i.\]
The application of the observed Fisher information formula in Lemma 4 completes the proof for the proposition.

B Bayesian implementation

For the Bayesian implementation, we may use the MCMC methods based on the complete-data log likelihood function (9). Either Gibbs sampling or the Metropolis-Hastings algorithm can be used for the posterior simulations. References such as Hurn et al. (2003); McLachlan and Peel (2000) give comprehensive reviews of such algorithms for mixture models.

In the current paper, we focus on the implementation of the two-part misrepresentation models using the BUGS language (Lunn et al., 2000). Owing to the excellent introductions by Scollnik (2001, 2002), the BUGS language has been widely used in the actuarial literature for implementing Bayesian models.

For the sake of illustration, let us consider the setting in Section 4 where there are four rating factors \((X_1, X_2, X_3, V)\) with \(S = \{1,2,3\}\) and \(F = \{1,2\}\). Rating factor \(V\) is subject to misrepresentation. For the parameters in \(\Psi\), we assume normal \(\text{Normal}(0,10)\) priors for the regression coefficients in \(\alpha\) and \(\beta\), an inverse gamma prior \(\text{IG}(0.001,0.001)\) for the shape parameter \(\sigma\), and a \(\text{Uniform}(0,1)\) prior for the misrepresentation prevalence parameters \(\lambda\) and \(\theta = P[V = 1]\).

Such a non-informative prior specification represents a situation where we have no prior knowledge on the parameters.

The following BUGS implementation of the two-part misrepresentation model utilizes the \textit{ones trick} in specifying the complete-data log likelihood. In the attempt to maximize the likelihood function, in the ones trick a Bernoulli trial is assumed for a vector of ones, with the probability of each observation being the likelihood divided by a large number.

```plaintext
model {

  # A large constant for the ones trick
  C <- 10000

  for (i in 1:n) {

    # Model for V_star
    V_star[i] ~ dbin(theta_star,1)
    V[i] <- V_star[i] + (1-V_star[i])*Z[i]
    Z[i] ~ dbin(lambda,1)

    # Define a dummy variable on whether Y > 0.01
    G[i] <- step(Y[i] - 0.01)
    logit(pi[i]) <- beta0 + beta1*X1[i] + beta2*X2[i] + betav*V[i]

  }
}
```
# Lognormal regression

\[ \mu[i] \leftarrow \alpha_0 + \alpha_1 \cdot X_1[i] + \alpha_2 \cdot X_2[i] + \alpha_3 \cdot X_3[i] + \alpha_v \cdot V[i] \]

\[ \logLN[i] \leftarrow \log(\text{dlnorm}(Y[i], \mu[i], \tau)) \]  # log likelihood of lognormal distribution

# Define the total likelihood

\[ \logLik[i] \leftarrow (1 - G[i]) \cdot \log(1 - \pi[i]) + G[i] \cdot (\log(\pi[i]) + \logLN[i]) \]

\[ \text{Lik}[i] \leftarrow \exp(\logLik[i]) \]

# Use the ones trick for the likelihood

\[ pp[i] \leftarrow \text{Lik}[i] / C \]

\[ \text{ones}[i] \sim \text{dbern}(pp[i]) \]

}

# Relationships of probabilities

\[ \theta_* \leftarrow \theta \cdot (1 - p) \]

\[ \lambda \leftarrow \theta \cdot p / (1 - \theta \cdot (1 - p)) \]

# Prior distributions

\[ \theta \sim \text{dunif}(0, 1) \]

\[ \alpha_0 \sim \text{dnorm}(0, 0.1) \]

\[ \alpha_1 \sim \text{dnorm}(0, 0.1) \]

\[ \alpha_2 \sim \text{dnorm}(0, 0.1) \]

\[ \alpha_3 \sim \text{dnorm}(0, 0.1) \]

\[ \alpha_v \sim \text{dnorm}(0, 0.1) \]

\[ p \sim \text{dunif}(0, 1) \]

\[ \tau \sim \text{dgamma}(0.001, 0.001) \]

\[ \beta_0 \sim \text{dnorm}(0, 0.1) \]

\[ \beta_1 \sim \text{dnorm}(0, 0.1) \]

\[ \beta_2 \sim \text{dnorm}(0, 0.1) \]

\[ \beta_v \sim \text{dnorm}(0, 0.1) \]
For the application studies, the Bayesian implementation gives similar results on the estimation of the parameters and their standard errors. Hence, we will not present the results here. For the simulation study, Bayesian MCMC is much slower than the frequentist methods. So, running repeated simulations with large sample sizes is computationally impossible. Thus, when there is no prior knowledge on the misrepresentation behaviors, the frequentist methods discussed in Section 3 are more convenient for the implementation of the proposed two-part misrepresentation models.