Astin Bulletin (2026), **56**, pp. 720–741
doi:10.1017/asb.2026.10115

International Actuarial Association
Association Actuarielle Internationale

RESEARCH ARTICLE

High-dimensional claim severity modeling with misrepresentation adjustment via deep learning

Pengcheng Zhang[1], Jianwei Gang[2], Shilong Li[1] and Xiaoyan Wang[3]

[1]School of Insurance, Shandong University of Finance and Economics, China
[2]School of Statistics and Data Science, Zhejiang Gongshang University, China
[3]College of Finance and Statistics, Hunan University, China
Corresponding author: Jianwei Gang; Email: gangjw@mail.zjgsu.edu.cn

Received: 29 November 2025 UTC; **Revised:** 8 June 2026 UTC; **Accepted:** 20 July 2026 UTC

Keywords: Claim severity; misrepresentation; deep learning; variable selection; KL divergence

Abstract

In insurance underwriting, applicants may deliberately misrepresent certain risk factors to secure lower premiums. However, from the insurer's perspective, verifying the true status of these risk factors is both time-consuming and resource-intensive. To address this challenge, we propose a novel deep learning framework for claim severity modeling that explicitly accounts for potential misrepresentation of a binary risk factor. To mitigate the identifiability issues arising from the unobservable nature of misrepresentation in neural networks, we introduce a regularization term based on Kullback–Leibler (KL) divergence. This term serves as a conservative anchor by penalizing deviations from a baseline assumption of honesty, unless such deviations are strongly supported by empirical evidence. Additionally, we introduce an innovative regularization term, the L_1 group minimax concave penalty (L_1-gMCP), within the neural network's loss function. This term facilitates effective variable selection in high-dimensional settings. Through comprehensive simulation studies, we demonstrate that our method maintains robust performance regardless of whether misrepresentation is present in the data. The proposed model excels in both prediction accuracy and the identification of relevant risk factors. We further validate our approach using real-world data from the 2014 Medical Expenditure Panel Survey, treating reported insurance status as a potentially misrepresented variable.

1. Introduction

In actuarial modeling, the intentional misrepresentation of risk profiles by policyholders poses a significant challenge. This issue is especially pronounced in health insurance, where applicants may conceal their smoking habits to obtain lower premiums, and in automobile insurance, where it is common to underreport annual mileage, particularly within digital underwriting platforms. From the insurer's perspective, the verification of such self-reported information can often be prohibitively expensive. Consequently, there is an urgent need for more effective methods to assess and validate risk profiles to mitigate this challenge.

Recent scholarly work has made significant strides in addressing misrepresentation within statistical frameworks. The foundational study by Xia and Gustafson (2016) established the identifiability conditions for Bayesian regression models with misclassified binary covariates that are unidirectional, where "unidirectional" refers to measurement errors that systematically bias responses in favor of policyholders' interests. This theoretical framework was later extended by Sun *et al*. (2017), who demonstrated moment identifiability for models incorporating misclassified ordinal covariates. Building on these advancements, Akakpo *et al*. (2019) adopted a frequentist approach, assuming log-normal distributions for the underlying loss variables.

Notably, the existing literature remains limited by its exclusive reliance on linear modeling frameworks. Given the complexity of nonlinear relationships between response variables and explanatory

© The Author(s), 2026. Published by Cambridge University Press on behalf of The International Actuarial Association.

https://doi.org/10.1017/asb.2026.10115 Published online by Cambridge University Press

factors, there is a compelling need to adopt more sophisticated analytical tools, particularly deep learning methodologies. Over the past decade, the actuarial science community has increasingly recognized the superior predictive capabilities of deep learning, particularly in capturing unspecified nonlinear effects and complex interactions. These advantages have led to successful applications in various insurance domains, including nonlife insurance pricing (Noll *et al*., 2020; Delong *et al*., 2021), claims reserving (Wüthrich, 2018; Al-Mudafer *et al*., 2022), mortality forecasting (Hainaut, 2018), telematics data analysis (Gao *et al*., 2019), and fraud analytics (Gomes *et al*., 2021), etc. A comprehensive review of the application of deep learning in actuarial science can be found in Richman (2021a) and (2021b). However, our comprehensive literature review uncovers that no existing study has developed specialized deep learning architectures for claims prediction that account for systematic misrepresentation in insurance data. This gap serves as the motivation for our current investigation.

Additionally, insurance datasets typically encompass a wide range of features, including policyholder demographics, coverage details, and historical claim records. The high dimensionality of input variables, combined with excessive noise in the data, complicates the identification of meaningful predictive signals. In linear models, regularization is a common technique for selecting significant predictors, with examples including the Least Absolute Shrinkage and Selection Operator (LASSO) penalty (Tibshirani, 1996), the Smoothly Clipped Absolute Deviation (SCAD) penalty (Fan and Li, 2001), and the Minimax Concave Penalty (MCP) (Zhang, 2010). However, these standard regularization techniques, which penalize individual weights, fail to fully eliminate a feature in deep neural networks (DNNs). This limitation arises because each input feature in a DNN typically connects to multiple neurons in the first hidden layer. To effectively "remove" a feature, all of its associated outgoing weights must be driven to zero simultaneously. To address this issue, researchers have proposed the group LASSO method, which employs an $L_{2,1}$-norm penalty to group the weights connecting a single input feature to all neurons in the first hidden layer. This approach encourages the weights within a group to be either entirely zero or nonzero, thereby achieving feature-level selection. This method has been explored by Scardapane *et al*. (2017) and Zhang *et al*. (2019), and its application to claims modeling in neural networks has been investigated by Shi and Shi (2024).

Building on these insights, we propose a novel approach that incorporates an L_1-gMCP into the weights connecting the input layer to the first hidden layer of the network architecture. This approach strategically combines the group MCP with LASSO regularization, enhancing the model's capability to perform feature selection and improve overall performance. This hybrid penalty offers two key advantages: (1) it preserves the sparsity-inducing properties of LASSO while mitigating its well-documented bias issues through the concave regularization of MCP and (2) it respects the natural grouping of network parameters, thereby enabling effective feature selection. For model inference, we develop a specialized proximal gradient method to efficiently address the nonconvex optimization problem arising from this composite penalty. This optimization framework ensures computational tractability while maintaining model performance.

Estimating latent misrepresentation variables in the absence of ground-truth labels presents a substantial challenge. In such cases, without proper constraints, the neural network is prone to identifiability issues, where random noise might be erroneously interpreted as fraudulent behavior. To address this concern, we introduce a Kullback–Leibler (KL) divergence-based regularization term. Rather than imposing a fixed prevalence of misrepresentation, this term functions as a sparsity constraint, anchoring the model's baseline assumption to honesty. This design mirrors the "presumption of innocence" principle in insurance underwriting, where penalizing applicants without robust evidence could lead to significant compliance risks and unjust premium penalties. By incorporating this regularization, we encourage the model to predict misrepresentation only when the input features provide sufficiently strong signals that override the assumption of honesty. This approach not only enhances the statistical plausibility of the model's predictions but also enhances fairness in underwriting decisions.

Our paper makes several key contributions to the existing literature. First, we introduce a comprehensive approach to modeling misrepresentation in high-dimensional settings. In addition to extending

https://doi.org/10.1017/asb.2026.10115 Published online by Cambridge University Press

the model within a linear framework, we propose a neural network architecture enhanced by our novel L_1-gMCP penalty for structured variable selection. Additionally, we incorporate a carefully designed KL divergence regularization term as a sparsity constraint, which anchors the model to a baseline assumption of honesty. Second, we develop an efficient optimization framework for our proposed neural network, leveraging local linear approximation (LLA) and proximal gradient descent (PGD). A detailed procedure for implementation is provided. Third, we demonstrate the superior performance of our proposed model over conventional approaches across multiple evaluation metrics, supported by extensive simulation studies. A real-world case study using health expenditure survey data is presented, offering valuable insights for practitioners in the insurance sector.

The remainder of this paper is organized as follows. Section 2 formulates the misrepresentation problem and develops our high-dimensional modeling approach within conventional statistical frameworks. Section 3 extends this framework to deep learning environments, providing a detailed description of the proposed architecture and learning strategy. Section 4 assesses the performance of the models through two comprehensive simulation studies. Section 5 applies the proposed methodology to analyze health expenditure data from the 2014 MEPS. Finally, Section 6 offers concluding remarks.

2. Conventional model

2.1. Specification

In insurance pricing and claims models, the misrepresentation of risk factors presents a significant challenge that can result in inaccurate premium calculations and inappropriate claim payments. Traditional statistical models typically presume that these risk factors are reported accurately; however, policyholders may intentionally conceal or misrepresent adverse information to secure lower premiums or higher payouts. Consequently, the development of models that can address and rectify the biases introduced by such misrepresentations is a critical objective in the insurance sector.

In our analysis, in addition to the conventional covariates, we introduce a binary risk factor v that is subject to misrepresentation. For this predictor, we assume that $v = 0$ corresponds to an outcome favorable to the individual (i.e. lower premium), whereas $v = 1$ leads to an unfavorable result. Due to the misstatement on this risk factor, we denote v^* to be the observed and reported status. Under the assumption of unidirectional misrepresentation, we have the following relationships:

$$\Pr(v = 1|v^* = 1) = 1, \quad \Pr(v = 1|v^* = 0) = q, \quad 0 < q < 1, \tag{2.1}$$

where q is referred to as the prevalence of misrepresentation (Akakpo *et al*., 2019). It is further assumed that, conditional on the true status v, the response variable y is independent of the reported status v^*. The probability function that accounts for misrepresentation then becomes:

$$f_Y(y|\boldsymbol{x}, v^*) = \left[f_1(y)\right]^{v^*} \left[qf_1(y) + (1-q)f_0(y)\right]^{1-v^*}, \tag{2.2}$$

where

$$f_1(y) = f_Y(y|\boldsymbol{x}, v = 1), \tag{2.3}$$

$$f_0(y) = f_Y(y|\boldsymbol{x}, v = 0), \tag{2.4}$$

and $\boldsymbol{x} = (x_1, \ldots, x_p)^\top$ denotes the covariate set.

The quantity of interest in our model is the log-transformed claim severity. In the linear framework, we assume that it follows a normal distribution, which aligns with the characteristics of the data and facilitates the analysis. The mean parameter μ of the normal distribution is incorporated into the model through an identity link function:

$$\mu = \beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta} + \beta_v v, \tag{2.5}$$

where β_0, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ and β_v represent the corresponding regression coefficients. We assume that the variance σ^2 of the normal distribution is known; otherwise, it can be inferred from the data.

https://doi.org/10.1017/asb.2026.10115 Published online by Cambridge University Press

To explore the potential impact of predictors on the prevalence parameter q, covariates can be included in q using a logit link function:

$$q = \frac{\exp\left(\alpha_0 + \boldsymbol{x}^\top \boldsymbol{\alpha}\right)}{1 + \exp\left(\alpha_0 + \boldsymbol{x}^\top \boldsymbol{\alpha}\right)}, \tag{2.6}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^\top$.

The individual likelihood function is then given by

$$L_i(\boldsymbol{\Theta}; y_i, \boldsymbol{x}_i, v_i^*) = \left[f_1(y_i)\right]^{v_i^*} \left[q_i f_1(y_i) + (1 - q_i) f_0(y_i)\right]^{1 - v_i^*}, \tag{2.7}$$

where $\boldsymbol{\Theta} = \{\beta_0, \boldsymbol{\beta}, \beta_v, \alpha_0, \boldsymbol{\alpha}\}$ represents the parameter set to be estimated.

2.2. High-dimensional modeling

The model specification above provides a flexible framework for accommodating misrepresentation in claims data. However, when the number of covariates p is large, the inclusion of all predictors in both the mean and prevalence components may lead to overfitting and diminished interpretability. Moreover, not all covariates necessarily influence the misrepresentation behavior or the claim severity, motivating the need for variable selection. To this end, we adopt a penalized likelihood approach that simultaneously performs parameter estimation and variable selection. Specifically, we impose L_1 penalty terms on the coefficients in both the mean and prevalence components, which encourage sparse solutions by shrinking irrelevant coefficients to zero. This formulation enables us to identify key risk factors that affect claim severity and those that drive misrepresentation behavior, while maintaining model parsimony in high-dimensional settings. The overall optimization problem is formulated as:

$$\underset{\boldsymbol{\Theta}}{\text{minimize}} -\frac{1}{N} \sum_{i=1}^{N} \log L_i(\boldsymbol{\Theta}; y_i, \boldsymbol{x}_i, v_i^*) + \lambda_Q \sum_{j=1}^{p} |\alpha_j| + \lambda_Y \sum_{j=1}^{p} |\beta_j|, \tag{2.8}$$

where λ_Q and λ_Y are tuning parameters. The misrepresented risk factor v is considered a significant predictor, and therefore, no penalty is applied to β_v to retain it in the model. For conciseness, the corresponding estimation procedure is provided in Appendix A.

3. Deep neural network with sparse inputs and misrepresentation

In severity modeling, the relationship between input features and claim amounts is often inherently nonlinear. This nonlinearity arises from several sources: complex interactions among risk factors (e.g., age and health status), and non-monotonic effects (e.g., driving experience versus accident severity). Linear or shallow models may struggle to capture such complexities, particularly when policyholders misrepresent their risk profiles. A more subtle yet equally critical limitation of linear frameworks is their vulnerability to functional form misspecification. Even in scenarios where all risk factors are reported truthfully, imposing an incorrect linear structure can lead to biased coefficient estimates and distorted inferences. This misspecification bias occurs because the linear model constrains the relationship between covariates and claim severity to a rigid additive form, which may not accurately reflect the true underlying data-generating process. When such bias propagates through the model, it can be mistakenly attributed to the effects of misrepresentation, potentially leading to erroneous conclusions about the prevalence of reporting behavior and the necessary pricing adjustments. Deep learning models, with their ability to learn hierarchical nonlinear representations, present a promising solution to address both of these challenges. By flexibly approximating the underlying nonlinear structure without imposing stringent parametric assumptions, they can better isolate the effects of misrepresentation from the complexities inherent in the data. This flexibility not only reduces the risk of misspecification bias

https://doi.org/10.1017/asb.2026.10115 Published online by Cambridge University Press

but also facilitates more accurate feature selection and risk prediction, particularly in contexts where the true functional form remains unknown.

3.1. Overview of DNN

We briefly review the DNN architecture. Let L be the number of hidden layers, and n_l, $l = 0, \ldots, L+1$, be the number of neurons in each layer. Here, n_0 and n_{L+1} correspond to the dimensions of the input and output layers, respectively. Let $\sigma(\cdot)$ denote the activation function in the hidden layers, and $\psi(\cdot)$ denote the activation function in the output layer. We define $\boldsymbol{W}^{(l)} \in \mathbb{R}^{n_l \times n_{l-1}}$ and $\boldsymbol{b}^{(l)} \in \mathbb{R}^{n_l}$, $l = 1, \ldots, L+1$ as the weight matrix and bias term, respectively. The total parameter set of the neural network is denoted by $\boldsymbol{\Theta} = \{\boldsymbol{W}^{(l)}, \boldsymbol{b}^{(l)} : l = 1, \ldots, L+1\}$. For clarity, we treat $\boldsymbol{\Theta}$ as a column vector in $\mathbb{R}^m$, where $m = \sum_{l=1}^{L+1} n_l\,(n_{l-1}+1)$.

To delve into the mathematical description of the DNN, we can break down its structure starting from the input layer and progressing towards the output layer:

$$\boldsymbol{a}^{(1)} = \boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}, \quad \boldsymbol{z}^{(1)} = \sigma\left(\boldsymbol{a}^{(1)}\right), \tag{3.1}$$

$$\boldsymbol{a}^{(l)} = \boldsymbol{W}^{(l)}\boldsymbol{z}^{(l-1)} + \boldsymbol{b}^{(l)}, \quad \boldsymbol{z}^{(l)} = \sigma\left(\boldsymbol{a}^{(l)}\right), \quad l = 2, \ldots, L, \tag{3.2}$$

$$\boldsymbol{a}^{(L+1)} = \boldsymbol{W}^{(L+1)}\boldsymbol{z}^{(L)} + \boldsymbol{b}^{(L+1)}, \quad \boldsymbol{z}^{(L+1)} = \psi\left(\boldsymbol{a}^{(L+1)}\right). \tag{3.3}$$

The choice of the output activation function $\psi(\cdot)$ should be aligned with the specific problem being tackled. For the binary classification problem, a sigmoid function $\psi(a) = 1/(1+\exp(-a))$ can be implemented; for the regression problem, an exponential or identity function can be applied.

3.2. DNN with misrepresentation

To effectively address the challenges associated with misrepresentation, we propose a modeling approach utilizing DNN. We define two neural networks, NN_Y and NN_Q, to model the log-transformed claim severity y_i based on $(\boldsymbol{x}_i, v_i)$, and the probability of misrepresentation q_i for individuals with $v_i^* = 0$ based on the features $\boldsymbol{x}_i$, respectively. The final output of the overall neural network, which integrates NN_Y and NN_Q, is as follows:

$$\mu(\boldsymbol{x}_i, v_i^*; \boldsymbol{\Theta}) = \begin{cases} \mu_Y(\boldsymbol{x}_i, v_i = 1; \boldsymbol{\Theta}_Y), & v_i^* = 1, \\ \mu_Q(\boldsymbol{x}_i; \boldsymbol{\Theta}_Q)\mu_Y(\boldsymbol{x}_i, v_i = 1; \boldsymbol{\Theta}_Y) + (1 - \mu_Q(\boldsymbol{x}_i; \boldsymbol{\Theta}_Q))\mu_Y(\boldsymbol{x}_i, v_i = 0; \boldsymbol{\Theta}_Y), & v_i^* = 0, \end{cases} \tag{3.4}$$

where $\boldsymbol{\Theta} = \{\boldsymbol{\Theta}_Q, \boldsymbol{\Theta}_Y\}$, μ_Y is the final output from NN_Y, and μ_Q is the final output from NN_Q.

The proposed loss function is defined as

$$\mathcal{L}(\boldsymbol{\Theta}) = \frac{1}{N}\sum_{i=1}^{N} \ell_0(y_i, \mu(\boldsymbol{x}_i, v_i^*; \boldsymbol{\Theta})) + \omega\,\mathrm{KL}\,(\bar{\mu}_Q || \varepsilon), \tag{3.5}$$

which consists of two key components: a standard prediction loss and a regularization term that penalizes deviations from a baseline assumption of honesty. The first term measures the prediction accuracy using squared error loss, defined as $\ell_0(y, \mu) = (y-\mu)^2$. The second term imposes a KL divergence constraint defined as:

$$\mathrm{KL}\,(\bar{\mu}_Q || \varepsilon) = \bar{\mu}_Q \log\left(\frac{\bar{\mu}_Q}{\varepsilon}\right) + (1-\bar{\mu}_Q)\log\left(\frac{1-\bar{\mu}_Q}{1-\varepsilon}\right), \tag{3.6}$$

where $\bar{\mu}_Q$ denotes the average predicted misrepresentation probability for the subgroup with $v_i^* = 0$, and $\varepsilon > 0$ is a small fixed constant that anchors the KL divergence to a baseline of near-zero misrepresentation probability. The regularization strength is governed by the tuning parameter ω, which is selected via

https://doi.org/10.1017/asb.2026.10115 Published online by Cambridge University Press

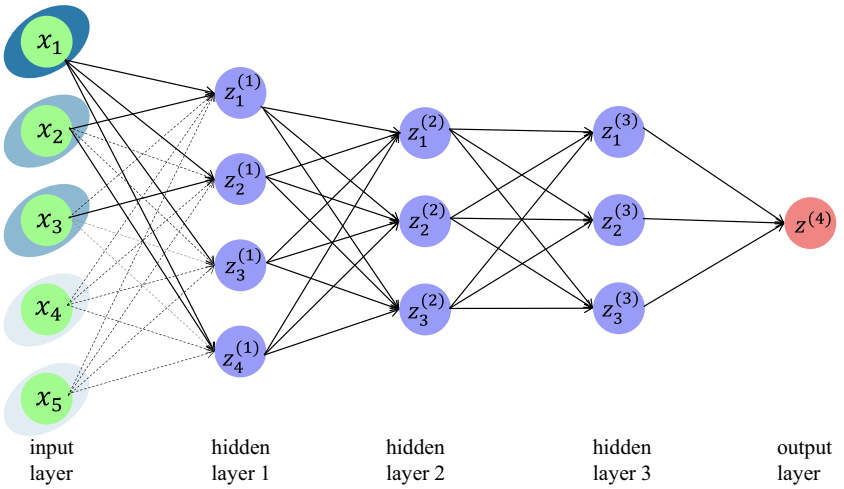


Figure 1. *An example of a sparse-input neural network, specifically a four-layer deep neural network with architecture* $\boldsymbol{m} = (5, 4, 3, 3, 1)$. *The heavy and dotted lines indicate nonzero and zero weights, respectively. Each shaded oval corresponds to a group of first-layer weights. The weights from the dark blue oval are both nonzero. Each medium blue oval contains a subset of nonzero weights, thereby exhibiting element-wise sparsity. All the weights in the light blue ovals are zero, so they exhibit group-level sparsity.*

validation. By setting ε to a small constant (e.g., 10^{-6}), this KL divergence regularization induces sparsity in the latent misrepresentation variables. When data provide limited evidence of misrepresentation, this term penalizes deviations from the baseline and drives NN_Q outputs toward zero, implementing a default honesty assumption. When empirical evidence is sufficiently strong, the data-fitting term overrides this anchor, allowing the model to detect genuine misrepresentation.

3.3. Sparse-input deep neural network

To perform variable selection in a DNN, we propose fitting a sparse-input deep neural network (SIDNN) by applying an L_1-gMCP to the weights connecting the input layer to the first hidden layer. The inclusion of this type of penalty term facilitates variable selection in neural networks. Let $w_{jk}^{(1)}$ represent the element in the j-th row and k-th column of $\boldsymbol{W}^{(1)}$. The parameters of the neural network can then be estimated by solving the following optimization problem:

$$\arg\min_{\boldsymbol{\Theta}} \left\{ \mathcal{L}_0(\boldsymbol{\Theta}) + P\left(\boldsymbol{W}^{(1)}; \lambda, \gamma\right) \right\}, \tag{3.7}$$

where $\mathcal{L}_0(\boldsymbol{\Theta})$ denotes a general loss function, and

$$P\left(\boldsymbol{W}^{(1)}; \lambda, \gamma\right) = \sum_{k=1}^{n_0} \rho\left(\sum\nolimits_{j=1}^{n_1} \left| w_{jk}^{(1)} \right|; \lambda, \gamma \right), \tag{3.8}$$

with $\rho(t; \lambda, \gamma) = \lambda \int_0^{|t|} (1 - \frac{x}{\gamma\lambda})_+ \mathrm{d}x$ as the MCP function.

The L_1-gMCP combines the group MCP with the LASSO penalty. The outer MCP promotes sparsity at the input layer by constraining the entire vector $w_{\cdot k}^{(1)} = (w_{1k}^{(1)}, \ldots, w_{n_1 k}^{(1)})^\top$ to be zero, thus facilitating variable selection. The inner LASSO penalty further promotes sparsity by forcing some of the individual weights $w_{jk}^{(1)}$ to be zero. The parameter λ governs the strength of the penalty, thus influencing the sparsity of the model. Meanwhile, the parameter γ regulates the concavity of the penalty function, helping to balance bias and variance. Figure 1 illustrates how the penalty term in (3.7) operates.

https://doi.org/10.1017/asb.2026.10115 Published online by Cambridge University Press

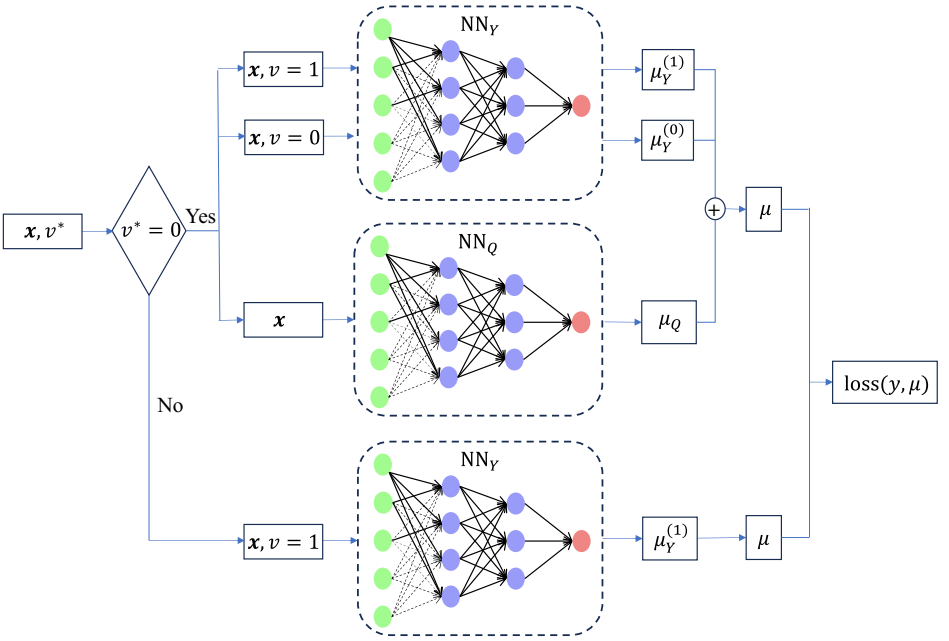


Figure 2. *SIDNN-based claims model with misrepresentation adjustment.*

3.4. SIDNN with misrepresentation

To achieve variable selection within the neural network framework addressing the misrepresentation problem, we introduce L_1-gMCP regularization terms in both NN_Y and NN_Q. The optimization problem can be formulated as follows:

$$\arg\min_{\boldsymbol{\Theta}} \left\{ \mathcal{L}(\boldsymbol{\Theta}) + \sum_{\mathcal{A}\in\{Y,Q\}} P\left(\boldsymbol{W}_{\mathcal{A}}^{(1)}; \lambda_{\mathcal{A}}, \gamma_{\mathcal{A}}\right) \right\}, \tag{3.9}$$

where $\mathcal{L}(\boldsymbol{\Theta})$ is defined in (3.5), and the penalty term is the same as in (3.8). The symbol $\mathcal{A}$ denotes the corresponding neural network. Note that v is considered a significant factor in our model, and thus no penalty is applied to the weights associated with v. Specifically, $\boldsymbol{W}_Y^{(1)}$ excludes the weights corresponding to edges connected to v. The overall structure of our neural network is depicted in Figure 2.

3.5. Learning

The objective function in (3.9) includes a non-smooth regularization term, which makes it unsuitable for direct optimization via gradient descent. To address this, we utilize the PGD method to obtain a sparse solution for the parameters. By initially applying the LLA technique, we transform the original L_1-gMCP problem into an adaptive LASSO problem, and then compute its proximal solution (Zou and Li, 2008).

3.5.1. LLA of L_1-gMCP

Given the value of $w_{\mathcal{A},jk}^{(1)}$ at iteration t, the LLA of the penalty term $P(\boldsymbol{W}_{\mathcal{A}}^{(1)}; \lambda_{\mathcal{A}}, \gamma_{\mathcal{A}})$ at $|w_{\mathcal{A},jk}^{(1)}| = |w_{\mathcal{A},jk}^{(1),t}|$ is:

$$P\left(\boldsymbol{W}_{\mathcal{A}}^{(1)}; \lambda_{\mathcal{A}}, \gamma_{\mathcal{A}}\right) \approx \sum_{k=1}^{n_0^{\mathcal{A}}} \sum_{j=1}^{n_1^{\mathcal{A}}} \delta_{\mathcal{A},k}^t \left| w_{\mathcal{A},jk}^{(1)} \right| + C_0, \tag{3.10}$$

https://doi.org/10.1017/asb.2026.10115 Published online by Cambridge University Press

where C_0 is a constant that does not depend on $|w^{(1)}_{\mathcal{A},jk}|$, and $\delta^t_{\mathcal{A},k}$ is the first-order partial derivative of $P(\boldsymbol{W}^{(1)}_{\mathcal{A}}; \lambda_{\mathcal{A}}, \gamma_{\mathcal{A}})$ evaluated at $|w^{(1),t}_{\mathcal{A},jk}|$, given by:

$$\delta^t_{\mathcal{A},k} = \left(\lambda_{\mathcal{A}} - \frac{\sum_{j=1}^{n_1^{\mathcal{A}}} \left| w^{(1),t}_{\mathcal{A},jk} \right|}{\gamma_{\mathcal{A}}} \right)_+, \tag{3.11}$$

where $(x)_+ = \max(x, 0)$. Note that $n_0^Y = n_0^Q = p$ represents the dimension of the input features $\boldsymbol{x}$, while n_1^Y and n_1^Q denote the number of neurons in the first hidden layers of NN_Y and NN_Q, respectively.

The parameter optimization problem in (3.9) is then transformed into the following adaptive LASSO form:

$$\arg\min_{\boldsymbol{\Theta}} \left\{ \mathcal{L}(\boldsymbol{\Theta}) + \sum_{\mathcal{A}\in\{Y,Q\}} \sum_{k=1}^{n_0^{\mathcal{A}}} \sum_{j=1}^{n_1^{\mathcal{A}}} \delta^t_{\mathcal{A},k} \left| w^{(1)}_{\mathcal{A},jk} \right| \right\}, \tag{3.12}$$

where $\mathcal{L}(\boldsymbol{\Theta})$ is given in (3.5).

3.5.2. PGD for LASSO penalty

We use the PGD method (Parikh and Boyd, 2014; Alvarez and Salzmann, 2016) to address the problem in (3.12). First, perform a quadratic approximation of $\mathcal{L}(\boldsymbol{\Theta})$ at $\boldsymbol{\Theta} = \boldsymbol{\Theta}^t$:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\Theta}) &\approx \mathcal{L}(\boldsymbol{\Theta}^t) + \nabla\mathcal{L}(\boldsymbol{\Theta}^t)^\top(\boldsymbol{\Theta} - \boldsymbol{\Theta}^t) + \frac{1}{2\eta} \left\| \boldsymbol{\Theta} - \boldsymbol{\Theta}^t \right\|_2^2 \\ &\approx \frac{1}{2\eta} \left\| \boldsymbol{\Theta} - (\boldsymbol{\Theta}^t - \eta\nabla\mathcal{L}(\boldsymbol{\Theta}^t)) \right\|_2^2 + C_1 \end{aligned} \tag{3.13}$$

where η is the step size for gradient descent, and C_1 is a constant that does not depend on $\boldsymbol{\Theta}$. Equation (3.13) transforms (3.12) into an optimization problem for a quadratic smooth function with L_1 regularization:

$$\arg\min_{\boldsymbol{\Theta}} \left\{ \frac{1}{2\eta} \left\| \boldsymbol{\Theta} - (\boldsymbol{\Theta}^t - \eta\nabla\mathcal{L}(\boldsymbol{\Theta}^t)) \right\|_2^2 + \sum_{\mathcal{A}\in\{Q,Y\}} \sum_{k=1}^{n_0^{\mathcal{A}}} \sum_{j=1}^{n_1^{\mathcal{A}}} \delta^t_{\mathcal{A},k} \left| w^{(1)}_{\mathcal{A},jk} \right| \right\}. \tag{3.14}$$

Then, the update of the parameter in $\mathcal{W} = \left\{ \boldsymbol{W}^{(1)}_Y, \boldsymbol{W}^{(1)}_Q \right\}$ is given by:

$$w^{(1),t+1}_{\mathcal{A},jk} = \mathcal{S}_{\delta^t_{\mathcal{A},k}\eta} \left(w^{(1),t}_{\mathcal{A},jk} - \eta \left. \frac{\partial\mathcal{L}(\boldsymbol{\Theta})}{\partial w^{(1)}_{\mathcal{A},jk}} \right|_{w^{(1)}_{\mathcal{A},jk} = w^{(1),t}_{\mathcal{A},jk}} \right), \tag{3.15}$$

where $\mathcal{S}$ denotes the soft-thresholding operator defined as $\mathcal{S}_\lambda(w) = \mathrm{sign}(w)(|w| - \lambda)_+$. For the parameter $\theta \in \boldsymbol{\Theta} \setminus \mathcal{W}$, the update is as follows:

$$\theta^{t+1} = \theta^t - \eta \left. \frac{\partial\mathcal{L}(\boldsymbol{\Theta})}{\partial\theta} \right|_{\theta=\theta^t}. \tag{3.16}$$

Algorithm 1 provides a summary of the pseudocode for the parameter estimation in our proposed model.

3.6. Summary

To help guide the reader, we briefly summarize the key stages of our methodological framework and the rationale for the penalty structures introduced. In Section 2, we presented a linear model that accounts for unidirectional misrepresentation. To handle high-dimensional covariates, we imposed ℓ_1 penalties on the coefficients in both the mean and prevalence components, enabling variable selection

https://doi.org/10.1017/asb.2026.10115 Published online by Cambridge University Press

Algorithm 1: Optimization algorithm for our proposed SIDNN with misrepresentation.

Input: Tuning parameters $\lambda_Q, \lambda_Y, \gamma_Q, \gamma_Y, \omega$, initialized model parameters $\boldsymbol{\Theta}^0$.
Output: Optimal parameters $\boldsymbol{\Theta}^*$.

while $t < T_{\max}$ **and** not converged **do**
  Calculate the loss $\mathcal{L}\left(\boldsymbol{\Theta}^t\right)$;
  Obtain the gradient $\nabla\mathcal{L}(\boldsymbol{\Theta}^t)$ through back-propagation;
  Compute $\delta^t_{\mathcal{A},k}$: $\delta^t_{\mathcal{A},k} = \left(\lambda_{\mathcal{A}} - \frac{\sum_{j=1}^{n_1^{\mathcal{A}}} \left|w^{(1),t}_{\mathcal{A},jk}\right|}{\gamma_{\mathcal{A}}}\right)_+$;
  Update the parameter in $\mathcal{W}$ by: $w^{(1),t+1}_{\mathcal{A},jk} = \mathcal{S}_{\delta^t_{\mathcal{A},k}\eta}\left(w^{(1),t}_{\mathcal{A},jk} - \eta \frac{\partial\mathcal{L}(\boldsymbol{\Theta})}{\partial w^{(1)}_{\mathcal{A},jk}}\Big|_{w^{(1)}_{\mathcal{A},jk}=w^{(1),t}_{\mathcal{A},jk}}\right)$;
  Update the parameter $\theta \in \boldsymbol{\Theta} \setminus \mathcal{W}$ by: $\theta^{t+1} = \theta^t - \eta \frac{\partial\mathcal{L}(\boldsymbol{\Theta})}{\partial\theta}\Big|_{\theta=\theta^t}$.
end
return $\boldsymbol{\Theta}^*$

and mitigating overfitting. Recognizing the limitations of the linear framework, particularly its vulnerability to misspecification bias when the true data-generating process is nonlinear, we introduced a deep learning approach in this section. Within this neural network architecture, standard L_1 penalties are insufficient for feature selection because each input connects to multiple hidden units. We therefore adopted group-based penalties that treat all weights from a single input as a group, allowing for feature-level sparsity. Building on this, we proposed a novel L_1-gMCP penalty that combines group MCP with LASSO regularization, achieving feature selection with reduced estimation bias. A proximal gradient algorithm is developed for efficient optimization.

4. Simulation studies

Before applying our model to real-world datasets, we conduct a series of simulation experiments to assess its performance. Two simulation scenarios are considered: Simulation 1 examines the setting where misrepresentation is present in the data, while Simulation 2 evaluates model performance when the reported status is truthful (i.e., no misrepresentation). By replicating these realistic data characteristics in controlled settings where the ground truth is known, we can validate the model's effectiveness in identifying misrepresentation. The insights gained from these simulations will provide a solid foundation for interpreting the results of our subsequent analysis of real-world data, where the underlying data-generating process is unknown.

In our studies, we aim to compare the performance of the following models, each of which is described below.

- M1: Linear regression with L_1 regularization.
- M2: Linear regression with misrepresentation adjustment and L_1 regularization, see (2.8).
- M3: Deep neural network with L_1-gMCP regularization.
- M4: Deep neural network incorporating misrepresentation and L_1-gMCP regularization.

To further disentangle the sources of improvement and to evaluate the contribution of our proposed L_1-gMCP structured sparsity penalty, we introduce an additional benchmark model, denoted as M5. M5 is a deep neural network that incorporates the misrepresentation adjustment and KL divergence, but replaces our proposed L_1-gMCP penalty with a standard element-wise L_1 penalty on $\boldsymbol{W}^{(1)}$. Unlike M4, M5 does not enforce group-level sparsity at the input layer; consequently, it cannot fully eliminate irrelevant features, as each input may retain residual connections to multiple hidden neurons.

https://doi.org/10.1017/asb.2026.10115 Published online by Cambridge University Press

The implementations of M1 and M2 are carried out using R, while M3, M4 and M5 are implemented with PyTorch. The setup for running M4 is as follows: The experiment is conducted on an Ubuntu 20.04 system, equipped with an Intel i9-13900K CPU (24 cores) and an NVIDIA RTX 3060 Ti GPU (8 GB), running the PyTorch 2.2.1+cu12 framework. A single simulation run for a given tuning-parameter configuration takes approximately 50 seconds to complete.

4.1. Setting

The data-generating process follows a common configuration across both simulations. The sample size is $N = 10,000$ with covariate dimension $p = 100$. The covariate vector $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ is generated from a Gaussian copula, where the element of the covariance matrix $\boldsymbol{\Sigma}$ is defined as $\Sigma_{jk} = 0.5^{|j-k|}$ for $1 \le j, k \le p$. The reported status v_i^* is generated from Bernoulli (0.2). The response variable y_i is generated from a normal distribution with mean $\mu_i = -3 + 1.5 \exp(x_{i1}) - 1.5(2x_{i2} - 1)^2 + 2 \sin(2\pi x_{i3} x_{i4}) - 2 \cos(2\pi x_{i5} x_{i6}) + 3v_i$ and standard deviation $\sigma = 1$. The key difference between the two simulations lies in the generation of the true status v_i:

- **Simulation 1 (with misrepresentation)**: For individuals with reported status $v_i^* = 1$, the true status $v_i = 1$. For those with $v_i^* = 0$, v_i is generated from Bernoulli(q_i), where $\text{logit}(q_i) = -3 + \exp(x_{i3}) - 2(2x_{i4} - 1)^2 + 3 \sin(2\pi x_{i5} x_{i6}) + 3 \cos(2\pi x_{i7} x_{i8})$. Under this setting, the true misrepresentation rate among the subgroup with $v_i^* = 0$ is 0.472.
- **Simulation 2 (without misrepresentation)**: We enforce $v_i = v_i^*$, ensuring no discrepancy between reported and true statuses. The true misrepresentation rate is zero.

We compare five models (M1–M5) across 100 replications for each simulation. In each replication, the data are randomly split into training, validation, and test sets in a 6:2:2 ratio. All models are trained on the training set, with tuning parameters selected using the validation set, and predictive performance is assessed on the test set. The neural network architectures in M3, M4, and M5 consist of four hidden layers with 128, 64, 32, and 32 neurons, respectively. This configuration is selected based on pilot experiments demonstrating stable performance with reasonable computational cost. A dropout rate of 0.1 is applied to enhance robustness and generalization. The Swish activation function, $f(x) = x/(1 + e^{-x})$, is used for its ability to capture complex nonlinear patterns while mitigating vanishing-gradient issues. The anchor parameter ε in M4 is set to 10^{-6}.

Evaluation metrics include mean squared error (MSE), mean absolute error (MAE), R^2, area under the curve (AUC), and $F1$ score. MSE, MAE, and R^2 assess prediction accuracy for y; AUC measures the effectiveness in identifying misreported individuals (only applicable in Simulation 1); and $F1$ score evaluates variable selection performance. Additionally, we introduce the predicted prevalence of misrepresentation ($\bar{q}$), defined as the average predicted probability of misreporting among individuals with $v_i^* = 0$ in the test set. To save space in the main text, the calculation methods for the corresponding metrics are provided in Appendix B.

4.2. Results

Table 1 reports the mean and standard deviation (in parentheses) of each metric across 100 replications under the misrepresentation scenario. Several key observations can be made from the results. First, neural network-based models (M3–M5) significantly outperform their linear counterparts (M1, M2) across all evaluation dimensions. This aligns with the well-established capacity of neural networks to approximate complex functional forms, as the simulated data contain exponential, quadratic, sine, and cosine transformations. Second, models incorporating misrepresentation correction (M2, M4) consistently outperform their unadjusted analogs (M1, M3) in two critical aspects: prediction accuracy for y and variable selection effectiveness for y. These results underscore the necessity of accounting for misrepresentation when such characteristics are present. Third, despite the conservative regularization anchor ($\varepsilon = 10^{-6}$), the predicted $\bar{q} \approx 0.416$ from M4 suggests a misrepresentation prevalence that aligns

https://doi.org/10.1017/asb.2026.10115 Published online by Cambridge University Press

Table 1. *Comparison results of several candidate models in Simulation 1.*

Model	MSE	MAE	R^2	AUC	$F1_q$	$F1_y$	$\bar{q}$
M1	5.576	1.903	0.212	–	–	0.314	–
	(0.172)	(0.033)	(0.014)	–	–	(0.106)	–
M2	5.524	1.891	0.220	0.767	0.383	0.316	0.552
	(0.165)	(0.032)	(0.014)	(0.014)	(0.160)	(0.113)	(0.010)
M3	2.465	1.250	0.651	–	–	0.649	–
	(0.208)	(0.059)	(0.031)	–	–	(0.194)	–
M4	2.388	1.233	0.662	0.785	0.497	0.847	0.416
	(0.231)	(0.062)	(0.033)	(0.029)	(0.081)	(0.074)	(0.107)
M5	2.706	1.311	0.618	0.772	–	–	0.458
	(0.184)	(0.046)	(0.025)	(0.036)	–	–	(0.103)

Table 2. *Comparison results of several candidate models in Simulation 2.*

Model	MSE	MAE	R^2	$F1_y$	$\bar{q}$
M1	3.659	1.529	0.437	0.360	–
	(0.111)	(0.026)	(0.015)	(0.133)	–
M2	3.772	1.550	0.419	0.442	0.310
	(0.115)	(0.026)	(0.014)	(0.184)	(0.013)
M3	1.473	0.965	0.773	0.733	–
	(0.237)	(0.077)	(0.035)	(0.074)	–
M4	1.506	0.976	0.768	0.765	0.001
	(0.322)	(0.091)	(0.047)	(0.103)	(0.009)
M5	1.905	1.091	0.707	–	0.017
	(0.349)	(0.097)	(0.054)	–	(0.073)

fairly closely with the true value of 0.472. This validates the role of the KL term as a stability constraint: when the data signal is sufficiently strong, the regularization does not impede the model's ability to capture the true underlying distribution. Fourth, the predictive accuracy of M5 for the response variable y is generally inferior to that of both M3 and M4. This shortcoming can be attributed to the element-wise L_1 penalty, which fails to completely eliminate irrelevant features. As a result, residual noisy connections remain, ultimately degrading prediction performance. Therefore, it is crucial to combine misrepresentation adjustment with the proposed group-level sparsity penalty in order to fully realize the advantages of our modeling framework.

Table 2 presents the comparative results under the no misrepresentation scenario. We focus on three performance criteria: prediction accuracy for y, variable selection effectiveness for y, and the influence of KL divergence regularization in M4. First, neural network-based models continue to demonstrate substantially better predictive performance than their linear counterparts, as anticipated given the nonlinear data-generating process. Second, M3 can be considered the most appropriate baseline in this scenario. Our proposed M4 achieves nearly identical performance to M3 in terms of both prediction accuracy and variable selection, indicating that the misrepresentation adjustment mechanism does not introduce spurious biases when no misrepresentation exists. Third, M4 demonstrates remarkable precision in estimating the absence of misrepresentation, with predicted $\bar{q} \approx 0.001$ that closely aligns with the ground truth of zero. This validates the role of the KL divergence term in correctly penalizing spurious misrepresentation signals. In contrast, M2 again overestimates q, reflecting its limitations in discerning true reporting behavior when the relationship between response and explanatory variables is complex. Fourth, M5 again underperforms M4 in terms of predictive accuracy for y. This further emphasizes the critical role of the structured L_1-gMCP penalty in enhancing model effectiveness in this context.

https://doi.org/10.1017/asb.2026.10115 Published online by Cambridge University Press

Table 3. *Summary of selected explanatory variables.*

Variable	Description	Type
UNINS14(v^*)	Reported status: 1 if uninsured, 0 if insured	Categorical
Demographic:		
SEX	Sex: Male, Female	Categorical
AGE14X	Age: 18-65	Numerical
RACEV1X	Race: White, Black, Native, Asian, Other	Categorical
MARRY14X	Marital Status: Married, Widowed, Divorced, Separated, Never married	Categorical
FAMSZEYR	Size of responding annualized family: 1–14	Numerical
Geographic:		
REGION14	Census Region: Northeast, Midwest, South, West	Categorical
Income:		
WAGEP14X	Person's wage income: $0–$395,558	Numerical
POVCAT14	Family income as a percentage of poverty: Poor, Near Poor, Low, Middle, High	Categorical
Health Status:		
RTHLTH53	Perceived health status, round 5/3: Excellent, Very good, Good, Fair, Poor, Inapplicable	Categorical
MNHLTH53	Perceived mental health status, round 5/3: Excellent, Very good, Good, Fair, Poor, Inapplicable	Categorical
ADSMOK42	Currently smoke: Yes, No, Inapplicable, Not ascertained	Categorical
Utilization:		
OBTOTV14	Total office-based visits: 0–391	Numerical
OPTOTV14	Total outpatient visits: 0–153	Numerical
ERTOT14	Total emergency room visits: 0–18	Numerical
IPDIS14	Total inpatient stays (including zero-night stays): 0–10	Numerical
RXTOT14	Total prescription medicines: 0–344	Numerical
DVTOT14	Total dental visits: 0–27	Numerical

5. Application

5.1. Data description

In this study, we use the 2014 consolidated data from the Medical Expenditure Panel Survey (MEPS), as previously examined by Akakpo *et al*. (2019) to investigate the presence of misrepresentation in insurance status. The dataset is available from the MEPS website https://meps.ahrq.gov/. A potential incentive for reporting inaccurate insurance status arises from the 2014 Affordable Care Act reforms imposing a tax penalty for lacking health insurance. Social desirability and financial concerns may drive MEPS respondents to misrepresent their insurance status.

The raw dataset contains information on 34,875 individuals. For our analysis, we retain individuals aged 18–65 years with positive health expenditures. Additionally, we exclude individuals whose health status and mental health status fall into rare categories. After processing, the dataset includes 15,865 records. Selected predictors are displayed in Table 3. Some categorical variables have multiple levels (e.g., race, marital status, and health status), which expand the dimensionality of the explanatory variables to 39 dimensions. The response variable of interest is the log-transformed total health care expenditure. Density plots of the response variable, categorized by reported insurance status,

https://doi.org/10.1017/asb.2026.10115 Published online by Cambridge University Press

Table 4. *Data partitioning.*

Set	Symbol	Number of individuals	Proportion (%)
Training	$\mathcal{T}_r$	9,519	60
Validation	$\mathcal{V}_a$	3,173	20
Test	$\mathcal{T}_e$	3,173	20

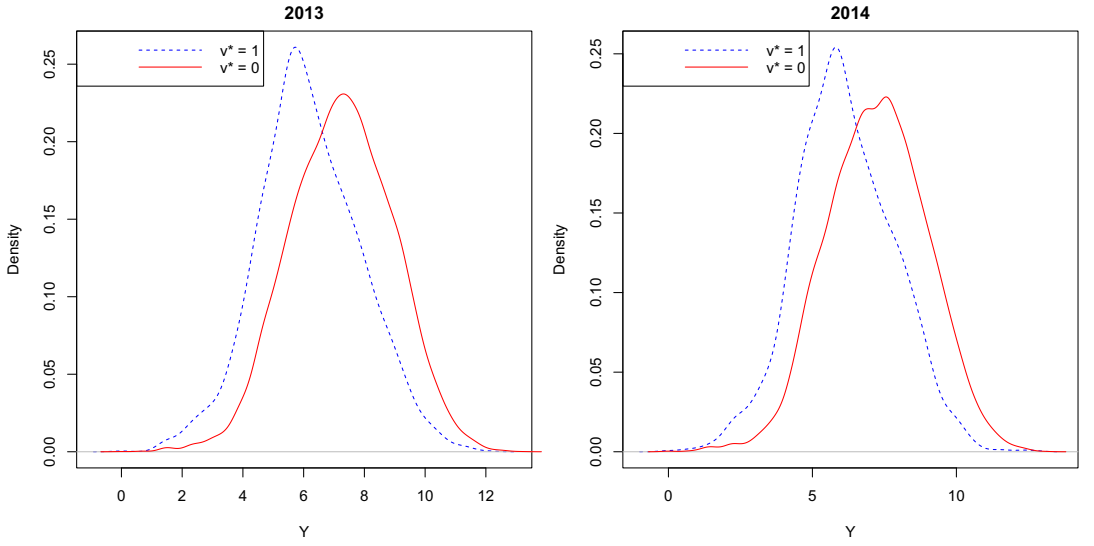


Figure 3. *Density plots for log-transformed positive total medical expenditure categorized by reported insurance status.*

are presented in Figure 3. We denote $v_i^* = 1$ if the respondent reports being uninsured, and $v_i^* = 0$ otherwise. A distinct bimodal distribution is evident for the reported insured group in 2014, whereas the corresponding distribution in 2013 lacks this clarity. This temporal difference provides preliminary evidence consistent with systematic misrepresentation of insurance status specifically in 2014.

In our analysis, we randomly split the data into a training, validation, and test set as seen in Table 4. The training set is used to develop the model, the validation set is used to select tuning parameters, and predictive performance is assessed on the test set. In addition to our proposed DNN model with misrepresentation and the L_1-gMCP penalty (M4), the other four candidate models (M1, M2, M3, M5) are identical to those used in the simulation studies. All numerical explanatory variables are scaled before they are fed into the model.

5.2. Model fitting

We provide a brief overview of the architectures used for the three deep learning models. To ensure comparability, M3, M4, and M5 adopt the same underlying network architecture. Specifically, each network is a feedforward neural network with four hidden layers, featuring 96, 64, 32, and 32 neurons, respectively. All hidden layers use the Swish activation function (as previously defined). To prevent overfitting, a dropout rate of 0.1 is applied to the output of each hidden layer.

The hyperparameter tuning for these three deep learning models follows a rigorous protocol. For the MCP penalty's concavity parameter γ, we explore values within the set $\{3, 6, 9, 12\}$. Other critical regularization hyperparameters (i.e., λ, ω) are determined through an exhaustive grid search, with optimal values selected from predefined candidate grids based on minimizing the validation MSE. The learning

https://doi.org/10.1017/asb.2026.10115 Published online by Cambridge University Press

Table 5. *Variable selection under four models.*

	Q				*Y*			
Variable	M1	M2	M3	M4	M1	M2	M3	M4
SEX: Female					√	√	√	
AGE14X					√	√		√
RACEV1X: Black					√	√		
RACEV1X: Native				√			√	√
RACEV1X: Asian				√	√	√		√
RACEV1X: Other								√
MARRY14X: Widowed								
MARRY14X: Divorced								
MARRY14X: Separated				√				
MARRY14X: Never married				√	√	√	√	
FAMSZEYR				√	√	√		
REGION14: Midwest				√	√	√		
REGION14: South				√		√		√
REGION14: West				√	√	√		√
WAGEP14X				√			√	
POVCAT14: Near Poor					√	√		
POVCAT14: Low				√				
POVCAT14: Middle				√	√	√	√	
POVCAT14: High				√	√	√	√	
RTHLTH53: Excellent				√	√	√	√	√
RTHLTH53: Very good				√	√	√	√	√
RTHLTH53: Good				√	√	√	√	√
RTHLTH53: Fair				√	√	√	√	√
RTHLTH53: Poor							√	√
MNHLTH53: Excellent				√			√	√
MNHLTH53: Very good				√			√	√
MNHLTH53: Good				√			√	√
MNHLTH53: Fair					√		√	√
MNHLTH53: Poor				√	√		√	√
ADSMOK42: Inapplicable				√			√	√
ADSMOK42: Yes				√			√	√
ADSMOK42: No				√	√	√	√	√
OBTOTV14		√			√	√	√	√
OPTOTV14		√			√	√	√	√
ERTOT14		√			√	√	√	√
IPDIS14		√			√	√	√	√
RXTOT14		√		√	√	√	√	√
DVTOT14		√		√	√	√	√	√

rate η is fixed at 0.001 with a batch size of 512, and ε is set to 10^{-6}. To mitigate overfitting, an early stopping mechanism is employed. Models are trained for a sufficient number of epochs, with final model selection based on the epoch that achieves the minimum MSE on the validation set.

Since M5 does not perform feature-level variable selection, we only present the variable selection results for models M1 through M4 in Table 5. Notably, the four models consistently identify healthcare service utilization as selected predictors of expenditure (y). This finding aligns well with established health economic theory, where increased healthcare utilization naturally leads to higher medical expenditures. In terms of predicting the prevalence of misrepresentation (q), we observe that M4

https://doi.org/10.1017/asb.2026.10115 Published online by Cambridge University Press

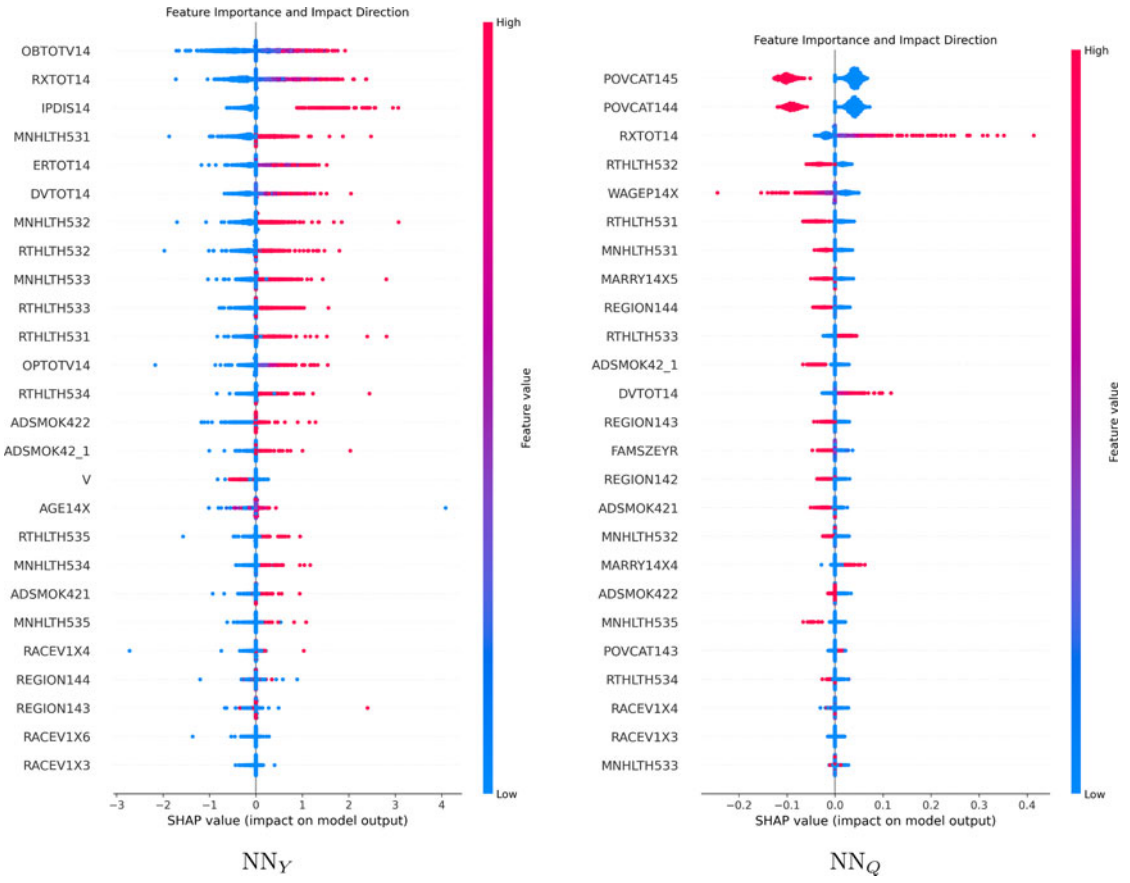


Figure 4. *SHAP summary plots of the most important variables considered in NN_Y and NN_Q under M4.*

incorporates a more comprehensive set of explanatory variables compared to M2. Importantly, both models share two common active predictors for q: RXTOT14 and DVTOT14.

To gain deeper insights into the decision-making mechanisms and interpret variable importance in M4, this study employs SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017). This cutting-edge method in explainable AI has gained significant recognition in recent years. The corresponding plots are displayed in Figure 4. In the summary plot, the vertical axis represents features ranked by their importance, while the horizontal axis quantifies each feature's contribution to model outputs through SHAP values. Each point corresponds to an observation, with the color intensity indicating feature values (red for high, blue for low). Positive SHAP values indicate that the feature increases the predicted outcome, while negative values suggest a decreasing effect. Our core model, M4, addresses misrepresentation through two specialized sub-networks (NN_Y and NN_Q). The SHAP analysis reveals distinct patterns learned by each sub-network, providing valuable interpretability into their respective decision-making processes. From the left panel, we can observe that healthcare utilization variables (e.g., OBTOTV14, RXTOT14, IPDIS14) are the most influential factors in the prediction, all demonstrating positive effects on medical costs. Notably, insurance status also plays a significant role: when the status is uninsured, it substantially lowers the baseline cost prediction. On the right panel, we find that socioeconomic status variables, such as POVCAT145 (high-income households), POVCAT144 (middle-income households), and WAGEP14X (wage income), are key determinants in identifying potential misrepresentation. Higher values of these features correspond to increasingly negative SHAP values, indicating that individuals with higher income levels exhibit lower predicted probabilities of

https://doi.org/10.1017/asb.2026.10115 Published online by Cambridge University Press

Table 6. *Comparison of criteria across five models.*

Model	MSE	MAE	R^2
M1	1.633	0.967	0.477
M2	1.570	0.941	0.497
M3	0.858	0.688	0.725
M4	**0.847**	**0.684**	**0.729**
M5	0.861	0.688	0.724

Table 7. *The Gini indices of health expenditures for five models on the test data.*

	Competing model					
Base model	M1	M2	M3	M4	M5	Minimax
M1	0.000	3.554	6.732	6.805	6.803	6.805
M2	−2.718	0.000	6.437	6.517	6.504	6.517
M3	−0.315	−0.313	0.000	1.086	0.735	1.086
M4	−0.166	−0.164	0.314	0.000	0.446	**0.446**
M5	−0.672	−0.649	0.865	1.163	0.000	1.163

potential misrepresentation. Conversely, healthcare utilization variables such as RXTOT14 (total prescription medicines) show positive associations, systematically increasing the predicted likelihood of misrepresentation.

5.3. Predictive performance

To assess the predictive performance of the five candidate models, we employ multiple diagnostic approaches. First, we evaluate and compare the performance of the five models using multiple metrics, including MSE, MAE, and R^2. The detailed results are presented in Table 6. Consistent with theoretical expectations, the three neural network-based models demonstrate superior predictive accuracy compared to their linear counterparts. This empirical finding underscores the necessity of employing nonlinear, complex modeling frameworks to capture the underlying patterns in this dataset. Notably, the performance gains observed when transitioning from M1 to M2 and from M3 to M4 suggest the presence of misreported cases within the data. This inference aligns with the conclusions drawn in Akakpo *et al.* (2019), further reinforcing the validity of our methodological approach. The comparison of M4 and M5 demonstrates the effectiveness of incorporating the L_1-gMCP compared to the standard L_1 penalty. To ensure the robustness of the comparison and mitigate the impact of data partitioning, we repeated the experiments using five distinct random splits. The aggregated results across all splits are reported in Appendix C. Across all splits, M4 consistently outperforms the other candidate models in terms of MSE, MAE, and R^2, with the same direction and similar magnitude of improvement. Notably, in Split 5, M4's marginal MAE advantage over M3 is obscured by rounding in the tabulated values. Collectively, these results support the robustness of the observed performance gain across data splits.

To evaluate the risk-differentiation ability of models predicting expenditures, we utilize the Gini indices, a well-established metric derived from Lorenz curve analysis. As demonstrated in Frees *et al.* (2011, 2014), Gini indices provide a robust measure of a model's ability to discriminate risk by quantifying the alignment between predicted and actual loss distributions. Our comparative analysis proceeds as follows: for each candidate model (M1–M5), we treat its predicted losses as the base distribution and compute Gini indices against competing predictions from alternative models. The complete matrix of pairwise Gini comparisons is presented in Table 7. To objectively identify the optimal model, we implement a minimax selection strategy: for each model, we determine its maximum Gini index across all possible competing model comparisons; we then select the model exhibiting the smallest maximum

https://doi.org/10.1017/asb.2026.10115 Published online by Cambridge University Press

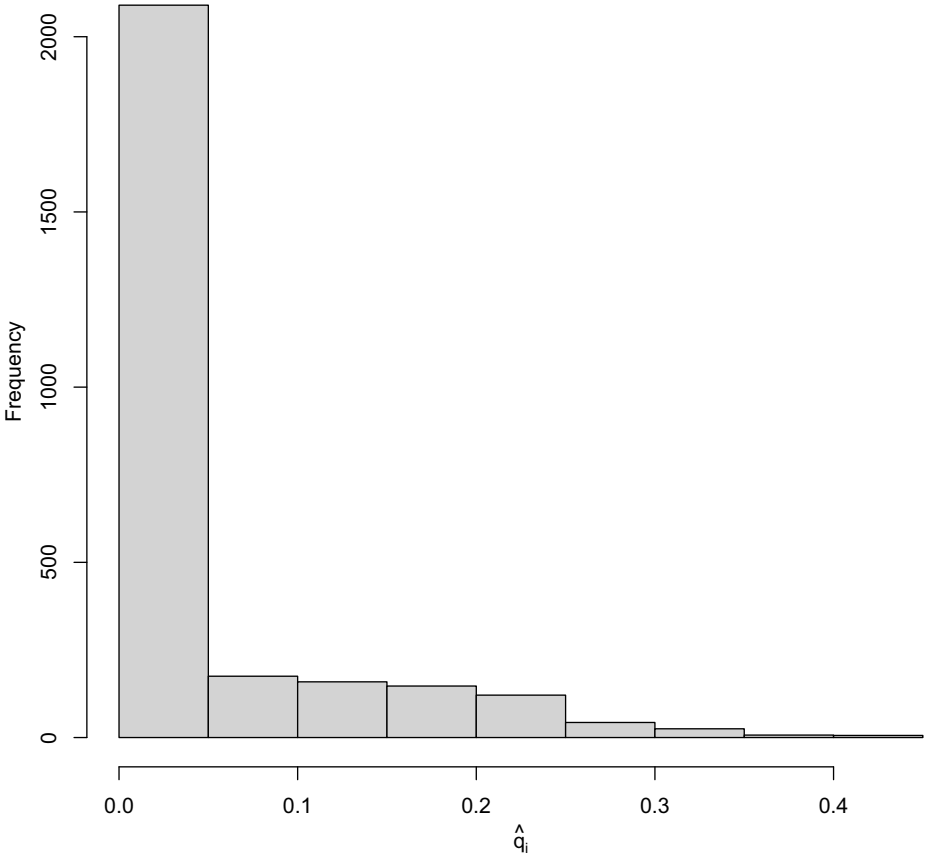


Figure 5. *Histogram of the predicted misrepresentation probabilities $\widehat{q}_i$ from M4 for individuals with reported insurance status $v_i^* = 0$.*

Gini value. This conservative approach identifies the most stable performer – the model least sensitive to variations in ranking basis. As Table 7 demonstrates, M4 emerges as superior with the lowest maximum Gini index (0.446) among all candidates.

Finally, Figure 5 presents a histogram of the predicted misrepresentation probabilities $\widehat{q}_i$ from M4 for individuals with reported insurance status $v_i^* = 0$. The distribution is highly skewed toward zero: the majority of individuals have $\widehat{q}_i$ very close to 0, indicating that the model classifies most respondents as truthful. A small proportion of individuals exhibit elevated $\widehat{q}_i$ values, suggesting that the model identifies a subset of potentially misreporting individuals based on their covariate patterns. The average predicted prevalence of misrepresentation among respondents with $v_i^* = 0$ is $\bar{q} \approx 4.3\%$. It is noteworthy that our estimate of misrepresentation probability is substantially lower than the estimate of 18% reported by Akakpo *et al*. (2019). One possible explanation is the difference in model specification: our method accommodates nonlinear covariate effects, whereas the comparison model uses a more restrictive parametric structure.

5.4. Summary

Although the application focuses on health care expenditures, the proposed framework offers broader implications for the insurance industry. First, compared with conventional linear models, the deep learning framework achieves a substantial improvement in claim severity prediction accuracy, capturing

https://doi.org/10.1017/asb.2026.10115 Published online by Cambridge University Press

complex nonlinearities that traditional approaches cannot accommodate. Second, the predicted misrepresentation probabilities q_i serve primarily as a screening tool to prioritize policies for external verification and manual underwriting review. Only after such validation can these estimates inform adjustments to the insured's risk profile. Consequently, rather than relying solely on potentially misstated self-reported information, this process mitigates adverse selection and supports actuarially fair pricing within a robust governance framework. Third, the KL divergence regularization embeds a "presumption of honesty" into the learning process, flagging misrepresentation only when strongly supported by the data. This design reduces the operational cost and reputational risk associated with investigating wrongly suspected policyholders, while directing audit and anti-fraud resources toward higher-risk cases. Fourth, the sparse structure induced by the L_1-gMCP penalty identifies the most influential drivers of claim severity. This enhances operational efficiency by streamlining data collection, simplifying underwriting workflows, and ensuring that risk assessment focuses on the variables with the greatest predictive value.

6. Concluding remarks

In this article, we addressed the challenge of severity modeling with misrepresentation adjustment in high-dimensional settings. For the linear case, we introduced an L_1 regularization term to facilitate variable selection. To extend beyond linearity, we proposed a neural network-based approach enhanced with a tailored penalty term to maintain interpretability while capturing complex, nonlinear relationships. Through extensive simulation studies conducted under both misrepresentation and no-misrepresentation scenarios, we demonstrated the robustness and adaptability of our neural network model with misrepresentation correction. Finally, we applied our methodology to real-world data from the 2014 MEPS. The empirical analysis further illustrated the practical advantages of our neural network model with misrepresentation adjustment.

While our framework accommodates high-dimensional settings, it currently assumes misreporting is confined to a single binary variable. A natural extension involves addressing multiple concurrent binary fraud indicators. This generalization, however, introduces several technical hurdles. First, the number of possible fraud patterns grows exponentially with the number of binary variables, intensifying the curse of dimensionality. Second, ensuring model identifiability necessitates stringent structural assumptions, without which estimation becomes unstable. Third, accounting for dependence among fraud indicators (e.g., correlated misreporting behaviors) requires more elaborate latent structures, substantially increasing computational complexity. Addressing these challenges remains an important direction for future research.

Code. The code is available from the corresponding author upon reasonable request.

Acknowledgements. The authors express their gratitude to the editor and four reviewers for their valuable comments and suggestions, which have significantly contributed to the improvement of this paper. The research of Pengcheng Zhang was supported by the National Natural Science Foundation of China (No. 12301607), the Shandong Provincial Natural Science Foundation (No. ZR2023QA091), the International Cooperation Platform and Project of Shandong University of Finance and Economics in 2024. The research of Jianwei Gang was supported by the Scientific Research Fund of Zhejiang Provincial Education Department (No. Y202558940). The research of Shilong Li was supported by the National Natural Science Foundation of China (No. 72171133). The research of Xiaoyan Wang was supported by the National Natural Science Foundation of China (No. 72271088).

Competing interests. The authors declare that they have no competing interests.

References

Akakpo R.M., Xia M. and Polansky A.M. (2019) Frequentist inference in insurance ratemaking models adjusting for misrepresentation. *ASTIN Bulletin: The Journal of the IAA* **49**(1), 117–146.

Al-Mudafer M.T., Avanzi B., Taylor G. and Wong B. (2022) Stochastic loss reserving with mixture density neural networks. *Insurance: Mathematics and Economics* **105**, 144–174.

https://doi.org/10.1017/asb.2026.10115 Published online by Cambridge University Press

Alvarez J.M. and Salzmann M. (2016) Learning the number of neurons in deep networks. *In Advances in Neural Information Processing Systems,* **29**.

Delong Ł., Lindholm M. and Wüthrich M.V. (2021) Gamma mixture density networks and their application to modelling insurance claim amounts. *Insurance: Mathematics and Economics* **101**, 240–261.

Fan J. and Li R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**(456), 1348–1360.

Frees E.W., Meyers G. and Cummings A.D. (2011) Summarizing insurance scores using a Gini index. *Journal of the American Statistical Association* **106**(495), 1085–1098.

Frees E.W., Meyers G. and Cummings A.D. (2014) Insurance ratemaking and a Gini index. *Journal of Risk and Insurance* **81**(2), 335–366.

Gao G., Meng S. and Wüthrich M.V. (2019) Claims frequency modeling using telematics car driving data. *Scandinavian Actuarial Journal* **2019**(2), 143–162.

Gomes C., Jin Z. and Yang H. (2021) Insurance fraud detection with unsupervised deep learning. *Journal of Risk and Insurance* **88**(3), 591–624.

Hainaut D. (2018) A neural-network analyzer for mortality forecast. *ASTIN Bulletin* **48**(2), 481–508.

Lundberg S.M. and Lee S.I. (2017) A unified approach to interpreting model predictions. *In Advances in Neural Information Processing Systems,* **30**.

Noll A., Salzmann R. and Wüthrich M.V. (2020) Case study: French motor third-party liability claims. Available at SSRN 3164764.

Parikh N. and Boyd S. (2014) Proximal algorithms. *Foundations and Trends in Optimization* **1**(3), 127–239.

Richman R. (2021a) AI in actuarial science–a review of recent advances-part 1. *Annals of Actuarial Science* **15**(2), 207–229.

Richman R. (2021b) AI in actuarial science–a review of recent advances-part 2. *Annals of Actuarial Science* **15**(2), 230–258.

Scardapane S., Comminiello D., Hussain A. and Uncini A. (2017) Group sparse regularization for deep neural networks. *Neurocomputing* **241**, 81–89.

Shi K. and Shi P. (2024) A sparse deep two-part model for nonlife insurance claims. *Variance* **17**(1).

Sun L., Xia M., Tang Y. and Jones P.G. (2017) Bayesian adjustment for unidirectional misclassification in ordinal covariates. *Journal of Statistical Computation and Simulation* **87**(18), 3440–3468.

Tibshirani R. (1996) Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **58**(1), 267–288.

Wüthrich M.V. (2018) Neural networks applied to chain-ladder reserving. *European Actuarial Journal* **8**, 407–436.

Xia M. and Gustafson P. (2016) Bayesian regression models adjusting for unidirectional covariate misclassification. *The Canadian Journal of Statistics* **44**, 198–218.

Zhang C.H. (2010) Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38**, 894–942.

Zhang H., Wang J., Sun Z., Zurada J.M. and Pal N.R. (2019) Feature selection for neural networks using group LASSO regularization. *IEEE Transactions on Knowledge and Data Engineering* **32**(4), 659–673.

Zou H. and Li R. (2008) One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics* **36**(4), 1509–1533.

Appendix A. MM algorithm for the high-dimensional linear model

A.1 Majorization step

Motivated by the EM algorithm, we construct a surrogate function $Q(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(t)})$ (defined up to an additive constant) for the objective function in (2.8) at iteration t as follows:

$$Q\big(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(t)}\big) = -\frac{1}{N}\sum_{i=1}^{N}\Big\{v_i^* \log f_1(y_i) + z_i^{(t)}(1-v_i^*)[\log q_i + \log f_1(y_i)] \\ + \big(1-z_i^{(t)}\big)(1-v_i^*)[\log(1-q_i) + \log f_0(y_i)]\Big\} + \lambda_Q \sum_{j=1}^{p}|\alpha_j| + \lambda_Y \sum_{j=1}^{p}|\beta_j|,$$

where

$$z_i^{(t)} = \frac{q_i f_1(y_i)}{q_i f_1(y_i) + (1-q_i) f_0(y_i)}\bigg|_{\boldsymbol{\Theta}=\boldsymbol{\Theta}^{(t)}}.$$

https://doi.org/10.1017/asb.2026.10115 Published online by Cambridge University Press

A.2 Minimization step

Update the parameter set $\boldsymbol{\Theta}$ to make $Q(\boldsymbol{\Theta}^{(t+1)}; \boldsymbol{\Theta}^{(t)}) \leq Q(\boldsymbol{\Theta}^{(t)}; \boldsymbol{\Theta}^{(t)})$. The Q function can be decomposed into two parts, $Q(\alpha_0, \boldsymbol{\alpha}; \boldsymbol{\Theta}^{(t)})$ and $Q(\beta_0, \boldsymbol{\beta}, \beta_v; \boldsymbol{\Theta}^{(t)})$, such that

$$Q\left(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(t)}\right) = Q\left(\alpha_0, \boldsymbol{\alpha}; \boldsymbol{\Theta}^{(t)}\right) + Q\left(\beta_0, \boldsymbol{\beta}, \beta_v; \boldsymbol{\Theta}^{(t)}\right),$$

where

$$Q\left(\alpha_0, \boldsymbol{\alpha}; \boldsymbol{\Theta}^{(t)}\right) = -\frac{1}{N}\sum_{i=1}^{N}(1-v_i^*)\left[z_i^{(t)}\log q_i + \left(1-z_i^{(t)}\right)\log\left(1-q_i\right)\right] + \lambda_Q\sum_{j=1}^{p}|\alpha_j|,$$

$$Q\left(\beta_0, \boldsymbol{\beta}, \beta_v; \boldsymbol{\Theta}^{(t)}\right) = -\frac{1}{N}\sum_{i=1}^{N}\left\{\left[v_i^* + z_i^{(t)}(1-v_i^*)\right]\log f_1(y_i) + \left(1-z_i^{(t)}\right)(1-v_i^*)\log f_0(y_i)\right\} + \lambda_Y\sum_{j=1}^{p}|\beta_j|.$$

The minimization step can thus be accomplished separately. Both problems can be addressed using a coordinate descent method.

- For $Q(\alpha_0, \boldsymbol{\alpha}; \boldsymbol{\Theta}^{(t)})$, using the quadratic approximation of its first term at $\alpha_0 + \boldsymbol{x}_i^\top\boldsymbol{\alpha} = \alpha_0^{(t)} + \boldsymbol{x}_i^\top\boldsymbol{\alpha}^{(t)}$ and the fact that $q_i(1-q_i) \leq 1/4$, we further obtain a surrogate function $Q'(\alpha_0, \boldsymbol{\alpha}; \boldsymbol{\Theta}^{(t)})$:

$$Q'\left(\alpha_0, \boldsymbol{\alpha}; \boldsymbol{\Theta}^{(t)}\right) \propto \frac{1}{2N}\sum_{i=1}^{N}(1-v_i^*)\left(\mu_i^{(t)} - \alpha_0 - \boldsymbol{x}_i^\top\boldsymbol{\alpha}\right)^2 + 4\lambda_Q\sum_{j=1}^{p}|\alpha_j|,$$

  where

$$\mu_i^{(t)} = 4\left(z_i^{(t)} - q_i^{(t)}\right) + \alpha_0^{(t)} + \boldsymbol{x}_i^\top\boldsymbol{\alpha}^{(t)}.$$

  The updates are as follows.

$$\alpha_0 \leftarrow \frac{\sum_{i=1}^{N}(1-v_i^*)\left(\mu_i^{(t)} - \boldsymbol{x}_i^\top\boldsymbol{\alpha}\right)}{\sum_{i=1}^{N}(1-v_i^*)},$$

$$\alpha_j \leftarrow \frac{1}{u_j}\mathcal{S}_{4\lambda_Q}(r_j), \quad j = 1, \ldots, p,$$

  where

$$r_j = \frac{1}{N}\sum_{i=1}^{N}(1-v_i^*)\left[\mu_i^{(t)} - \alpha_0 - \sum_{k\neq j}x_{ik}\alpha_k\right]x_{ij},$$

$$u_j = \frac{1}{N}\sum_{i=1}^{N}(1-v_i^*)x_{ij}^2,$$

  and $\mathcal{S}$ is the soft-thresholding operator.
- For $Q(\beta_0, \boldsymbol{\beta}, \beta_v; \boldsymbol{\Theta}^{(t)})$, the form can be simplified as:

$$Q\left(\beta_0, \boldsymbol{\beta}, \beta_v; \boldsymbol{\Theta}^{(t)}\right) \propto \frac{1}{2N}\frac{1}{\sigma^2}\sum_{i=1}^{N}\left[s_i^{(t)}\left(y_i - \beta_0 - \boldsymbol{x}_i^\top\boldsymbol{\beta} - \beta_v\right)^2 + \left(1-s_i^{(t)}\right)\left(y_i - \beta_0 - \boldsymbol{x}_i^\top\boldsymbol{\beta}\right)^2\right] + \lambda_Y\sum_{j=1}^{p}|\beta_j|,$$

https://doi.org/10.1017/asb.2026.10115 Published online by Cambridge University Press

where $s_i^{(t)} = v_i^* + z_i^{(t)}(1 - v_i^*)$. The updates are as follows.

$$\beta_0 \leftarrow \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta} - s_i^{(t)}\beta_v\right),$$
$$\beta_v \leftarrow \frac{\sum_{i=1}^{N} s_i^{(t)}(y_i - \beta_0 - \boldsymbol{x}_i^\top \boldsymbol{\beta})}{\sum_{i=1}^{N} s_i^{(t)}},$$
$$\beta_j \leftarrow \frac{1}{u_j'}\mathcal{S}_{\sigma^2\lambda_\gamma}\left(r_j'\right), \quad j = 1, \ldots, p,$$

where

$$r_j' = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{k\neq j} x_{ik}\beta_k - s_i^{(t)}\beta_v\right)x_{ij},$$
$$u_j' = \frac{1}{N}\sum_{i=1}^{N} x_{ij}^2,$$

Appendix B. Calculation of the evaluation metrics

The formulas for MSE, MAE, and R^2 are provided as follows:

$$\text{MSE} = \frac{1}{|\mathcal{T}_e|}\sum_{i\in\mathcal{T}_e}(y_i - \widehat{y}_i)^2, \quad \text{MAE} = \frac{1}{|\mathcal{T}_e|}\sum_{i\in\mathcal{T}_e}|y_i - \widehat{y}_i|, \quad R^2 = 1 - \frac{\sum_{i\in\mathcal{T}_e}(y_i - \widehat{y}_i)^2}{\sum_{i\in\mathcal{T}_e}(y_i - \bar{y})^2},$$

where $\mathcal{T}_e$ represents the test set, and $|\cdot|$ denotes the number of observations in the set. AUC is computed as the area under the receiver operating characteristic curve. The process for calculating $F1$ score is given as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad F1 = 2 * \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}},$$

where

- True positives (TP) denote the number of truly active predictors correctly included in the model.
- False positives (FP) refer to the number of inactive predictors erroneously selected.
- False negatives (FN) correspond to the number of truly active predictors mistakenly excluded.

Appendix C. Results across multiple data splits

To address the concern that the performance gains of the proposed model might be sensitive to a particular train–validation–test split, we conduct additional experiments using five different random splits of the dataset. For each split, we strictly follow the same data partition procedure described in the main text. Table C1 reports the key performance metrics of the proposed model and all baseline methods across the five splits. In each split, the proposed model outperforms all competing methods across all evaluation metrics. The performance advantage remains stable regardless of the specific partition used. These results demonstrate that the superior performance of the proposed model is not an artifact of a specific data partition but is robust across different sampling variations. This robustness provides stronger evidence that the proposed model generalizes well to unseen data.

https://doi.org/10.1017/asb.2026.10115 Published online by Cambridge University Press

Table C1. *Performance comparison across five random data splits.*

Split	Metric	M1	M2	M3	M4	M5
Split 1	MSE	1.795	1.748	0.902	0.886	0.916
	MAE	0.999	0.977	0.698	0.694	0.700
	R^2	0.445	0.459	0.721	0.726	0.717
Split 2	MSE	1.565	1.520	0.822	0.806	0.828
	MAE	0.943	0.921	0.676	0.671	0.675
	R^2	0.472	0.487	0.723	0.728	0.721
Split 3	MSE	1.570	1.517	0.862	0.855	0.873
	MAE	0.956	0.934	0.685	0.680	0.688
	R^2	0.486	0.503	0.718	0.720	0.714
Split 4	MSE	1.643	1.578	0.893	0.885	0.903
	MAE	0.970	0.945	0.695	0.692	0.697
	R^2	0.485	0.505	0.720	0.723	0.717
Split 5	MSE	1.580	1.531	0.871	0.867	0.877
	MAE	0.957	0.936	0.696	0.696	0.701
	R^2	0.485	0.502	0.716	0.718	0.715

https://doi.org/10.1017/asb.2026.10115 Published online by Cambridge University Press