The New England Journal of Statistics in Data Science Volume 4, 18–36 (2026)
DOI: https://doi.org/10.51387/26-NEJSDS071

Variational Inference of Extremes for Rare Event Modeling: Theory and Applications

Chen QIAN, Chenyang TAO, Jingbin XU, and Feng GUO*

Abstract

Severe event class imbalance is observed in many fields, such as insurance fraud, severe weather, traffic safety, and rare disease, and has far-reaching significance when the generalization of minority event classes is of primary interest. While existing solutions mostly focus on differential sampling or sample re-weighting approaches to alleviate the imbalance issue, we take a novel alternative view on promoting generalization. Our proposal is formulated under the generative Bayesian framework, positing that predictors are the stochastic proxies of latent causes with limited samples, whose exceedance leads to extreme events. To accurately capture the extended tail, our solution adopts the Gumbel distribution as prior and is modeled with a variational inference framework. Following the assertion that exceedance leads to extremes, we devise a disentangled additive monotonic neural architecture to predict the risk. The proposed model acknowledges representation uncertainties while embracing improved interpretability, generalization, and robustness. We provide theoretical insights to show the merits of the proposed approach. To verify the effectiveness in empirical settings, we conducted studies on various real-world data against the state-of-the-art counterparts, with encouraging results reported.

KEYWORDS AND PHRASES: Rare event modeling, Variational inference, Generative Bayesian models, Extreme value theory.

1. INTRODUCTION

Rare event modeling (REM) deals with situations with an extremely low prevalence of events but influential consequences. Common examples of high social-economic value events include vehicle collisions in traffic safety [10], autonomous vehicle testing [34], life-threatening rare diseases in healthcare, and fraudulent financial activities [32]. Accurate and robust modeling of rare events is critical for the identification of the relevant risk factors so as to predict, and hopefully prevent, associated negative outcomes via intervention. As an example, it is estimated that an adult driver in the U.S. will, on average, travel approximately 6.8 million miles before experiencing a single traffic crash [10]. Given that traffic crashes are relatively rare events compared to normal driving, the precise identification and prediction of such incidents are of utmost importance for reducing fatalities and enhancing overall safety [39].

Characterized by severe event class imbalance and the lack of minority labels, rare event modeling falls outside the comfort zone of standard statistical approaches [46]. Without explicit statistical adjustments, the imbalance drives a learning agent to bias toward the majority class; at the same time, the absence of adequate minority examples causes unprotected models to wrongly capture spurious features that do not generalize. Without adequate generalization ability, the implemented REM model would substantially diminish effectiveness in detecting rare events, which are usually associated with risks that people aim to avoid. In the context of traffic crashes, inaccurate predictions by the REM model could potentially result in severe accidents occurring.

To establish reliability for REMs, the implemented model should have sufficient robustness to handle future out-of-sample predictions for the extreme event, i.e. the data which has not been exposed to during the training of REM. [48] point out that learning from less represented events is vulnerable to data perturbations. These generated misjudgments might lead to safety-critical situations. For instance, if practitioners employ the REM model to assess the safety functions of autonomous vehicles, including decisions about whether to execute maneuvers to prevent traffic crashes based on crash predictions, a less robust REM model could potentially permit self-driving cars to take unforeseen actions, possibly resulting in collisions [5].

Pioneered by the work of [23], there has been extensive research to overcome the potential pitfalls of rare-event modeling in the classical statistical regression framework. Prominent examples include coefficient bias correction [23] and penalized maximum likelihood estimation [14]. However, difficulties arise when applying such correction techniques in real-world settings: (*i*) the estimation efficiency for rare event modeling is dictated by the number of events of interest given fixed dimensions, and high-dimensional inputs render the conclusions unreliable and ungeneralizable [46, 43]; (*ii*) heavy tails are commonly expected in rare event

*Corresponding author.

data, and most existing solutions fail to account for such characteristics [48].

Statistical re-sampling and re-weighting are the two most popular sample adjustment strategies to mitigate severe event class unbalancing issues. Re-sampling approaches alter the learner's exposure frequency by over-sampling the minority class and down-sampling the majority class. The re-weighting approach adjusts the relative importance by assigning more weight to the less-represented examples. In line with these practices, recent work represented by [4] and [29] also considers modifications to derive class-sensitive loss that properly penalizes the minority event class, allowing the model to learn efficient representation without getting trapped in the complications associated with sample adjustments.

The re-sampling and re-weighting statistical approaches have limited capability to handle the situation when inputs are outside the norm. Re-sampling typically involves either over-sampling or under-sampling [3]. The former is often subjected to the loss of estimation efficiency [46], the latter is often associated with compromised generalization [4]. The re-weighting schemes are often criticized for the numerical instability [35]. Recent works by [4, 29] modify the hinge loss and entropy loss to capitalize the minority class prediction. However, research points out this might result in overfitting issues due to training bias and label noise [35].

An alternative way to overcome the difficulties of rare event modeling is to solicit inductive bias, thereby imposing structural constraints to suppress the spurious associations and identify robust predictive features. Many such models are framed under the few-shot learning setup, capitalizing on its superior ability to generalize in the small sample regime. These models often operate under relatively strong assumptions that the majority event classes sufficiently characterize the feature encoder and meta-predictors, leaving only a handful of parameters to tune for the minority classes, thereby boosting efficiency and generalization. Sharing an aggregating knowledge across different learning tasks defined in majority classes contributes significantly to the success of few-shot learning on generalization [13]. Few-shot learning relies on the assumption that all classes in the data share common features and leverages knowledge transfer from majority classes to enhance learning for minority classes. However, in reality, this assumption may not always hold true, as different classes can have varying data generation processes. For example, traffic crashes and normal driving segments often display distinct characteristics [16]. The failure to meet the assumptions of few-shot learning methods can potentially lead to a degradation in their performance [11].

Adding to the difficulty is the challenge of discerning spurious correlations as rare events do not always associate with stable, recognizable patterns. Instead, feature irregularities are common for rare events, which makes it unattainable to confidently model the rare event behavior. In such scenarios, rare event modeling can be formulated by "establishing the norm" using the majority of examples and to test if input is out of the ordinary, i.e., the anomaly detection approach with abundant tools developed based on various heuristics [8, 37]. Unsupervised anomaly detection approaches often involve density modeling, which has been a long-standing challenge for high-dimensional inputs [8, 37]

Despite the varying degree of empirical success of existing rare event modeling schemes, a few limitations have been largely overlooked. First, assumptions made by different modeling schemes are often at odds, implying performance depends on whether underlying assumptions have been satisfied. Second, representation has not been sufficiently valued in existing research, potentially taking a toll on the minority class generalization. In light of the above, it is imperative to develop robust and interpretable learning strategies that can fully recognize the unique characteristics of rare-event data.

In recognition of the limitations of current rare-event modeling discussed above, we present a novel learning framework, named *Variational Inference for Rare Event Modeling* (VI-REM), that explicitly addresses some of the weaknesses of existing schemes. Our main contributions include (i) formulation of a variational representation learning scheme based on information theoretic model, allowing disentangled extreme representations for rare-event prediction; (ii) development of a robust and interpretable prediction arm that joins the strength of a generalized additive model and an isotonic neural net; (iii) presentation of theoretical insights that justify the empirical gains through multiple numerical experiments of VI-REM.

2. BACKGROUND

Denote the input features as $\mathbf{x} \in \mathbb{R}^d$, the latent variables as $\mathbf{z} \in \mathbb{R}^k$, and class label $y \in \{0, 1\}$ for which $y = 1$ represents the rare-event of interest with probability of occurrence p, and $y = 0$ as *Normality*. Training instances are given by n paired samples $\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^n$ with sample size n_1 in class *Event* and $n_0 = n - n_1$ in class *Normality*. The goal is to predict the conditional likelihood of an event given the features, i.e., $\Pr(y|\mathbf{x})$ with accuracy and robustness. The rare-event problem arises when $p \ll 1$.

2.1 Information Theoretic Generative Model

The generative Bayesian model has gained popularity due to its scalability, interpretability, and generalizability [24]. Under the full Bayesian setup, the generative process for the data likelihood of y given $\mathbf{x}$ can be obtained by marginalization over latent variables $\mathbf{z}$:

$$\Pr(y|\mathbf{x}) = \frac{\int \Pr(\mathrm{y}, \mathbf{x}, \mathbf{z}) d\mathbf{z}}{\Pr(\mathbf{x})} = \int \Pr(\mathrm{y}|\mathbf{x}, \mathbf{z}) \Pr(\mathbf{z}|\mathbf{x}) d\mathbf{z} \quad (2.1)$$

A common assumption in such generative process is $\mathbf{z}$ contains sufficient information of $\mathbf{x}$ for y, i.e $\Pr(\mathrm{y}|\mathbf{x}, \mathbf{z}) =$

$\Pr(y|\mathbf{z})$. Similar assumption can be found in [41]. With this, Equation 2.1 can be simplified as

$$\Pr(y|\mathbf{x}) = \int \Pr(\mathrm{y}|\mathbf{z}) \Pr(\mathbf{z}|\mathbf{x}) d\mathbf{z} \tag{2.2}$$

For REM which belongs to supervised learning tasks, the core of the generative Bayesian model is to reconstruct the conditional probability of response variable y given predictors $\mathbf{x}$, which is $\Pr(y|\mathbf{x})$, with high accuracy [46, 43].

For REM, maximizing over Equation 2.2 is challenging because of the computational intractability of the integration as well as the difficulty to find the latent $\mathbf{z}$ given high-dimensional $\mathbf{x}$.

To address the computational intractability issue for latent $\mathbf{z}$, the Variational Information Bottleneck (VIB) [1] utilizes the information-theoretic methods to find a meaningful representation $\mathbf{z}$ from $\mathbf{x}$ by maximizing the following equations:

$$\arg\max_{\mathbf{z}} I(\mathbf{z}, y) - \beta I(\mathbf{z}, \mathbf{x}) \tag{2.3}$$

where $I(\mathbf{z}, y)$ measures the mutual information between $\mathbf{z}$ and y, and the parameter $\beta > 0$ is a hyper parameter.

Under the generative Bayesian setup, maximizing Equation 2.4 is identical to minimize the following loss function based on Variational Inference (VI) [24]:

$$\begin{aligned} \mathrm{Loss}_{\mathrm{VIB}} = &- \mathcal{D}_{\mathrm{KL}}\Big(q_\phi(\mathbf{z}|\mathbf{x}) \| \Pr(\mathbf{z})\Big) \\ &+ \beta \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\Big[\log p_\theta(\mathbf{y}|\mathbf{z})\Big] \end{aligned} \tag{2.4}$$

where $\Pr(\mathbf{z})$ denotes a prior distribution on $\mathbf{z}$, $p_\theta(\mathbf{y}|\mathbf{z})$ denotes a conditional distribution of $\mathbf{y}$ given $\mathbf{z}$ parameterized by θ, and $q_\phi(\mathbf{z}|\mathbf{x})$ is a family of distributions parameterized by ϕ to approximate the true posterior density [24].

The first term $\mathcal{D}_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \| \Pr(\mathbf{z}))$ is the KL divergence between $\Pr(\mathbf{z})$ and $q_\phi(\mathbf{z}|\mathbf{x})$. It shows the complexity of the posterior distribution. Smaller values mean that $q_\phi(\mathbf{z}|\mathbf{x})$ is similar to $\Pr(\mathbf{z})$, which implies a simpler posterior. The second term $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{y}|\mathbf{z})]$ shows how well latent $\mathbf{z}$ can reconstruct label $\mathbf{y}$, i.e. the task performance.

The parameter β is used to keep a balance between the performance on the task and the complexity of the latent representation $\mathbf{z}$ [50]. If $\beta = 1$, then Variational Information Bottleneck in Equation 2.4 is the same as maximizing the evidence lower bound (ELBO) of $\Pr(y|\mathbf{x})$, given latent variable $\mathbf{z}$ [41]; if $\beta \neq 1$ Variational Information Bottleneck will be the β-Variational Auto Encoder (VAE) [21] in the version of supervised learning.

For supervised learning tasks, maximizing over the VIB objective can improve model's generalizability [41], because the $\mathcal{D}_{\mathrm{KL}}\Big(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}, \mathbf{y}) \| \Pr(\mathbf{z})\Big)$ regularizes the Kullback–Leibler (KL) divergence between approximated posterior $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}, \mathbf{y})$ and prior distribution $\Pr(\mathbf{z})$ and helps to prevent $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}, \mathbf{y})$ being overfitted because of simplicity in prior distribution. The generative Bayesian model also shows superior robustness over the deterministic models, especially in handling testing data with considerable noises [1].

2.2 Challenges and Motivation

When applying the generative Bayesian model for REM, there are three challenges from statistical perspectives. The first challenge is setting an appropriate prior $\Pr(\mathbf{z})$ for the less represented events. Within the current state of the art (SOTA) approaches, most work uses the isotropic Gaussian [41, 24] or non-parametric based approaches [45]. However, neither of them could be a good fit in REM, where the events of interest are always associated with heavy tail and rarity of observations.

The second challenge in REM comes from the model's generalization ability to unseen data [48], as the events of interest do not always associate with robust patterns. The disentanglement learning approach proves its effectiveness in representing information with stability, especially for the patterns that are less stable [2]. Promising work includes generalized additive models (GAM) [19] and β-Variational Auto Encoder (VAE) [21]. A feasible approach for disentanglement is to assume sets of statistically independent variables for modeling [7]. In this work, we follow the work in [21] to assume sets of statistically independent variables assuring the model's disentanglement in REM. This setup helps to decouple the representations into individual features that each capture a unique aspect of data.

The third challenge arises from how we can build an interpretable prediction model. Isotonic regression, also known as monotonic regression, is an important class of constrained regression techniques that works under the assumption of monotonicity [18, 47]. Research has verified its appropriateness in preventing overfitting and enhancing model interpretability [18]. Specifically, for a scalar predictor x, the relationship of x toward $y = g(x)$ is assumed to be non-decreasing with regard to x; i.e., $g(x_1) \leq g(x_2)$ if $x_1 \leq x_2$. Such relation is ubiquitous in practice, and monotonicity allows extra flexibility compared to linear relations yet retains a similar level of interpretability of the predictive features and modeling robustness. In this paper, we leverage the recent introduction of the unconstrained monotonic neural net (UMNN) [47] to approximate flexible mapping functions in the generative Bayesian model.

3. METHODOLOGY

We formulate VI-REM under the latent variable model framework, in which the event class labels can be predicted from the latent $\mathbf{z}$. The key idea of the VI-REM is to amortize the difficulty of direct prediction of rare-events to the representation learning stage. The premise is the extreme latent representations lead to extreme events. The approach partitions the representation space into normal and extreme

regions, where the prevalence of target events in the latter far exceeds those in the former. This alleviates the majority bias issue that afflicts conventional schemes, as the event distribution is more balanced for the extreme region.

3.1 Extreme Prior Distribution

The Extreme Value Theory (EVT) studies the statistical behavior of extremes [9]. To model the extreme behavior of statistical features, *maximum model* within the domain of EVT describes the maximal value in a sequence of independent observations. Modeling extremes poses a challenge due to the limited number of observations. However, extreme value theory (EVT) asserts that, under mild conditions [9], extreme values actually belong to the Generalized Extreme Value (GEV) family of distributions. For a sequence of independent and identically distributed random variables, denoted as $X_1, X_2, \ldots, X_N$, the asymptotic distribution of the maximum value $\max(X_1, X_2, \ldots, X_N)$ follows:

$$\lim_{N\to\infty} \Pr\left(\max\{X_1, \ldots, X_N\} \leq x\right) = \mathrm{GEV}(\mu_1, \sigma_1, \xi) \quad (3.1)$$

where $\mathrm{GEV}(\mu_1, \sigma_1, \xi)$ denotes a GEV distribution with μ_1 as the location, σ_1 as the scale and ξ as the extreme value index. Equation 3.1 is known as the Fisher–Tippett–Gnedenko theorem [9], allowing for accurate modeling of extreme values with very limited examples.

The GEV has three forms depending on the parameter ξ: Fréchet distribution ($\xi > 0$), Gumbel distribution($\xi = 0$), and reversed Weibull distribution ($\xi < 0$). The Fréchet and Weibull distributions have bounded support, e.g. $x > \mu_1 - \frac{\sigma_1}{\xi}$ for Fréchet distribution. To cover the whole space, we use the Gumbel distribution, denoted as $\mathcal{G}(\mu_1, \sigma_1)$, as the prior because the support for Gumbel is $x \in \mathbb{R}$.

In VI-REM, we assume extreme latent in the representation space lead to the happening of *Event* cases. As these extreme representation usually deviate from the center of normality, we leverage *maximum model* in Equation 3.1 to represent these extreme cases. We extend standard Gaussian prior [24, 41] via incorporating a Gumbel based prior to accommodate heavy tails. We model the normal regular representations with a Gaussian distribution, and the extreme representation with a Gumbel distribution. We assume the representation space of latent variable of z can be partitioned into the into normal and extreme regions. So that the probability density can be:

$$\begin{aligned}\Pr(z) = \Pr(z \in &\textit{Normality}) \times \mathcal{N}(\mu_2, \sigma_2) \\ &+ \Pr(z \notin \textit{Normality}) \times \mathcal{G}(\mu_1, \sigma_1)\end{aligned} \quad (3.2)$$

where $\mathcal{N}(\mu_2, \sigma_2)$ is a Gaussian distribution with parameter μ_2, σ_2. For simplicity, we set $\lambda = \Pr(z \in \textit{Normality})$. We define the extreme prior (EP) as a weighted mixture of a Gumbel distribution and Gaussian distribution in equation 3.2, and its associated probability density function is formulated as:

$$\begin{aligned}p_{\boldsymbol{\kappa}}^{\mathrm{EP}}(z) = &\frac{\lambda}{\sqrt{2\pi\sigma_2^2}} \exp\left\{-\frac{(z-\mu_2)^2}{2\sigma_2^2}\right\} \\ &+ \frac{1-\lambda}{\sigma_1} \exp\left\{-\frac{z-\mu_1}{\sigma_1} - \exp\left(-\frac{z-\mu_1}{\sigma_1}\right)\right\}\end{aligned} \quad (3.3)$$

To induce more flexibility into the model, we set the parameters $\boldsymbol{\kappa} = (\lambda, \mu_1, \sigma_1, \mu_2, \sigma_2)$ of EP prior are all learnable during the optimization process. The parameters μ_1, μ_2 controls the center of Normal and Gumbel distribution. The parameter σ_1 controls the shape of distribution tail. Figure 1a displays the scatter plot for different EP distributions. Theorem 3.1 states that as long as $\sigma_1, \sigma_2 > 0$, the proposed EP prior is a proper prior. If the tailed behavior does not actually lead to rarity, the EP can fall back to a standard Gaussian distribution.

Proposition 3.1. The proposed EP distribution is a proper prior, satisfying:

$$\int_{z\in R} p_{\boldsymbol{\kappa}}^{\mathrm{EP}}(z)dz = 1 \quad (3.4)$$

where $\boldsymbol{\kappa} = (\lambda, \mu_1, \sigma_1, \mu_2, \sigma_2)$ are the parameters for the EP prior and $\sigma_1, \sigma_2 > 0$.

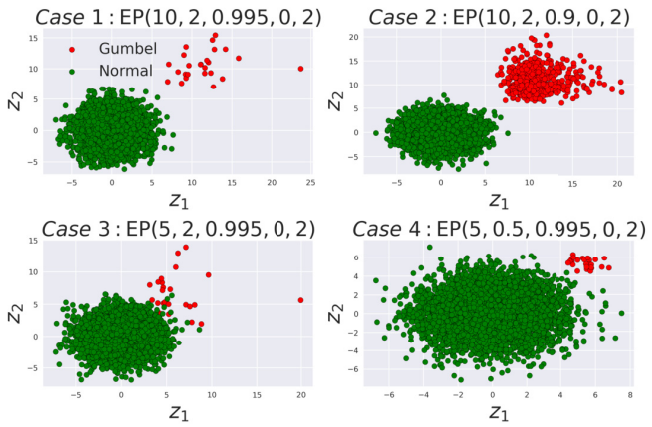
Figure 1b compares the performance of different estimators in modeling the tail part of the distributions. The fitted Normal distribution considerably underestimates the tail portion, while non-parametric kernel density estimation tends to overfit the observed samples. In scenarios involving heavy-tailed distributions, which are commonly observed in risk-related analyses [9], traditional parametric methods (such as fitted Normal distribution) and non-parametric methods (like kernel density estimation) may struggle to accurately capture the tail probabilities due to the limited number of rare events. However, extreme value theory-based estimation (EP distribution) excels in handling heavy-tailed scenarios effectively. It not only addresses the challenges posed by limited rare events but also demonstrates substantial generalization ability in accurately describing the tails of distributions.

EP distribution can be generalized to multi-dimensional latent components by applying the above setup to each individual dimension. We assume mutual independence among multi-dimensional latents $\mathbf{z} \in \mathbb{R}^k$, so as the EP distribution for $\mathbf{z}$ with parameter $\boldsymbol{\omega} \in \mathbb{R}^{5k}$ can be extended as where $\mathbf{z}_i \in \mathbb{R}$:

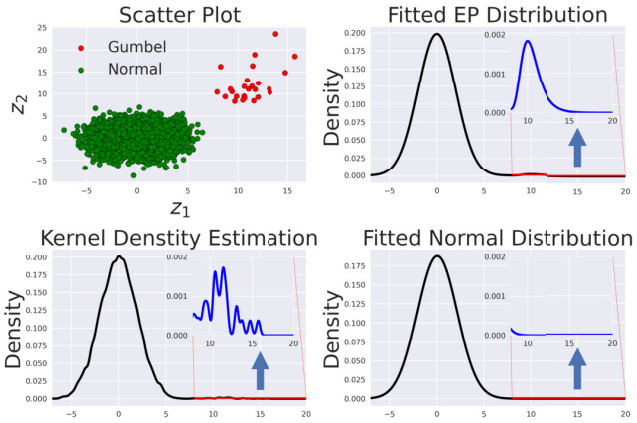
$$p_{\boldsymbol{\omega}}^{\mathrm{EP}}(\mathbf{z}) = \prod_{i=1}^{k} p_{\boldsymbol{\kappa}_i}^{\mathrm{EP}}(\mathbf{z}_i) \quad (3.5)$$

where $\boldsymbol{\kappa}_i = (\lambda_i, \mu_{1,i}, \sigma_{1,i}, \mu_{2,i}, \sigma_{2,i})$ and $\boldsymbol{\omega} = (\boldsymbol{\kappa}_1, \ldots, \boldsymbol{\kappa}_k)$.

The decomposition above assumes that the prior distribution is independent, which implies a possible disentanglement of the latent factor $\mathbf{z}$ [21]. One can also consider



(a) Scatter Plot for Different EP Distributions



(b) Different Estimators for $z \sim \text{EP}(10, 0.2, 0.995, 0, 2)$

Figure 1: Illustration of the Extreme Prior distribution. z_1 and z_2 are simulated independently in the figure. (a) The scatter plot for different EP distributions and the red points come from the Gumbel distribution part and the green points come from the Normal distribution part. For case 1, both of the latent variables $z_1, z_2 \sim \text{EP}(10, 2, 0.995, 0, 2)$. (b) We set both the latent factors $z_1, z_2 \sim \text{EP}(10, 0.2, 0.995, 0, 2)$ and as z_1, z_2 comes from the same distribution, we only show the results for the probability density estimation for z_1. We consider the EP distribution, kernel density estimation (a non-parametric approach), and Normal distribution to fit the distribution of z_1. The black line represents the estimated entire distribution functions, and the blue displays the distribution function for the tailed part.

dependent priors for $\mathbf{z}$. For example, a Gaussian process-based prior $\Pr(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathcal{K}zz)$ with a kernel matrix $\mathcal{K}zz$ [6] to account for dependence in the prior distribution, or to introduce a hierarchical structure to the prior with $\Pr(\mathbf{z}) = \Pr(\mathbf{z}_1) \prod_{i=2}^{k} \Pr(\mathbf{z}_i \mid \mathbf{z}_{i-1})$ [42]. However, more complex priors will increase the number of parameters to be optimized, thus to increase the overfitting risk in REM [45]. We focus on the independent prior in this work, and defer more flexible priors to future research.

3.2 Monotonic Additive Neural Network

To facilitate interpretability and generalization for REM, it is beneficial to impose constraints on $p_{\boldsymbol{\theta}}(y|\mathbf{z})$. We combine two prominent techniques: the generalized additive model (GAM) [19] and isotonic regression for $p_{\boldsymbol{\theta}}(y|\mathbf{z})$, and we name it as Monotonic Additive Neural Network (MANN). Let y come from a Bernoulli(p) distribution and $\psi_{\boldsymbol{\theta}}(y|\mathbf{z})$ be the Logit function and the $p_{\boldsymbol{\theta}}(y|\mathbf{z})$ can be the following:

$$p_{\boldsymbol{\theta}}(y|\mathbf{z}) = \frac{1}{1 + \exp\Big(- \psi_{\boldsymbol{\theta}}(y|\mathbf{z})\Big)} \tag{3.6}$$

We assume sets of statistically independent variables [7] in MANN to assure model's disentanglement. Each f_j takes input from the j-th dimension of the latent to model the generation of y. This additive decomposition encourages disentangled representation. The Logit function $\psi_{\boldsymbol{\theta}}(y|\mathbf{z})$ is based on GAM:

$$\psi_{\boldsymbol{\theta}}(y|\mathbf{z}) = \log\left(\frac{p_{\boldsymbol{\theta}}(y|\mathbf{z})}{1 - p_{\boldsymbol{\theta}}(y|\mathbf{z})}\right) = \alpha(\boldsymbol{\theta}) + \sum_{j=1}^{k} f_j(\mathbf{z}_j) \tag{3.7}$$

where each of the individual function is in Equation 3.8, $j = 1, ..k$ and k is the dimension of latent variables.

$$f_j(\mathbf{z}_j) = \beta_j(\theta) \int_0^{\mathbf{z}_j} \exp\Big(h_j(t; \boldsymbol{\theta})\Big) \mathrm{d}t \tag{3.8}$$

It is easy to see all f_j are monotonic, with the direction dictated by the sign of β_j. We model $h_j(t; \boldsymbol{\theta})$ above using MLP based deep neural networks, and compute $f_j(\mathbf{z}_j)$ via Clenshaw-Curtis numerical integration [47].

We use a toy examples to illustrate the advantages of monotonicity in dealing with REM. We set y from a Bernoulli distribution and the conditional probability follows

$$\Pr(y|z_1, z_2) = \frac{1}{1 + \exp(-z_1^3 - 0.5 z_2^3)}$$

where the random variables $z_1 \sim \text{EP}(2, 0.5, 0.99, -2, 0.5)$ and $z_2 \sim \text{EP}(1, 0.5, 0.99, -1, 0.5)$. Through this, the probability for $\Pr(y = 1) = 0.02$. We set a 2×500 fully connected layers for both MANN and MLP. From Figure 2, the MANN models shows better fittings for the low data density region.

3.3 Variational Inference Learning

Our VI-REM model combines the MANN and EP prior distribution. We establish VI-REM on the Multilayer Perceptron (MLP), which is flexible to handle different learning tasks. We adopt the setup from VIB [1], the loss function to

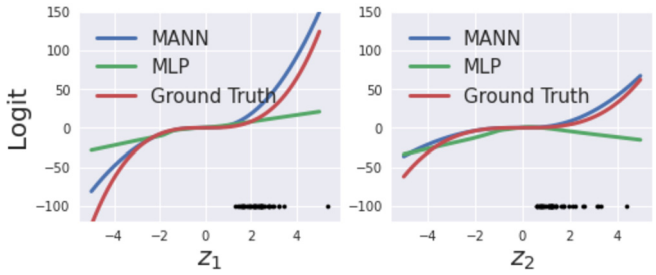


Figure 2: The marginal relationship of z_1, z_2 with the Logit value of condition probability, that is $\log\left(\frac{\Pr(y|z_1)}{1-\Pr(y|z_1)}\right)$ and $\log\left(\frac{\Pr(y|z_2)}{1-\Pr(y|z_2)}\right)$ for z_1 and z_2. The black points represent the observed samples of the Gumbel part.

be minimized for VI-REM is:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\omega}) = \mathcal{D}_{\mathrm{KL}}\Big(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p^{\mathrm{EP}}_{\boldsymbol{\omega}}(\mathbf{z})\Big) \\ -\beta\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\Big(\log p_{\boldsymbol{\theta}}(y|\mathbf{z})\Big)\end{aligned} \tag{3.9}$$

where β is empirically set to be n_0/n_1 for models to be better optimized. $p^{\mathrm{EP}}_{\boldsymbol{\omega}}(\mathbf{z})$ is the EP distribution with parameters $\boldsymbol{\omega}$.

The $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ is a family of distributions to approximate the true posterior density [24]. Practitioners usually refer $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ as Encoder structure as it reduces the original high-dimensional features $\mathbf{x}$ into lower dimensions $\mathbf{z}$. The $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ can be set as isotropic Gaussian distributions through:

$$q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) = \mathcal{N}\Big(\mathbf{z} \mid \mu_\phi(\mathbf{x}), \sigma^2_\phi(\mathbf{x}) \times \mathbf{I}\Big)$$

where $\mathbf{I}$ is the identity matrix. The parameters $\mu_\phi(\mathbf{x}), \sigma^2_\phi(\mathbf{x}) = g_{\boldsymbol{\phi}}(\mathbf{x})$ and $g_\phi(\mathbf{x})$ is established based on MLP parameterized by $\boldsymbol{\phi}$. This is a standard choice in literature of VI [25]. The usage of isotropic Gaussian distributions imply the individual components of the posterior $q_\phi(\mathbf{z}_1|\mathbf{x})$, ..., $q_\phi(\mathbf{z}_K|\mathbf{x})$ are mutually independent [25]. For more complex cases with dependent posterior, we can use distributions with tractable likelihood, such as normalizing flows [36] or inverse autoregressive flows [26]. Our results show that isotropic Gaussian distributions perform well. We use them in the following analysis. $\log p_{\boldsymbol{\theta}}(y|\mathbf{z})$ is based on MANN parameterized by $\boldsymbol{\theta}$:

$$\log p_{\boldsymbol{\theta}}(y|\mathbf{z}) = -\log\Big\{1+\exp\big(-\alpha(\boldsymbol{\theta}) - \sum_{j=1}^{k} f_j(\mathbf{z}_j)\big)\Big\} \tag{3.10}$$

Updating the parameters $\boldsymbol{\theta}$, $\boldsymbol{\phi}$, $\boldsymbol{\omega}$ in VI-REM is based on stochastic gradient descent. Updating the parameter $\boldsymbol{\phi}$ is the same as the strategies in [41, 24]. As for parameters $\boldsymbol{\omega} \in \mathbb{R}^{5k}$ for $p^{\mathrm{EP}}_{\boldsymbol{\omega}}(\mathbf{z})$ and we calculate the derivatives with Equation 3.5:

$$\begin{aligned}\nabla_{\boldsymbol{\omega}}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\omega}) &= \nabla_{\boldsymbol{\omega}}\mathcal{D}_{\mathrm{KL}}[q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p_{\boldsymbol{\omega}}(\mathbf{z})] \\ &= -\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\Big(\sum_{i=1}^{k}\nabla_{\boldsymbol{\omega}}\log p_{\boldsymbol{\kappa}_i}(\mathbf{z}_i)\Big)\end{aligned} \tag{3.11}$$

The parameters $\lambda_1, \ldots, \lambda_k \in (0,1)$ determine the weight for the Gumbel distribution within the EP prior. Their derivatives can be calculated as follows:

$$\begin{aligned}\nabla_{\lambda_i}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\omega}) &= -\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}_i|\mathbf{x})}\Big(\frac{\nabla_{\lambda_i}p_{\boldsymbol{\kappa}_i}(\mathbf{z}_i)}{p_{\boldsymbol{\kappa}_i}(\mathbf{z}_i)}\Big) \\ &= -\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}_i|\mathbf{x})}\Big\{\frac{h_{1,i}-h_{2,i}}{\lambda_i h_{1,i}+(1-\lambda_i)h_{2,i}}\Big\}\end{aligned} \tag{3.12}$$

where $i = 1, \ldots, k$, and k is the dimension of the latent variable $\mathbf{z}$; $\mathbf{z}_i$ is the ith component of latent $\mathbf{z}$; λ_i is the parameter λ for latent $\mathbf{z}_i$; $h_{1,i} = \frac{1}{\sqrt{2\pi\sigma^2_{2,i}}}\exp\left(\frac{-(\mathbf{z}_i-\mu_{2,i})^2}{2\sigma^2_{2,i}}\right)$; and $h_{2,i} = \frac{1}{\sigma_{1,i}}\exp\left(\frac{-(\mathbf{z}_i-\mu_{1,i})}{\sigma_{1,i}} - \exp\left(\frac{-(\mathbf{z}_i-\mu_{1,i})}{\sigma_{1,i}}\right)\right)$.

During the training process, parameter λ_i might collapse to zero due to sparse parameter regime. To avoid this issues, we set parameter λ_i through a logistic function as $\lambda_i = \frac{1}{1+\exp(-\Lambda_i)}$. This transformation enables us to update the weight for $\lambda_i \in (0,1)$ by updating the real-valued parameter $\Lambda_i \in \mathbb{R}$. The logistic function helps prevent the parameter λ_i from collapsing to zero. To obtain λ_i, we utilize the chain rule on Equation 3.12 to compute the derivatives of $\nabla_{\Lambda_i}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\omega})$. The effectiveness of this approach is confirmed through numerical studies in Sections 5 and 6.

The calculation of derivatives $\boldsymbol{\theta}$ involves the derivatives of each individual function $f_j(\mathbf{z}_j)$ in Equation 3.8. Based on Leibniz integral rule, we can have:

$$\nabla_{\boldsymbol{\theta}} f_j(\mathbf{z}_j) = \int_{s_0}^{\mathbf{z}_j}\nabla_{\boldsymbol{\theta}}\exp\Big(h_j(t;\boldsymbol{\theta})\Big)\mathrm{d}t + \nabla_{\boldsymbol{\theta}}\beta_j(\boldsymbol{\theta}) \tag{3.13}$$

Based on Equation 3.13, the derivatives of the Logit function $\psi_{\boldsymbol{\theta}}(y|\mathbf{z})$ with respect to $\boldsymbol{\theta}$ can be obtained through:

$$\begin{aligned}\nabla_{\boldsymbol{\theta}}\psi_{\boldsymbol{\theta}}(y|\mathbf{z}) = \nabla_{\boldsymbol{\theta}}\alpha(\boldsymbol{\theta}) + \sum_{j=1}^{k}\nabla_{\boldsymbol{\theta}}\beta_j(\boldsymbol{\theta}) \\ + \sum_{j=1}^{k}\int_{s_0}^{\mathbf{z}_j}\nabla_{\boldsymbol{\theta}}\exp\Big(h_j(t;\boldsymbol{\theta})\Big)\mathrm{d}t\end{aligned} \tag{3.14}$$

With $\mathbf{z} \sim q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$, the derivatives w.r.t parameters $\boldsymbol{\theta}$ are:

$$\begin{aligned}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\omega}) &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\Big[\nabla_{\boldsymbol{\theta}}\log p_\theta(y|\mathbf{z})\Big] \\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\Big[\frac{\exp(-\psi_\theta(y|\mathbf{z}))}{1+\exp(-\psi_\theta(y|\mathbf{z}))}\times\nabla_{\boldsymbol{\theta}}\psi_{\boldsymbol{\theta}}(y|\mathbf{z})\Big]\end{aligned} \tag{3.15}$$

With the optimized parameters $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ through the introduced learning process, we can generate the predictions for $\Pr(y|\mathbf{x})$

through a Monte Carlo estimation where L is number of generated samples, where $\mathbf{z}^{(l)} \sim q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$:

$$\Pr(y|\mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\Big(\log p_{\boldsymbol{\theta}}(y|\mathbf{z})\Big) \approx \frac{1}{L}\sum_{l=1}^{L} p_{\boldsymbol{\theta}}(y|\mathbf{z}^{(l)}) \quad (3.16)$$

3.4 Theoretical Justifications of VI-REM

We theoretically discuss modeling the tail in REM can improve generalization power. The generalization power of $\Pr(y|\mathbf{x})$ involve $q_\phi(\mathbf{z}|\mathbf{x})$ and $p_\theta(y|\mathbf{z})$ as in Equation 3.16. The asymptotic behavior of $q_\phi(\mathbf{z}|\mathbf{x})$ through VI has been extensively discussed in [49]. This paper focuses on $p_\theta(y|\mathbf{z})$.

We use *generalization gap* to measure model's generalization ability, which is defined as the difference between the model's performance on training data and on unseen testing data. Denote $\mathrm{P}_n\big(g(\mathbf{z})\big)$ and $\Pr\big(g(\mathbf{z})\big)$ respectively as the empirical measure and underlying ground truth for some function $g(\mathbf{z})$. Define $\mathcal{F}$ as the set of all monotonic functions, let f be the candidate function belonging to function space $\mathcal{F}$, and $f_0 \in \mathcal{F}$ is the ground truth. We use $\mathrm{L}_f(\mathbf{z}, y)$ to denote the loss measured by function f taking value at $\mathbf{z}$ for the response y. We identify $\Omega_\eta \triangleq \{\|\mathbf{z}\| < \eta\}$ as an η-ball where the bulk of the distribution mass resides, i.e., $\Pr(\Omega_\eta) \geq 1-\epsilon$, $\epsilon \ll 1$, and consequently $\Omega_\eta^C \triangleq \mathbb{R}^d \backslash \Omega$ contains the extremes.

We point out the theoretical generalization gap with the introduction of cutoff points in Theorem 3.2. The following regularized assumptions are needed to establish Theorem 3.2:

C.1 The function $f(\mathbf{z})$ follows the Lipschitz property:

$$|f(\mathbf{z}) - f(\mathbf{z}')| \leq k_n \|\mathbf{z} - \mathbf{z}'\|$$

where k_n is some positive constant.

C.2 The ground truth function f_0 satisfies the condition:

$$f_0 = \arg\min_{f\in\mathcal{F}} \mathbb{E}\Big(\Pr\big(\mathrm{L}_{f_0}(\mathbf{z}, y)\big)\Big)$$

C.3 Function $f(\mathbf{z})$ satisfies the condition for nonzero $\mathbf{z}$:

$$\sup_{f\in\mathcal{F}} |f(\mathbf{z}) - f_0(\mathbf{z})| \leq \sup_{f\in\mathcal{F}} \|f - f_0\| \|\mathbf{z}\|$$

Theorem 3.2. *Under the assumptions in C.1–C.3 described above, the following generalization gap holds*

$$\begin{aligned}
&\mathbb{E}\Big(\sup_{f\in\mathcal{F}} \big|\mathrm{P}_n\big(\mathrm{L}_f(\mathbf{z}, y)\big) - \Pr\big(\mathrm{L}_{f_0}(\mathbf{z}, y)\big)\big|\Big) \\
&\leq \sqrt{\frac{\log|\mathcal{F}| + \log(1/\delta)}{2n}} + \gamma h^{1/2}\eta^{1/2} \sup_{f\in\mathcal{F}} \|f - f_0\|^{1/2}_{\mathbf{z}\in\Omega_\eta} \\
&+ \gamma \sup_{\mathbf{z}\in\Omega_\eta^C} \|\mathbf{z}\|^{1/2}(1-h)^{1/2} \sup_{f\in\mathcal{F}} \|f - f_0\|^{1/2}_{\mathbf{z}\in\Omega_\eta^C}
\end{aligned} \quad (3.17)$$

where $\delta = |\mathcal{F}|e^{-2n\phi^2}$ and $|\mathcal{F}|$ is the cardinality of space $\mathcal{F}$. ϕ, γ are some constant, $I_{\mathbf{z}\in\Omega_\eta:\|\mathbf{z}\|\leq\eta}$ is parameter λ in EP distribution, and h is the empirical value of $I_{\mathbf{z}\in\Omega_\eta:\|\mathbf{z}\|\leq\eta}$.

Through Theorem 3.2, we decompose the generalization gap into estimation error, the approximation error for the non-tailed part, and the approximation error for the tailed part over a certain cutoff point as stated in the following Remark.

Remark 1. *Assume $f_1, f_2 \in \mathcal{F}$ are two functions belonging the space $\mathcal{F}$. We set f_1 as the $\log p_{\boldsymbol{\theta}}(y|\mathbf{z})$ which is obtained by maximizing Equation 3.9. Similarly, we set f_2 as $\log p_{\boldsymbol{\theta}'}(y|\mathbf{z}')$ is obtained by optimizing*

$$\beta\mathbb{E}_{q_{\boldsymbol{\phi}'}(\mathbf{z}'|\mathbf{x})}\Big(\log p_{\boldsymbol{\theta}'}(y|\mathbf{z}')\Big) - \mathcal{D}_{\mathrm{KL}}\Big(q_{\boldsymbol{\phi}'}(\mathbf{z}'|\mathbf{x}) \mid p^{Gauss}(\mathbf{z}')\Big)$$

where $p^{Gauss}(\mathbf{z})$ is a isotropic Gaussian distribution. Denote $\Delta_{\Omega_\eta^C} = ||f_2 - f_0||^{1/2}_{\mathbf{z}\in\Omega_\eta^C} - ||f_1 - f_0||^{1/2}_{\mathbf{z}\in\Omega_\eta^C}$ and $\Delta_{\Omega_\eta} = ||f_1 - f_0||^{1/2}_{\mathbf{z}\in\Omega_\eta} - ||f_2 - f_0||^{1/2}_{\mathbf{z}\in\Omega_\eta}$. As long as the follows holds:

$$\Delta_{\Omega_\eta^C}/\Delta_{\Omega_\eta} > \frac{\eta^{1/2}}{r \sup_{\mathbf{z}\in\Omega_\eta^C} \|\mathbf{z}\|^{1/2}} \Big(\frac{h}{1-h}\Big)^{1/2} \quad (3.18)$$

where r is some constant. Then based on conditions C.1–C.3 and Theorem 3.2, the generalization gap over function f_1, f_2 can have the following:

$$\begin{aligned}
&\mathbb{E}\Big(\sup_{f_1\in\mathcal{F}} \big|\mathrm{P}_n\big(\mathrm{L}_{f_1}(\mathbf{z}, y)\big) - \Pr\big(\mathrm{L}_{f_0}(\mathbf{z}, y)\big)\big|\Big) \\
&< \mathbb{E}\Big(\sup_{f_2\in\mathcal{F}} \big|\mathrm{P}_n\big(\mathrm{L}_{f_2}(\mathbf{z}', y)\big) - \Pr\big(\mathrm{L}_{f_0}(\mathbf{z}', y)\big)\big|\Big)
\end{aligned} \quad (3.19)$$

Remark 1 supports that the proposed extreme prior can improve model generalization. The misrepresented tailed part in REM amplifies the approximation error, thus restricting the model's generalizability. Intuitively from Figure 1b, utilizing the heavy-tailed prior distribution remedies the misrepresentation problem, thus helps to tighten the generalization bound.

4. MODEL IMPLEMENTATIONS

4.1 Model Interpretation

The interpretation of VI-REM, i.e., the relationship between $\mathbf{x}$ and label $\mathbf{y}$, needs to be made through latent factor $\mathbf{z}$. The MANN component in recognizing the $p_{\boldsymbol{\theta}}(y|\mathbf{z})$ makes the relationship between latent factors $\mathbf{z}$ and response y easy to assess. However, interpreting the relationship between $\mathbf{x}$ and $\mathbf{z}$ is challenging due to the black-box neural network. We frame the model interpretability as a variable selection process, i.e., identifying the most relevant $\mathbf{x}$ for each of the latent factors.

The core of this interpretation approach is the evaluation of the statistical dependence between $\mathbf{x}$ and $\mathbf{z}$. Recent studies suggest that distance correlation [44] is a powerful non-parametric tool to measure the statistical correlation between two random variables. The MANN structure guarantees that the latent factors $\mathbf{z}$ in VI-REM are largely statistically independent. Based on this fact, we can use the

Distance Correlation [44] to select the critical $\mathbf{x}$ in determining $\mathbf{z}_1$, ...$\mathbf{z}_k$.

Considering the high dimensionality of the input features $\mathbf{x}$, we adopt the Distance Correlation Sure Independence Screening (DC-SIS) framework [27], which is the SOTA approach to use Distance Correlation to select high-dimensional features. The essence of a DC-SIS is to rank variables based on the Distance Correlation. Under our setup, we can use the Distance Correlation between $\mathbf{z}_i$ and each of the input features $\mathbf{x}_1$, ..., $\mathbf{x}_d$ to rank the importance of input features and select the most relevant ones. However, when directly using the DC-SIS to the proposed VI-REM model, the $\mathcal{O}(n^2)$ computational complexity for Distance Correlation is intolerable for a large sample size [28], especially when hundreds of thousands of training data. Therefore, we adopted a Divide and Conquer algorithm [28] to overcome the computation challenge.

For latent factor $\mathbf{z}_j$, we equally spilt the n samples into G groups with $s = n/G$ data points in each subgroup. For each subgroup $g \in [1, \dots G]$ with s samples, we measure the Distance Correlation of $\mathbf{x}_l^g$ and $\mathbf{z}_j^g$, where $\mathbf{x}_l^g$ represents the input feature $\mathbf{x}_l$ within subgroup g and $\mathbf{z}_j^g$ represents the latent factor $\mathbf{z}_j$ within subgroup g. For simplicity, denote $\mathbf{u} = \mathbf{x}_l^g$ and $\mathbf{v} = \mathbf{z}_j^g$, d_u, d_v as the dimension for $\mathbf{u}$ and $\mathbf{v}$, and $\mathbf{u}^{(i)}$, $\mathbf{v}^{(i)}$ as the i th sample of $\mathbf{u}$ and $\mathbf{v}$ correspondingly. The detailed formula for estimation of distance correlation is as follows:

$$\widehat{\mathrm{DC}}_l^g = \widehat{\mathrm{dcorr}}(\mathbf{x}_l^g, \mathbf{z}_j^g) = \frac{\widehat{\mathrm{dcov}}(\mathbf{u}, \mathbf{v})}{\sqrt{\widehat{\mathrm{dcov}}(\mathbf{u}, \mathbf{u})\widehat{\mathrm{dcov}}(\mathbf{v}, \mathbf{v})}} \tag{4.1}$$

where

$$\widehat{\mathrm{dcov}}^2(\mathbf{u}, \mathbf{v}) = \widehat{S}_1^g + \widehat{S}_2^g - 2\widehat{S}_3^g \tag{4.2}$$

and $\widehat{S}_1^g$, $\widehat{S}_2^g$, $\widehat{S}_3^g$ for $\widehat{\mathrm{dcov}}^2(\mathbf{u}, \mathbf{v})$ are as follow:

$$\begin{aligned}
\widehat{S}_1^g &= \frac{1}{s^2}\sum_{i=1}^{s}\sum_{j=1}^{s}\left\|\mathbf{u}^{(i)} - \mathbf{u}^{(j)}\right\|_{d_u}\left\|\mathbf{v}^{(i)} - \mathbf{v}^{(j}\right\|_{d_v} \\
\widehat{S}_3^g &= \frac{1}{s^3}\sum_{i=1}^{s}\sum_{j=1}^{s}\sum_{l=1}^{s}\left\|\mathbf{u}^{(i)} - \mathbf{u}^{(l)}\right\|_{d_u}\left\|\mathbf{v}^{(j)} - \mathbf{v}^{(l)}\right\|_{d_v} \\
\widehat{S}_2^g &= \frac{1}{s^2}\sum_{i=1}^{s}\sum_{j=1}^{s}\left\|\mathbf{u}^{(i)} - \mathbf{u}^{(j)}\right\|_{d_u}\frac{1}{s^2}\sum_{i=1}^{s}\sum_{j=1}^{s}\left\|\mathbf{v}^{(i)} - \mathbf{v}^{(j)}\right\|_{d_v}
\end{aligned} \tag{4.3}$$

For each subgroup, we conduct the DC-SIS [27] to select the set with the most contribution of input features $\mathbf{x}$ toward the latent factor $\mathbf{z}_j$. In practice for each of the subgroup g, the selection process for a given size $q < s = n/G$ can be:

$$\widehat{\mathcal{D}}_g = \left\{ l : \widehat{\mathrm{DC}}_l^g \text{ is among top } q \text{ largest of } d \right\}, \tag{4.4}$$

where $l \in [1, \dots, d]$ and d is the dimension of input features $\mathbf{x}$.

After conducting these G (in total we have G subgroups) independent calculations and selections, the final component of $\mathbf{z}_j$ denoted by $\widehat{\mathcal{D}}_{\mathbf{z}_j}$ is established by: $\widehat{\mathcal{D}}_{\mathbf{z}_j} = \cap_{g\in\{1,\dots G\}}\widehat{\mathcal{D}}_g$. In practice, we can simply uses the most common selected relevant features $\mathbf{x}_1$, ..., $\mathbf{x}_d$ in each of the subgroup procedures. Repeat the above process for all $\mathbf{z}_j$ and calculate the $\widehat{\mathcal{D}}_{\mathbf{z}_j}$; i.e., the identified component of latent factors.

Theorem 4.1. *Denote $\mathcal{D}^*_{\mathbf{z_j}}$ as the true components associated with latent $\mathbf{z_j}$, and $\widehat{\mathcal{D}}_{\mathbf{z_j}}$ as the set identified by the previously mentioned model's interpretation strategy. Then:*

$$\begin{aligned}
\Pr\left(\mathcal{D}^*_{\mathbf{z_j}} \subseteq \widehat{\mathcal{D}}_{\mathbf{z_j}}\right) \geq \Big\{1 - \mathcal{O}\{d\exp[-c_1 n^{\alpha-2\alpha(u+v)}] \\
+ dn^{\alpha}\exp(-c_2 n^{v})\}\Big\}^{G} \to 1
\end{aligned} \tag{4.5}$$

where $G = n^{1-\alpha}$, $\alpha > 0$, d is the dimension of input features $\mathbf{x}$, c_1, c_2, u, v are some constant and $u + v < 1/2$.

Theorem 4.1 shows that as long as the sample size n is getting sufficiently large, the true risk modulators for the learnt latents have high confidence to be selected by the given procedure latent factor identification.

4.2 Robustness Evaluations

A robust prediction model should be less vulnerable to attacks and remain valid when testing data is out of domain [5]. Building a robust model is critical for deploying deep learning models in the application of safety-critical situation [48]. Robustness measures the model's ability to handle data perturbations and helps understand the boundary of successful applications. Mathematically for a model f and its predictions over x $f(x)$, robustness evaluations aims to find how the model f can handle $f(x')$, where x' has minor perturbations to x. if $f(x')$ does not show much difference with $f(x)$, then model f is robust and trustworthy to deploy [48].

Researchers have proposed various approaches under different application scenarios to generate x' to evaluate robustness. Regarding natural language processing tasks, [51] summarizes the current well-established methods and suggests mixing the words from different documents [22, 51] could be a simple and efficient approach. For computer vision studies, image blurring and image rotation are all useful in practice. In this work, we use the rotation operator to investigate whether the model can efficiently extract invariant features from rare events.

4.3 Extension to General Setups

The previous VI-REM model mostly focuses on binary classification, and we generalize it to multi-label classification in risk analysis. Suppose we have L + 1 classified labels, and we treat one dominant class as class *Normality*, the rest L minority classes as *Event*. For example, traffic crashes are rare and can be classified from the most severe fatal crashes to injury crashes to minor property-damage only crashes. To mimic the general settings in risk analysis, we assume the severity of the L minority class ranges from low to high.

Without generality, we set 1, ..., L as labels for the L minority classes, and class 1 as the least severe *Event* and class L is the most severe one. As the response variable, the severity of an event increases from class 1 toward class L and belongs to the ordinal variables, we applied the setup from the Ordered Logit Model (OLM). Under the OLM setup, we want to model $\Pr(y \geq m|\mathbf{x})$ given different event severity where $m = 1,,,.L$. So that under the VI-REM, we modify the Logit function in Equation 3.6 through:

$$\psi_{\boldsymbol{\theta}}\big(y \geq m|\mathbf{z}\big) = \log\left(\frac{p_{\boldsymbol{\theta}}(y \geq m|\mathbf{z})}{1 - p_{\boldsymbol{\theta}}(y \geq m|\mathbf{z})}\right) \tag{4.6}$$

and the Logit function in the OLM based VI-REM follows:

$$\psi_{\boldsymbol{\theta}}\big(y \geq m|\mathbf{z}\big) = \boldsymbol{\alpha_m}(\boldsymbol{\theta}) + \sum_{j=1}^{k} f_j(\mathbf{z}_j) \tag{4.7}$$

where the coefficients $\boldsymbol{\alpha_m}(\boldsymbol{\theta})$ are intercepts for different event categories and $0 < \boldsymbol{\alpha_1}(\boldsymbol{\theta}) < \cdots < \boldsymbol{\alpha_L}(\boldsymbol{\theta})$. The individual function $f_j(\mathbf{z}_j)$ function follows the same as Equation 3.8.

4.4 Experiment Setup

We considered the following machine learning approaches for comparison: (1) Sampling MLP, short as **S-MLP**, is a re-sampling scheme based on Multilayer Perceptron (MLP) over-samples from the less-represented class [3]. The hyperparameter r in the S-MLP is used to control the over-sample ratio, which is equal to the expected n_1 over n_0 after over-sampling. We use grid search cross-validation to select the sampling ratio $r \in (0.01, 0.05, 0.1, 0.2)$. (2) Focal MLP, short as **F-MLP**, uses a re-weighting based loss that adapts the cross entropy loss putting more penalization on the less-represented class [29] based on MLPs. We use grid search cross-validation to select the hyperparameters $\alpha \in (0.25, 0.5, 0.75)$, $\gamma \in (2, 5)$. (3) **MAML** is a few-shot learning scheme from [13] and it focuses on improving the generalization performance for less represented data samples and we use 5-shot learning.

We also consider the Deep-SVDD [37], short as **D-SVDD**, which is a unsupervised anomaly detection approach. D-SVDD will train a MLP based neural network to represent the majority of the data. The data points falling outside the majority will be considered as outliers. We also consider **GBDT**, a tree based classifier optimizing objective function through a multiple boosting process, and it is a good fit for REM [31].

In terms of the evaluation metrics, we use the F1 score to evaluate the model's classification ability, and we set 0.5 as the decision threshold. Another adopted evaluation metric is the Precision-Recall AUC (PR-AUC), which focuses on the trade-off between precision and recall ability [20]. Besides the F1 and PR-AUC score for classification ability, we also care about the model's ability with uncertainty prediction as the model results provide support for decision-making [40]. We use the negative log likelihood (NLL), Brier score [40]. Considering the extreme imbalanced label in REM, we modify the Brier score as the Brier Skill Score (BSS) in the following:

$$\text{BSS} = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}\Big(\Pr(y^{(i)}|\mathbf{x}^{(i)}) - y^{(i)}\Big)^2}{\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{1}{n}\sum_{i}^{n} y^{(i)} - y^{(i)}\Big)^2} \tag{4.8}$$

where $\Pr(y^{(i)}|\mathbf{x}^{(i)})$ is predicted probability of label $y^{(i)}$ and $\text{BSS} \in [0, 1]$. Similar as the BSS, we modify the NLL with the relative NLL (short as R-NLL) given the imbalanced label:

$$\text{R-NLL} = \frac{\sum_{i=1}^{n}\text{NLL}\Big(\Pr(y^{(i)}|\mathbf{x}^{(i)}), y^{(i)}\Big)}{\sum_{i=1}^{n}\text{NLL}\Big(\frac{1}{n}\sum_{i}^{n} y^{(i)}, y^{(i)}\Big)} \tag{4.9}$$

where $\text{NLL}\Big(\Pr(y^{(i)}|\mathbf{x}^{(i)}), y^{(i)}\Big)$ measure the NLL between $\Pr(y^{(i)}|\mathbf{x}^{(i)})$ and label $y^{(i)}$.

5. BENCHMARK DATASET

We investigated VI-REM's generalization performance on unstructured image and natural language processing tasks and evaluated robustness. We conducted the experiments via Pytorch on an NVIDIA GeForce RTX 4080 Laptop GPU. To ensure fair comparisons and convenience of visualization, we set the number of latents to 2 in all experiments. As the two considered benchmark datasets are well established in practice, we design an ablation study to understand the functionality of each component in VI-REM. For model VI-REM-V1, we alternate the EP distribution with the Gaussian distribution. Model VI-REM-V2 further replaces the MANN component with the fully connected MLPs. Model VI-REM-V3 alternates the weight $\beta = n_0/n_1$ in Equation 3.9 with a small number. For simplicity, we set the weight $\beta = 5$ in VI-REM-V3.

5.1 Benchmark Dataset: Imbalanced Digit Recognition

We applied the VI-REM to MNIST[1] dataset. The MNIST dataset covers handwritten digit numbers 0 to 9, corresponding to class 0 to class 9 as the label. Following a similar setup in [3], in each task, one of the ten classes is set as *Event*, while the samples from the remaining nine classes represent *Normality*. In the model training part, 100 images are randomly selected from the *Event* class, and nearly 600,000 images are selected from *Normality* class. As the selection of *Event* cases involves randomness, we repeat the experiment ten times for each task. In total, we conduct 100 individual

[1] http://yann.lecun.com/exdb/mnist/

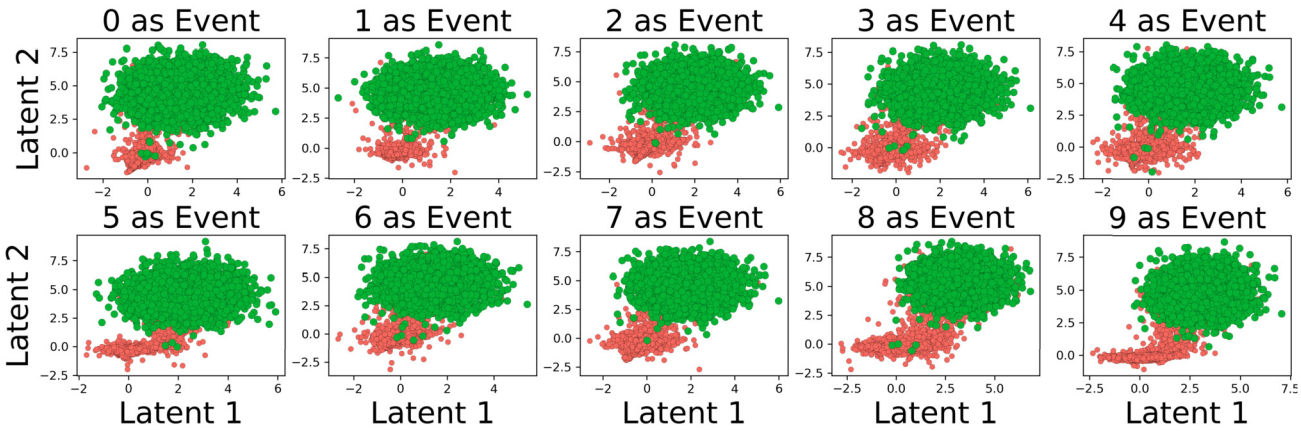


Figure 3: Learned representations via VI-REM under different tasks on original testing. Red points represent class *Event*, green points represent class *Normality*. Red points represent the class *Event*, green points represent the class *Normality*.

runs of experiments. For model testing part, we create balanced testings consist of 5,000 images from the class *Event* and 5,000 images from the class *Normality*. Regarding robustness evaluations, we applied the random rotation operator to the original testing data. Each of the original testing images will rotate at a different angle, and the rotated angle is a random value from a uniform distribution $[-60°, 60°]$.

For VI-REM, the experiment sets $500 \times 500 \times 2$ fully connected (FC) MLPs for $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ and $500 \times 500 \times 500$ for MANN in recognizing $p_{\boldsymbol{\theta}}(y|\mathbf{z})$. We used Leaky ReLU for the activation function for both $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ and $p_{\boldsymbol{\theta}}(y|\mathbf{z})$. For S-MLP, F-MLP and MAML, we set the neural network to be $500 \times 500 \times \times 2 \times 500 \times 500 \times 500 \times 2$ FC layers; for D-SVDD, we set the neural network to be $1000 \times 1000 \times 1000 \times 2$ FC layers. For the GBDT model, we set the maximum depth of GBDT to be 10 and number of estimators in GBDT to be 50.

Figure 3 shows the learned representations via VI-REM under different tasks under the original testing. It's clear that all the class *Event* deviate from the distribution and mostly lie on the tail part. This pattern is consistent with the EP prior from the VI-REM model. Table 1 reports the average of the 100 experiment results for both original testings and robustness evaluations. The proposed VI-REM model consistently outperformed the competing models regarding model accuracy and robustness. Under random rotation tasks, the VI-REM shows significantly improved model robustness compared to other SOTAs. Through ablation study, the MANN structure largely increases the model's generalization ability under original testings, while the EP distribution contributes much larger in robustness.

The dimension of latent variable is an important hyper parameter in the VI-REM. To examine how the latent variable dimension affects the performance of our model, we vary it from 2 to 3 and 5. The results are reported in Table 2, and the column "Dim" displays the dimension for the latent variables. We find that setting the dimension of the latent variable can generate satisfying results on both original testing and rotation testing.

Table 1. Digit Number Detection Overall Performance Comparison.

		R-NLL↑	BSS↑	PR-AUC↑	F1 Score↑
Original	GBDT	0.28	−0.49	0.75	0.24
	MAML	**1.32**	0.47	**0.98**	0.82
	S-MLP	0.74	0.31	**0.98**	0.78
	F-MLP	0.83	0.40	**0.98**	0.81
	D-SVDD	0.38	−0.80	0.60	0.01
	VI-REM	1.13	**0.51**	0.96	**0.86**
	VI-REM-V1	1.25	0.54	0.95	**0.86**
	VI-REM-V2	0.63	−0.22	0.78	0.45
	VI-REM-V3	0.55	−0.43	0.70	0.32
Rotation	GBDT	0.24	−0.76	0.65	0.24
	MAML	0.23	−0.82	0.69	0.12
	S-MLP	0.31	−0.50	0.77	0.33
	F-MLP	0.25	−0.70	0.78	0.22
	D-SVDD	0.38	−0.81	0.54	0.01
	VI-REM	**0.87**	**0.21**	**0.89**	**0.71**
	VI-REM-V1	0.34	−0.45	0.72	0.36
	VI-REM-V2	0.26	−0.75	0.65	0.17
	VI-REM-V3	0.23	−0.92	0.57	0.05

↑ for higher is better and ↓ for lower is better.

Table 2. Digit Number Detection Overall Performance Comparison with Different Number of Latent Variables.

	Dim	R-NLL↑	BSS↑	PR-AUC↑	F1 Score↑
Original	2	1.13	0.51	0.96	0.86
	3	0.33	0.31	0.96	0.77
	5	0.48	0.56	0.96	0.87
Rotation	2	0.87	0.21	0.89	0.71
	3	0.13	−0.42	0.77	0.38
	5	0.13	−0.35	0.77	0.43

↑ for higher is better and ↓ for lower is better.

5.2 Benchmark Dataset: Imbalanced News Detection

This experiment applied the VI-REM to the standard natural language processing (NLP) benchmark dataset *AG News Classification*.[2] The AG News dataset covers various news articles from "Sports", "World", "Business", and "Sci/Tech" these four different categories. For implementing the applications, we used pre-trained fastText [33] and SWEM [38] to extract raw text representations. For the original input text data, we firstly apply the pre-trained fastText model to transform raw word input into the word embedding, where we set the dimensionality to 300. Based on the extracted word embedding, we used the average operator from SWEM [38] to combine word embedding and generate a 300×1 vector for an individual news article. In terms of robustness evaluations, we randomly select 20% words from class *Event* and class *Normality*, and exchange them. Figure 4 gives an illustrative example of the data processing and robustness evaluations.

We created an extremely imbalanced text classification task. For the training task, we treated 30,000 "Sports" news articles as *Normality* and 100 randomly selected news articles from other categories as *Event* types. Within each task, we repeat the random samplings for *Event* types ten times to alleviate randomness. In task 1, we use the news articles from the "World" category as class *Event* while in task 2 and 3, we set "Business" and "Sci/Tech" separately as *Event* types. For out-of-sample model performance evaluation, 1,900 news articles for each category made up the testing sample. For VI-REM, we set a $500 \times 500 \times 2$ FC neural layer for $q_\phi(\mathbf{z}|\mathbf{x})$ and $500 \times 500 \times 1$ for the individual function $f_j(\mathbf{z}_j)$ of MANN in $p_\theta(y|\mathbf{z})$. For S-MLP, F-MLP and MAML, we set the neural network as $500 \times 500 \times 2 \times 500 \times 500 \times 1$ FC layers; for D-SVDD, we set a $1000 \times 1000 \times 1000 \times 2$ network.

Table 3 reports average performance under the original testings and robustness evaluations, and VI-REM shows better generalization ability and more robust model performance in dealing with data perturbations. The MANN structure largely increases the model's generalization ability through the ablation study in Table 3. We also check how the model performances will change when setting a different number of latent dimension in Table 4.

6. REAL DATA APPLICATIONS

Two real data applications are considered. First application is to detect the fraudulent credit card transactions. The second application identifies traffic crash events based on driving data.

[2] https://www.kaggle.com/amananandrai/ag-news-classification-dataset

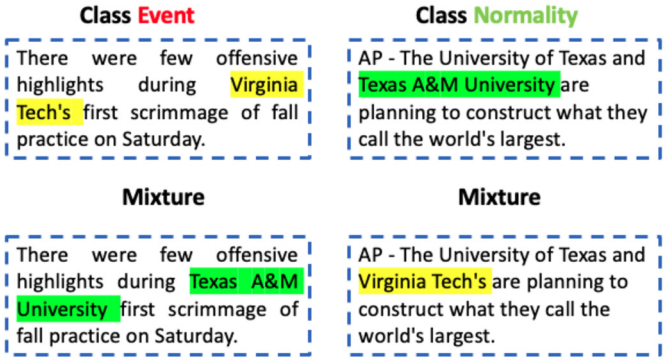


Figure 4: Illustration of Robustness Evaluations for Imbalanced News Detection.

Table 3. News Detection Model performance Comparisons.

	Method	R-NLL ↑	BSS ↑	PR-AUC↑	F1-Score ↑
Original	GBDT	0.06	−0.61	0.81	0.33
	MAML	0.60	0.44	**0.99**	0.82
	F-MLP	0.34	0.43	**0.99**	0.83
	S-MLP	0.35	0.43	0.98	0.83
	D-SVDD	0.45	−0.69	0.60	0.02
	VI-REM	**1.64**	**0.58**	0.98	**0.87**
	VI-REM-V1	1.43	0.52	0.98	0.84
	VI-REM-V2	0.68	0.20	0.95	0.63
	VI-REM-V3	0.50	−0.03	0.92	0.50
Mixture	GBDT	0.05	−0.95	0.66	0.06
	MAML	0.34	0.02	**0.97**	0.65
	F-MLP	0.11	−0.49	0.86	0.39
	S-MLP	0.12	−0.44	0.85	0.43
	D-SVDD	0.42	−0.75	0.57	0.01
	VI-REM	**0.92**	**0.24**	0.95	**0.74**
	VI-REM-V1	0.80	0.15	0.95	0.69
	VI-REM-V2	0.42	−0.18	0.87	0.47
	VI-REM-V3	0.32	−0.38	0.85	0.32

↑ for higher is better and ↓ for lower is better.

Table 4. News Detection Overall Performance Comparison with Different Number of Latent Variables.

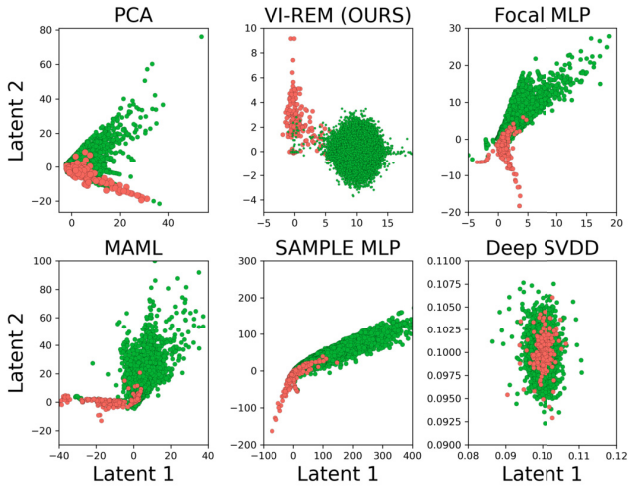
	Dim	R-NLL↑	BSS↑	PR-AUC↑	F1 Score↑
Original	2	1.64	0.58	0.98	0.87
	3	1.06	0.42	0.96	0.85
	5	0.91	0.06	0.99	0.27
Mixture	2	0.92	0.24	0.95	0.74
	3	0.65	0.04	0.91	0.70
	5	0.69	−0.17	0.95	0.19

↑ for higher is better and ↓ for lower is better.

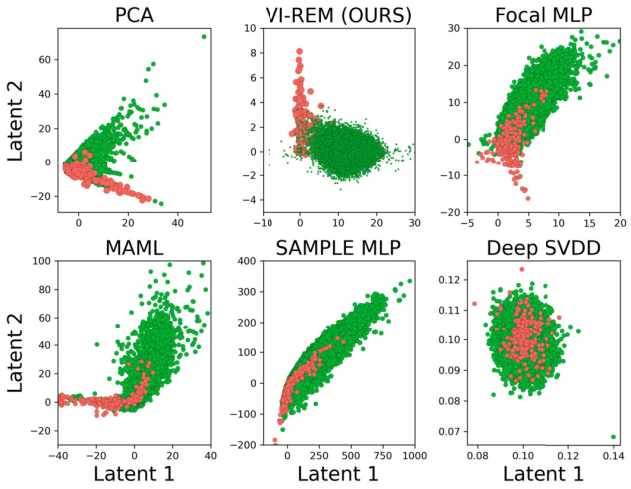
6.1 Application Case 1: Fraud Detection

We implemented the VI-REM model on the fraud detection applications[3] where fraudulent credit card transactions are the rare events of interest. The dataset to which the VI-REM was applied consists of about 284,000 financial transactions associated with 28 predictive features, among which only 492 examples are labeled as *Event* resulting in

[3] https://www.kaggle.com/mlg-ulb/creditcardfraud



(a) Learnt Representation for the Original Testing Data



(b) Learnt Representation under "Noisy 1" Testings

Figure 5: The latent representation by different model in testing Set for Fraud Detection. PCA is short as Principal Component Analysis. For VI-REM model, the latents comes from $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{z})$. The red points represent the fraud transactions, the green points represent the normal ones.

an extremely low prevalence of 0.2%. We adopted a 5:5 train test split. To further evaluate model robustness in handling massive data noise, we created an artificially noisy setting by injecting Gaussian noise $\mathcal{N}(0,1)$ to the testing data, labeled as "Noisy 1" in Table 5. Further, we impose Gaussian noise $\mathcal{N}(0,2)$ to testing data and we refer it as "Noisy 2" in Table 5. The neural network structure is the same as Imbalanced News Detection.

Figure 5a and Figure 5b display the learnt representations for the fraud detection dataset in the original testing and noisy testings. Figure 5a and Figure 5b confirm that for the VI-REM model, the learnt representation for the *Event* cases lie on the tailed part while *Normality* mostly lie in the center.

Table 5 reports results for both original and noisy settings. The proposed VI-REM model consistently outperformed the competing models in terms of accuracy and robustness. Under noisy testing tasks, which sometimes can be associated with data quality issues in real-world settings, the VI-REM shows significantly improved model robustness compared to all SOTAs. VI-REM's supreme performance on robustness to data noises indicates it can learn ubiquitous discriminative information. We notice that under "Noisy 1" testings, our model VI-REM still has reliable prediction performance while the other competitors are like a random guesses.

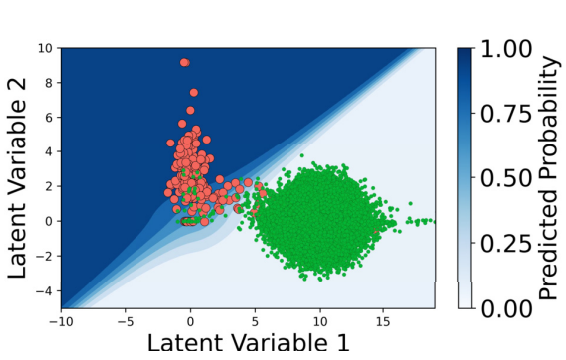
Next, we investigate the relationship between latents $\mathbf{z}$ and label y. Figure 6a plots the decision boundary and associated predicted probability for determining the *Event* cases under original testings. It is clear that the larger distance with the distribution center, the more likely it will be class *Event*. Figure 6b details the relationship between latent factors $\mathbf{z}_1$, $\mathbf{z}_2$ and label y. Latent $\mathbf{z}_1$ has a negative relationship with the occurrences of the events while $\mathbf{z}_2$ has a positive relationship.

Table 5. Fraud Detection Model Comparison.

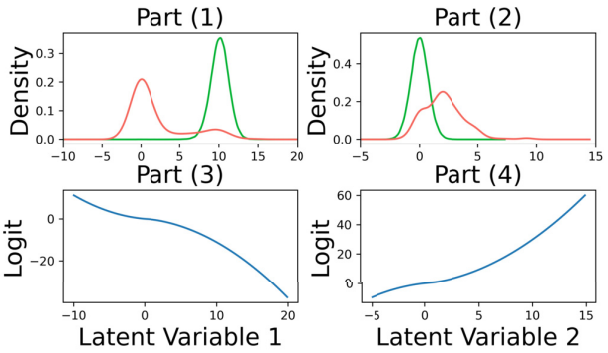
	Method	R-NLL ↑	BSS ↑	PR-AUC↑	F1-Score ↑
Original	GBDT	0.54	0.42	0.64	0.71
	MAML	1.06	−0.31	0.71	0.49
	F-MLP	2.23	0.63	0.79	0.79
	S-MLP	0.89	0.56	0.76	0.77
	D-SVDD	1.09	−0.01	0.03	0.00
	VI-REM	**3.72**	**0.68**	**0.80**	**0.81**
Noisy 1	GBDT	0.07	−5.49	0.36	0.17
	MAML	0.54	−2.16	0.67	0.26
	F-MLP	0.76	−0.61	0.68	0.44
	S-MLP	0.31	−1.56	0.46	0.32
	D-SVDD	0.99	−0.01	0.02	0.00
	VI-REM	**2.83**	**0.57**	**0.72**	**0.73**
Noisy 2	GBDT	0.02	−30.2	0.33	0.04
	MAML	0.30	−4.46	**0.53**	0.17
	F-MLP	0.28	−2.82	0.41	0.22
	S-MLP	0.13	−5.10	0.21	0.14
	D-SVDD	0.55	−0.26	0.01	0.00
	VI-REM	**1.32**	**0.16**	**0.53**	**0.56**

↑ for higher is better and ↓ for lower is better.

We identify the relationship between input features $\mathbf{x}$ and latent $\mathbf{z}$. We consider the 28 individual features and the interaction of any two of the input features. In total, $\binom{28}{1} + \binom{28}{2} = 406$ combinations are considered for model interpretations. We implement the strategy discussed in Section 4.1 where we $q = 5$ for each round of selections. Table 6 reports the major components of latent factor $\mathbf{z}_1$ and $\mathbf{z}_2$ respectively, where V_{22} represents the 22th variable of the input features. The most contributed part of $\mathbf{z}_1$ and $\mathbf{z}_2$ do not overlap, confirming the disentanglement. We also fit a



(a) Decision Boundary of VI-REM in Fraud Detection



(b) Latent Factors of the VI-REM in Fraud Detection

Figure 6: Model interpretations for VI-REM in fraud detection. (a)The decision boundary and associated predicted probability for determining the *Event* cases under original testings. (b) Latent factors plot. Part (1) and part (3) of displays the probability density functions of latent variable 1 and latent variable 2, where the red lines represent the *Event* class and the green lines represent the *Normality* class.Part (3) and (4) presents the marginal relationship of the latent factors $\mathbf{z}_1$, $\mathbf{z}_2$ with the Logit value of condition probability, that is $\log\left(\frac{\Pr(y|\mathbf{z}_1)}{1-\Pr(y|\mathbf{z}_1)}\right)$ and $\log\left(\frac{\Pr(y|\mathbf{z}_2)}{1-\Pr(y|\mathbf{z}_1)}\right)$ for $\mathbf{z}_1$ and $\mathbf{z}_2$.

Table 6. Model Interpretations for Latent Factor $\mathbf{z}_1$ and $\mathbf{z}_2$ in Application Case 1 and Application Case 2.

		Selected Variable	Coefficient
Case 1	Latent Variable $\mathbf{z}_1$	$V_{22} \times V_{26}$	−0.001
	Latent Variable $\mathbf{z}_1$	$V_{16} \times V_{24}$	0.024
	Latent Variable $\mathbf{z}_1$	$V_{16} \times V_{18}$	−0.041
	Latent Variable $\mathbf{z}_1$	$V_{13} \times V_{24}$	−0.006
	Latent Variable $\mathbf{z}_1$	$V_{16} \times V_{23}$	0.011
Case 1	Latent Variable $\mathbf{z}_2$	$V_3 \times V_{13}$	0.010
	Latent Variable $\mathbf{z}_2$	$V_7 \times V_{13}$	−0.002
	Latent Variable $\mathbf{z}_2$	$V_{13} \times V_{19}$	0.040
	Latent Variable $\mathbf{z}_2$	$V_3 \times V_{16}$	0.008
	Latent Variable $\mathbf{z}_2$	$V_{13} \times V_{20}$	0.023
Case 2	Latent Variable $\mathbf{z}_1$	acc-x$_{26}$ × acc-y$_{26}$	−0.747
	Latent Variable $\mathbf{z}_1$	acc-z$_{26}$ × acc-z$_{29}$	−2.042
	Latent Variable $\mathbf{z}_1$	acc-z$_{26}$ × acc-z$_{28}$	0.133
	Latent Variable $\mathbf{z}_1$	acc-z$_{25}$ × acc-z$_{27}$	−0.973
	Latent Variable $\mathbf{z}_1$	acc-z$_{26}$ × acc-z$_{30}$	1.896

linear regression between $\mathbf{z}_1$ and $\mathbf{z}_2$ and their selected major components, and it is reported as column "Coefficients" in Table 6. For example, the fitted coefficient between the combination of V_{16}, V_{24} and latent $\mathbf{z}_1$ is 0.024, indicating they have a positive relationship.

6.2 Application Case 2: Traffic Crash Identification

Traffic crashes are rare events, with an average rate of one crash 6.8 per million vehicle miles traveled in the US [10]. The naturalistic driving study (NDS) provides an unprecedented opportunity to evaluate crash risk [10]. NDSs are characterized by continuously recording driving information, such as three-dimensional Inertial Measurement Unit acceleration and multi-channel video recordings. Accurately identifying crashes with robustness benefits further understanding of driving behavior with the overall goal of reducing traffic accidents [10], identifying high risky segments at road network [52] or improving the validation efficiency for autonomous vehicles [34].

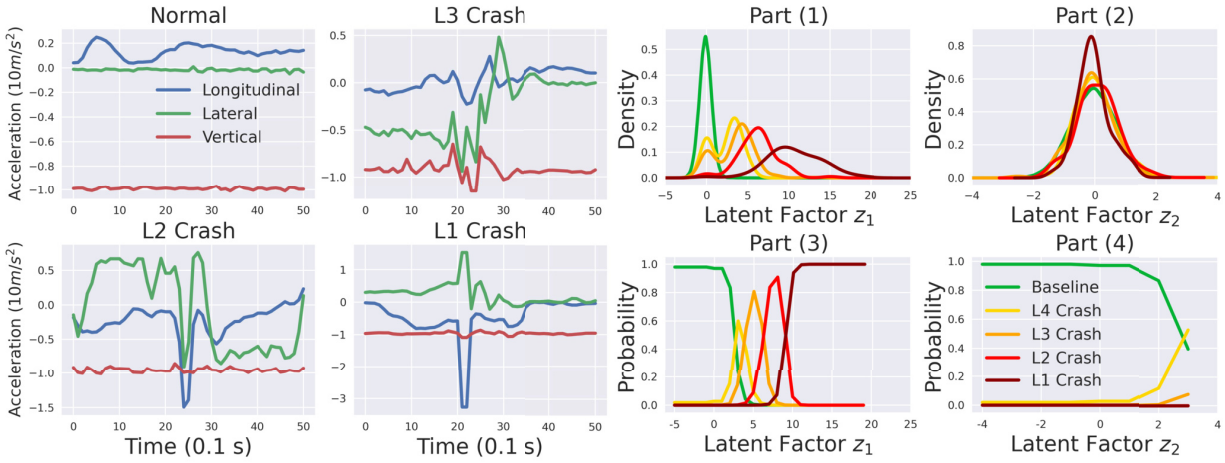
This experiment applied the VI-REM to the Second Strategic Highway Research Program (SHRP2) NDS [10], which is largest NDS to-date, with more than 1 million hours of continuous driving data. SHRP2 Insight website[4] gives a more detailed introduction. We applied the VI-REM to identify traffic crashes based on the high-frequency longitudinal, lateral and vertical tri-axial acceleration for the vehicles.

SHRP2 crash events have four different severity levels [10], ranging from level 4 to level 1. Level 4 (L4) crash events are low-risk tire strikes, level 3 (L3) crash events are minor crashes, level 2 (L2) crash events are police reportable crashes with at least 1,500 dollars worth of damage, and level 1 (L1) are fatal crash events. Our data consists of 100 level 1, 150 level 2, 578 level 3, and 588 level 4 crash events. After a random selection process, we also collected nearly 120,000 safe driving segments from SHRP2. These selected safe driving segments have similar maximum and minimum speeds with crashes. Table 7 summarizes the detail of different events and the last column calculates the imbalance ratio.

Table 7. Summary Description of Level 1 to Level 4 Crash Events and Normal Drivings.

	Description	Number	Ratio
L1 Crash	Fatal Crash	100	1200:1
L2 Crash	Police Reportable Crashes	150	800:1
L3 Crash	Minor Crash and MC	578	208:1
L4 Crash	Low-risk Tire Strikes	588	204:1
Normal Driving	Routine Safe Driving	119,996	1:1

[4] https://insight.shrp2nds.us/



(a) Demo plot of SHRP2 Driving Segments (b) Latent Factors for Model Interpretation

Figure 7: Several plots for identifying traffic crash events in SHRP2 data (a) Several driving segments are associated with the three-dimensional accelerations under different event categories. (b) Part (1) and Part (2) represent the kernel density estimation of latent $\mathbf{z}_1$ and $\mathbf{z}_2$ given different event categories. Part (3) and Part (4) display the associated probability of different event categories given different values on $\mathbf{z}_1$ and $\mathbf{z}_2$.

We applied the ordinal-based VI-REM model to the SHRP2 data, as the label of events has different severity levels. To standardize the data, we put the most volatile time in the middle of each driving segment. Each driving segment has 5.1 seconds and as the collection frequency for acceleration is 10 HZ, so the length for each segment is $5.1 \times 10 = 51$. As the acceleration has three directions, the dimensionality for the input features will be $51 \times 3 = 153$. Figure 7a shows several driving segments. Our experiment adopted a 5/5 split for training and testing, and the neural network structure for each model is the same as Imbalanced Digit Recognition.

Our work transforms the assessment of an ordinal model into assessing the performance of a classification model based on different selections on the severity level. For example:

$$\begin{aligned} &\text{PR-AUC}\left(\widehat{\Pr}(Y \geq j), \mathcal{I}(Y \geq j)\right) \\ &\text{and} \quad j \in [\text{L1, L2, L3, L4}] \end{aligned} \tag{6.1}$$

where $Y = (y_1, \ldots, y_N)$ and $\widehat{\Pr}(Y \geq j)$ is the predicted probability from model in Equation 4.7. Similar transformations will also be applied to the other metrics. Table 8 compares the performances of different models for identifying different severity traffic crashes, where we skip the Deep-SVDD model as it is difficult to extend to ordinal data conditions. Table 8 shows VI-REM's supreme performances over other competitive models, especially for severe level 1, level 2 and level 3 crash events. Overall for identifying crash events (level 4 up to level 1), our proposed VI-REM can have 0.77 in F1 score, while the most competitive F1 score from benchmark model is 0.68 by Sampling MLP.

Analytical modeling for identifying traffic crashes is sensitive to kinematic noise (such as signal connection issues or different driving environments), where model performance can largely deteriorate as noise increases [30]. Considering this, where model robustness is fundamentally important for decision making, we created an artificially noisy setting where Gaussian noise $\mathcal{N}(0, 0.05)$ was injected to the input accelerations. The model performance metrics are presented in Table 9. As the table shows, the proposed VI-REM significantly outperforms alternative models based on the evaluation metrics. Compared with benchmarks, the VI-REM shows much more robust disincentive power under noisy testing conditions.

As for the model interpretations, Figure 7b shows the mutual relationship between $\mathbf{z}$ and label y. $\mathbf{z}_1$ has larger discriminative over different crash events, and larger values in $\mathbf{z}_1$ are associated with more considerable severity. We implement the same strategy in application case 1 to identify the relationship between input features $\mathbf{x}$ and latent $\mathbf{z}$. We report the identified major components of $\mathbf{z}_1$ in Table 6, as latent $\mathbf{z}_2$ does not have strong distinguishing ability. In Table 6, acc-x_{26} represents the longitudinal acceleration at point 26, and acc-y, acc-z represent the lateral and vertical acceleration respectively. All the major contributing input features are around the middle of the driving segment through the model interpretations. This phenomenon is consistent with our experiment setup. We also find that vertical accelerations could be the most influential factor for identifying traffic crashes and determining their associated severity level. This observation is consistent with various empirical studies [30]

Table 8. Performance Comparisons in Traffic Crash Identifications.

	Method	R-NLL ↑	BSS ↑	PR-AUC↑	F1-Score ↑
VI-REM	Y ≥ **L4**	1.58	**0.58**	**0.70**	**0.77**
	Y ≥ **L3**	**2.50**	**0.52**	**0.69**	**0.68**
	Y ≥ **L2**	**3.32**	**0.59**	**0.77**	**0.74**
	Y ≥ **L1**	**2.71**	**0.56**	**0.78**	**0.71**
GBDT	Y ≥ **L4**	0.65	−0.23	0.49	0.52
	Y ≥ **L3**	0.48	−0.86	0.27	0.31
	Y ≥ **L2**	0.62	−0.61	0.38	0.36
	Y ≥ **L1**	0.94	−0.21	0.50	0.42
MAML	Y ≥ **L4**	**1.91**	0.42	0.64	0.56
	Y ≥ **L3**	2.16	0.44	0.67	0.64
	Y ≥ **L2**	1.67	0.40	0.65	0.61
	Y ≥ **L1**	1.42	0.30	0.53	0.49
S-MLP	Y ≥ **L4**	1.43	0.26	0.57	0.65
	Y ≥ **L3**	1.36	0.08	0.63	0.66
	Y ≥ **L2**	1.68	0.24	0.54	0.54
	Y ≥ **L1**	1.44	0.09	0.57	0.51
F-MLP	Y ≥ **L4**	0.71	−0.01	0.70	0.68
	Y ≥ **L3**	1.25	0.40	0.65	0.60
	Y ≥ **L2**	1.15	0.55	0.72	0.68
	Y ≥ **L1**	0.66	0.49	0.62	0.58

↑ for higher is better and ↓ for lower is better.

Table 9. Robustness Evaluations in Traffic Crash Identifications.

	Method	R-NLL ↑	BSS ↑	PR-AUC↑	F1-Score ↑
VI-REM	Y ≥ **L4**	1.48	**0.48**	**0.69**	**0.73**
	Y ≥ **L3**	**2.43**	**0.47**	**0.68**	**0.66**
	Y ≥ **L2**	**3.28**	**0.59**	**0.77**	**0.74**
	Y ≥ **L1**	**2.57**	**0.57**	**0.78**	**0.72**
GBDT	Y ≥ **L4**	0.38	−1.09	0.42	0.38
	Y ≥ **L3**	0.38	−1.42	0.24	0.23
	Y ≥ **L2**	0.38	−0.84	0.31	0.31
	Y ≥ **L1**	0.67	−0.64	0.43	0.38
MAML	Y ≥ **L1**	**1.85**	0.40	0.58	0.56
	Y ≥ **L2**	1.99	0.39	0.64	0.62
	Y ≥ **L3**	1.60	0.38	0.61	0.60
	Y ≥ **L4**	1.43	0.30	0.52	0.49
S-MLP	Y ≥ **L4**	0.87	0.05	0.58	0.61
	Y ≥ **L3**	0.66	−0.65	0.62	0.63
	Y ≥ **L2**	0.87	−0.42	0.45	0.48
	Y ≥ **L1**	0.63	−1.34	0.45	0.29
F-MLP	Y ≥ **L4**	0.55	−0.53	0.64	0.60
	Y ≥ **L3**	1.00	0.27	0.62	0.60
	Y ≥ **L2**	1.14	0.55	0.71	0.68
	Y ≥ **L1**	0.65	0.48	0.61	0.61

↑ for higher is better and ↓ for lower is better.

6.3 Discussion on the Running Time

Before concluding our paper, we present the average running time of our method VI-REM and other competitive methods across different tasks in Table 10. Column "CV" represents the task for imbalanced digit recognition in Section 5.1, column "NLP" represents the task for imbalanced detection in Section 5.2, and column "Fraud" represents the task for fraud detection in Section 6.1, and "Crash" represents the task in Section 6.2. Our experiments show that MAML is the most robust among the competitive methods. In terms of efficiency, VI-REM is faster than MAML, especially under the MLP and CV task. We also find that most of the running time for the VI-REM model comes from the MANN part. Designing a more efficient regularized neural networks could be an interesting topic to explore in the future. The GBDT method performs fast under CV and Fraud task, but may experience longer iteration (with an average of 2113 seconds) to converge when dealing with complex text data. Under the OLM setup in Section 6.2, the VI-REM is even faster than S-MLP.

Table 10. Overall Computation Time across Different Methods on Different Tasks. The unit is $\times 100$ Second.

Model	CV	NLP	Fraud	Crash
GBDT	1.05	21.13	1.01	6.44
MAML	4.82	1.53	2.78	3.79
F-MLP	0.30	0.11	0.35	0.32
S-MLP	0.51	0.14	0.32	3.53
D-SVDD	0.25	0.19	0.32	-
VI-REM	3.17	0.80	2.40	3.04
VI-REM-V1	2.95	0.77	2.10	2.89
VI-REM-V2	0.28	0.20	0.16	0.27

7. SUMMARY AND DISCUSSION

This study proposes a Variational Inference toward extreme approach for rare event modeling (short as VI-REM). Our model induces uncertainty to representation learning through a novel EP distribution and addresses the overfitting issue through MANN, which leads to a more robust model for rare events. We provide theoretical properties on the efficacy of the proposed model. Extensive empirical experiments are conducted to confirm the model performance over various application scenarios with promising results, especially with respect to model generalization, robustness and interpretability.

The VI-REM framework greatly improves the generalizability and interpretability of rare event modeling, two challenging issues associated with limited numbers of events. The application of VI-REM could accurately depict the risk associated with rare-events and allow the general public and decision-makers to set realistic expectations for rare events. The identification of adverse events at an early stage may allow mitigation of the damage and loss associated with the events. The features identified through the model are crucial for researchers and practitioners to identify the causes of rare events and take proper countermeasures to prevent and reduce the occurrence of future adverse events.

APPENDIX A. PROOF OF THEOREM 3.2

Proof. As the label y is binary variable, we set a cross entropy loss function, and we can specify $\mathrm{L}_f(\mathbf{z}, y)$ as:

$$\mathrm{L}_f(\mathbf{z}, y) = y \log f(\mathbf{z}) + (1-y)\log(1-f(\mathbf{z})) \qquad \text{(A.1)}$$

So as when $y = 1$ or when $y = 0$, we can have for $\mathbf{z}$, $\mathbf{z}'$

$$\begin{aligned} |\mathrm{L}_f(\mathbf{z}, y) - \mathrm{L}_f(\mathbf{z}', y)\| &\le |\log f(\mathbf{z}) - \log f(\mathbf{z}')| \\ &= \left| \log\left(1 + \left(\frac{f(\mathbf{z})}{f(\mathbf{z}')} - 1\right)\right)\right| \\ &\le \frac{|f(\mathbf{z}) - f(\mathbf{z}')|}{\min\big(f(\mathbf{z}), f(\mathbf{z}')\big)} \end{aligned} \tag{A.2}$$

Based on condition C.1, the function $\mathrm{L}_f(\mathbf{z}, y)$ follows:

$$\|\mathrm{L}_f(\mathbf{z}, y) - \mathrm{L}_f(\mathbf{z}', y)\| \le k_n' \|\mathbf{z} - \mathbf{z}'\| \tag{A.3}$$

where k_n' is some constant $k_n' = \frac{k_n}{\min\left(f(\mathbf{z}), f(\mathbf{z}')\right)}$

Decompose generalization bound with triangle inequality:

$$\begin{aligned} &\mathbb{E}\left(\sup_{f\in\mathcal{F}} \left| \mathrm{P}_n\big(\mathrm{L}_f(\mathbf{z}, y)\big) - \Pr\big(\mathrm{L}_{f_0}(\mathbf{z}, y)\big)\right|\right) \\ &\quad\le \mathbb{E}\left(\sup_{f\in\mathcal{F}} \left| \mathrm{P}_n\big(\mathrm{L}_f(\mathbf{z}, y)\big) - \Pr\big(\mathrm{L}_f(\mathbf{z}, y)\big)\right|\right) \\ &\quad+ \mathbb{E}\left(\sup_{f\in\mathcal{F}} \left| \Pr\big(\mathrm{L}_f(\mathbf{z}, y)\big) - \Pr\big(\mathrm{L}_{f_0}(\mathbf{z}, y)\big)\right|\right) \end{aligned} \tag{A.4}$$

where we know that f_0 minimize of the loss function form assumption C.2, then we can have the upper bound for:

$$\begin{aligned} &\mathbb{E}\left(\sup_{f\in\mathcal{F}} \left| \Pr\big(\mathrm{L}_f(\mathbf{z}, y)\big) - \Pr\big(\mathrm{L}_{f_0}(\mathbf{z}, y)\big)\right|\right) \\ &\quad\le \mathbb{E}\left(\Big(\Pr - \mathrm{P}_n\Big)\Big(\sup_{f\in\mathcal{F}} \big|\mathrm{L}_f(\mathbf{z}, y) - \mathrm{L}_f(\mathbf{z}, y)\big|\Big)\right) \end{aligned} \tag{A.5}$$

where we denote this upper bound to be

$$\mathbb{E}\Big(\mathrm{G}(\mathbf{z})\Big) = \mathbb{E}\Big((\Pr - \mathrm{P}_n) \sup_{f\in\mathcal{F}} \big|\mathrm{L}_f(\mathbf{z}, y) - \mathrm{L}_f(\mathbf{z}, y)\big|\Big) \tag{A.6}$$

Thus the generalization gap can be transferred as:

$$\begin{aligned} &\mathbb{E}\left(\sup_{f\in\mathcal{F}} \left| \mathrm{P}_n\big(\mathrm{L}_f(\mathbf{z}, y)\big) - \Pr\big(\mathrm{L}_{f_0}(\mathbf{z}, y)\big)\right|\right) \\ &\le \mathbb{E}\Big(\sup_{f\in\mathcal{F}} \left| \mathrm{P}_n\big(\mathrm{L}_f(\mathbf{z}, y)\big) - \Pr\big(\mathrm{L}_f(\mathbf{z}, y)\big)\right|\Big) + \mathbb{E}\Big(\mathrm{G}(\mathbf{z})\Big) \end{aligned} \tag{A.7}$$

And the function $\mathrm{G}(\mathbf{z})$ can be decomposed as:

$$\mathrm{G}(\mathbf{z}) = (\Pr - \mathrm{P}_n) \sup_{f\in\mathcal{F}} \big|\mathrm{L}_f(\mathbf{z}, y) - \mathrm{L}_f(\mathbf{z}, y)\big| \times (I_n^+ + I_n^-) \tag{A.8}$$

where I_n^- as the empirical value of $I_{\mathbf{z}\in\Omega_\eta : \|\mathbf{z}\| < \eta}$ and I_n^+ as the empirical value of $I_{\mathbf{z}\in\Omega_\eta^C : \|\mathbf{z}\| \ge \eta}$. It is obvious $I_n^+ + I_n^- = 1$. Thus the expectation generalization bound can be formed as:

$$\begin{aligned} \mathbb{E}\Big(\mathrm{G}(\mathbf{z})\Big) &= \mathbb{E}\Big((\Pr - \mathrm{P}_n) \sup_{f\in\mathcal{F}} \big|\mathrm{L}_f(\mathbf{z}, y) - \mathrm{L}_f(\mathbf{z}, y)\big|\} \times I_n^+\Big) \\ &\quad+ \mathbb{E}\Big((\Pr - \mathrm{P}_n) \sup_{f\in\mathcal{F}} \big|\mathrm{L}_f(\mathbf{z}, y) - \mathrm{L}_f(\mathbf{z}, y)\big|\} \times I_n^-\Big) \\ &= \mathbb{E}\Big(\mathrm{G}^+(\mathbf{z})\Big) + \mathbb{E}\Big(\mathrm{G}^-(\mathbf{z})\Big) \end{aligned} \tag{A.9}$$

From Lemma 2 and Lemma 3 in [12] and Equation A.3, we can have the following bound, where $j \in [+, -]$:

$$\mathbb{E}\Big(\mathrm{G}^j(\mathbf{z})\Big) \le 4k_n' \mathbb{E}\left(\sup_{f\in\mathcal{F}} \Big|\mathrm{P}_n \epsilon\big(f(\mathbf{z}) - f_0(\mathbf{z})\big) I_n^j\Big|\right) \tag{A.10}$$

where k_n' is some positive constant satisfying Lipschitz property, ϵ is the Rademacher sequence.

Similar to the Lemma 5 proof in [12], Cauchy–Schwarz inequality can bound it further by:

$$\begin{aligned} &4k_n' \mathbb{E}\left(\sup_{f\in\mathcal{F}} \Big|\mathrm{P}_n \epsilon\big(f(\mathbf{z}) - f_0(\mathbf{z})\big) I_n^j\Big|\right) \\ &\quad\le 4k_n' \mathbb{E}\left(\big(\mathrm{P}_n \epsilon I_n^j\big)^{\frac{1}{2}} \sup_{f\in\mathcal{F}} |f(\mathbf{z}) - f_0(\mathbf{z})|^{\frac{1}{2}}\right) \end{aligned} \tag{A.11}$$

Based on the property in Rademacher sequence, we can have the following bound where c is some positive constant:

$$\mathbb{E}\big(\mathrm{P}_n \epsilon\big) \le c^2/n \tag{A.12}$$

Denote $I_n^+ = h$ and $I_n^- = 1 - h$. Follows the assumption C.3, we can have the bound toward $\mathbb{E}\big(\mathrm{G}^-(\mathbf{z})\big)$:

$$\mathbb{E}\Big(\mathrm{G}^-(\mathbf{z})\Big) \le 4k_n' c n^{-\frac{1}{2}} \|\mathbf{z}\|^{1/2} (1-h)^{1/2} \sup_{f\in\mathcal{F}} \|f - f_0\|_{\mathbf{z}\in\Omega_\eta^C}^{1/2} \tag{A.13}$$

Similarly, we can have the bound toward $\mathbb{E}\big(\mathrm{G}^+(\mathbf{z})\big)$:

$$\mathbb{E}\Big(\mathrm{G}^+(\mathbf{z})\Big) \le 4k_n' c n^{-\frac{1}{2}} h^{1/2} \eta^{1/2} \|f - f_0\|_{\mathbf{z}\in\Omega_\eta}^{1/2} \tag{A.14}$$

For simplicity, we set $\gamma = 4k_n' c n^{-1/2}$ and we can have:

$$\begin{aligned} \mathbb{E}\Big(\mathrm{G}(\mathbf{z})\Big) &\le \gamma h^{1/2} \eta^{1/2} \sup_{f\in\mathcal{F}} \|f - f_0\|_{\mathbf{z}\in\Omega_\eta}^{1/2} \\ &\quad+ \gamma \sup_{\mathbf{z}\in\Omega_\eta^C} \|\mathbf{z}\|^{1/2} (1-h)^{1/2} \sup_{f\in\mathcal{F}} \|f - f_0\|_{\mathbf{z}\in\Omega_\eta^C}^{1/2} \end{aligned} \tag{A.15}$$

And similar to the classification generalization bound:

$$\begin{aligned} &\mathbb{E}\left(\sup_{f\in\mathcal{F}} \left| \mathrm{P}_n\big(\mathrm{L}_f(\mathbf{z}, y)\big) - \Pr\big(\mathrm{L}_f(\mathbf{z}, y)\big)\right|\right) \\ &\quad\le \sqrt{\frac{\log|\mathcal{F}| + \log(1/\delta)}{2n}} \end{aligned} \tag{A.16}$$

where $\delta = |\mathcal{F}|e^{-2n\phi^2}$ where ϕ is some constant. Combine Equation A.15 and Equation A.16 based on Equation A.7, we can have Theorem 3.2. □

APPENDIX B. PROOF OF THE REMARK 1

Proof. To simplify the proof, we denote the generalization gap for function f as $\mathcal{E}_f$. Then, the difference of the generalization bound by f_1, f_2 can be expressed as:

$$\mathcal{E}_{f_1} - \mathcal{E}_{f_2} \simeq \Delta_2 - \Delta_1 \tag{B.1}$$

where $\Delta_2 = \gamma h^{1/2}\eta^{1/2}\Delta_\Omega$. and

$$\Delta_1 = \max_{\mathbf{z}\in\Omega_\eta^C}\Big(\sup\|\mathbf{z}\|^{1/2}, \sup\|\mathbf{z}'\|^{1/2}\Big)\gamma(1-h)^{1/2}\Delta_{\Omega_\eta^C}$$

As $\sup_{\mathbf{z}\in\Omega_\eta^C}\|\mathbf{z}\|^{1/2}$ and $\sup_{\mathbf{z}'\in\Omega_\eta^C}\|\mathbf{z}'\|^{1/2}$ are at the same magnitude, we assume the following condition:

$$\max_{\mathbf{z}\in\Omega_\eta^C}\Big(\sup\|\mathbf{z}\|^{1/2}, \sup\|\mathbf{z}'\|^{1/2}\Big) = r\sup_{\mathbf{z}\in\Omega_\eta^C}\|\mathbf{z}\|^{1/2} \tag{B.2}$$

where r is some constant. So that:

$$\mathcal{E}_{f_1} - \mathcal{E}_{f_2} \simeq \gamma h^{1/2}\eta^{1/2}\Delta_\Omega - r\sup_{\mathbf{z}\in\Omega_\eta^C}\|\mathbf{z}\|^{1/2}\gamma(1-h)^{1/2}\Delta_{\Omega_\eta^C} \tag{B.3}$$

Intuitively, as long as $\Delta_2 < \Delta_1$, the overall model performance will be improved. As long as the Equation 3.18 satisfies, we can have the statement in Remark 1. □

APPENDIX C. PROOF OF THEOREM 4.1

Proof. For each round of the selections for model interpretation as in Equation 4.4, we have the following from[27]:

$$\begin{aligned}\Pr\{\mathcal{D}_{\mathbf{z}}^* \subseteq \widehat{\mathcal{D}}_g\} \geq 1 - \mathcal{O}\Big\{&d\exp[-c_1 s^{1-2(u+v)}] \\ &+ d\times s\exp(-c_2 s^{\gamma})\Big\}\end{aligned} \tag{C.1}$$

where s is the sample size group g, d is the dimension of input features, and c_1, c_2, u, v are some constant and $u+v < 1/2$.

Thus:

$$\begin{aligned}\Pr\Big(\mathcal{D}_{\mathbf{z_j}}^* \subseteq \widehat{\mathcal{D}}_{\mathbf{z_j}}\Big) &= \Pr\{\mathcal{D}_{\mathbf{z}}^* \subseteq \widehat{\mathcal{D}}_1\}\times\cdots\times\Pr\{\mathcal{D}_{\mathbf{z}}^* \subseteq \widehat{\mathcal{D}}_G\} \\ &\geq \Big\{1 - \mathcal{O}\big(d\exp[-c_1 s^{1-2(u+v)}] \\ &\quad + d\times s\exp(-c_2 s^{\gamma})\big)\Big\}^G\end{aligned} \tag{C.2}$$

Replacing $s = n^{-\alpha}$ where $\alpha > 0$, then we have Theorem 4.1. □

ACKNOWLEDGEMENTS

This study is partially funded by the Safety through Disruption (Safe-D) National University Transportation Center, a grant from the U.S. Department of Transportation – Office of the Assistant Secretary for Research and Technology, University Transportation Centers Program.

Accepted 31 July 2024

REFERENCES

[1] Alemi, A. A., Fischer, I., Dillon, J. V. and Murphy, K. (2016). Deep variational information bottleneck. *arXiv:1612.00410*.

[2] Bengio, Y., Courville, A. and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**(8) 1798–1828.

[3] Byrd, J. and Lipton, Z. C. (2018). What is the effect of importance weighting in deep learning? *arXiv:1812.03372*.

[4] Cao, K., Wei, C., Gaidon, A., Arechiga, N. and Ma, T. (2019). Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems* 1565–1576.

[5] Carlini, N. and Wagner, D. (2017). Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy* 39–57. IEEE.

[6] Casale, F. P., Dalca, A., Saglietti, L., Listgarten, J. and Fusi, N. (2018). Gaussian process prior variational autoencoders. *Advances in neural information processing systems* **31**.

[7] Chen, R. T., Li, X., Grosse, R. B. and Duvenaud, D. K. (2018). Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems* **31**.

[8] Chen, Y., Zhou, X. S. and Huang, T. S. (2001). One-class SVM for learning in image retrieval. In *Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205)* **1** 34–37. IEEE.

[9] Coles, S., Bawa, J., Trenner, L. and Dorazio, P. (2001). *An introduction to statistical modeling of extreme values* **208**. Springer. https://doi.org/10.1007/978-1-4471-3675-0. MR1932132

[10] Dingus, T. A., Guo, F., Lee, S., Antin, J. F., Perez, M., Buchanan-King, M. and Hankey, J. (2016). Driver crash risk factors and prevalence evaluation using naturalistic driving data. *Proceedings of the National Academy of Sciences* **113**(10) 2636–2641.

[11] Dong, Q., Gong, S. and Zhu, X. (2018). Imbalanced deep learning by minority class incremental rectification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **41**(6) 1367–1381.

[12] Fan, J., Song, R. et al. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics* **38**(6) 3567–3604. https://doi.org/10.1214/10-AOS798. MR2766861

[13] Finn, C., Abbeel, P. and Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning* 1126–1135. PMLR.

[14] Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* 27–38. https://doi.org/10.1093/biomet/80.1.27. MR1225212

[15] Fortuin, V. (2022). Priors in bayesian deep learning: A review. *International Statistical Review* **90**(3) 563–591. https://doi.org/10.1111/insr.12502. MR4524825

[16] Guo, F. (2019). Statistical methods for naturalistic driving studies. *Annual Review of Statistics and its Application* **6** 309–328. https://doi.org/10.1146/annurev-statistics-030718-105153. MR3939523

[17] Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H. and Bing, G. (2017). Learning from class-imbalanced data:

Review of methods and applications. *Expert Systems with Applications* **73** 220–239.
[18] Han, Q., Wang, T., Chatterjee, S., Samworth, R. J. et al. (2019). Isotonic regression in general dimensions. *The Annals of Statistics* **47**(5) 2440–2471. https://doi.org/10.1214/18-AOS1753. MR3988762
[19] Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models* **43**. CRC press. MR1082147
[20] He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* **21**(9) 1263–1284.
[21] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S. and Lerchner, A. (2017). Beta-VAE: learning basic visual concepts with a constrained variational framework. *The International Conference on Learning Representations* **2**(5) 6.
[22] Jin, D., Jin, Z., Zhou, J. T. and Szolovits, P. (2020). Is Bert really robust? A strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI Conference on Artificial Intelligence* **34** 8018–8025.
[23] King, G. and Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis* **9**(2) 137–163.
[24] Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv:1312.6114*.
[25] Kingma, D. P., Welling, M. et al. (2019). An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning* **12**(4) 307–392.
[26] Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I. and Welling, M. (2016). Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems* **29**.
[27] Li, R., Zhong, W. and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association* **107**(499) 1129–1139. https://doi.org/10.1080/01621459.2012.695654. MR3010900
[28] Li, X., Li, R., Xia, Z. and Xu, C. (2020). Distributed feature screening via componentwise debiasing. *Journal of Machine Learning Research* **21**. MR4071207
[29] Lin, T. -Y., Goyal, P., Girshick, R., He, K. and Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision* 2980–2988.
[30] Liu, L., Racz, D., Vaillancourt, K., Michelman, J., Barnes, M., Mellem, S., Eastham, P., Green, B., Armstrong, C., Bal, R. et al. (2023). Smartphone-based hard-braking event detection at scale for road safety services. *Transportation research part C: emerging technologies* **146** 103949.
[31] Lusa, L. et al. (2017). Gradient boosting for high-dimensional prediction of rare events. *Computational Statistics & Data Analysis* **113** 19–37. https://doi.org/10.1016/j.csda.2016.07.016. MR3662388
[32] Massoli, F. V., Falchi, F., Kantarci, A., Akti, Ş., Ekenel, H. K. and Amato, G. (2021). MOCCA: Multilayer one-class classification for anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*. https://doi.org/10.1109/tnnls.2021.3130074. MR4442308
[33] Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C. and Joulin, A. (2017). Advances in pre-training distributed word representations. *arXiv:1712.09405*.
[34] O'Kelly, M., Sinha, A., Namkoong, H., Tedrake, R. and Duchi, J. C. (2018). Scalable end-to-end autonomous vehicle testing via rare-event simulation. *Advances in Neural Information Processing Systems* **31** 9827–9838.
[35] Ren, M., Zeng, W., Yang, B. and Urtasun, R. (2018). Learning to reweight examples for robust deep learning. *arXiv:1803.09050*.
[36] Rezende, D. and Mohamed, S. (2015). Variational inference with normalizing flows. In *International conference on machine learning* 1530–1538. PMLR.
[37] Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S. A., Binder, A., Müller, E. and Kloft, M. (2018). Deep one-class classification. In *International Conference on Machine Learning* 4393–4402.
[38] Shen, D., Wang, G., Wang, W., Min, M. R., Su, Q., Zhang, Y., Li, C., Henao, R. and Carin, L. (2018). Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms. *arXiv:1805.09843*.
[39] Shi, L., Qian, C. and Guo, F. (2022). Real-time driving risk assessment using deep learning with XGBoost. *Accident Analysis & Prevention* **178** 106836.
[40] Snoek, J., Ovadia, Y., Fertig, E., Lakshminarayanan, B., Nowozin, S., Sculley, D., Dillon, J., Ren, J. and Nado, Z. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems* 13969–13980.
[41] Sohn, K., Lee, H. and Yan, X. (2015). Learning structured output representation using deep conditional generative models. *Advances in Neural Information Processing Systems* **28**.
[42] Sønderby, C. K., Raiko, T., Maaløe, L., Sønderby, S. K. and Winther, O. (2016). Ladder variational autoencoders. *Advances in neural information processing systems* **29**.
[43] Sur, P. and Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences* **116**(29) 14516–14525. https://doi.org/10.1073/pnas.1810420116. MR3984492
[44] Székely, G. J., Rizzo, M. L., Bakirov, N. K. et al. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics* **35**(6) 2769–2794. https://doi.org/10.1214/009053607000000505. MR2382665
[45] Tomczak, J. and Welling, M. (2018). VAE with a VampPrior. In *International Conference on Artificial Intelligence and Statistics* 1214–1223. PMLR.
[46] Wang, H. (2020). Logistic regression for massive data with rare events. In *International Conference on Machine Learning* 9829–9836. PMLR.
[47] Wehenkel, A. and Louppe, G. (2019). Unconstrained monotonic neural networks. *Advances in Neural Information Processing Systems* **32**.
[48] Wu, T., Liu, Z., Huang, Q., Wang, Y. and Lin, D. (2021). Adversarial robustness under long-tailed distribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 8659–8668.
[49] Zhang, F., Gao, C. et al. (2020). Convergence rates of variational posterior distributions. *Annals of Statistics* **48**(4) 2180–2207. https://doi.org/10.1214/19-AOS1883. MR4134791
[50] Zhang, Q., Yu, S., Xin, J. and Chen, B. (2022). Multi-view information bottleneck without variational approximation. In *ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing* 4318–4322. IEEE.
[51] Zhang, W. E., Sheng, Q. Z., Alhazmi, A. and Li, C. (2020). Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology* **11**(3) 1–41.
[52] Zheng, L., Sayed, T. and Mannering, F. (2021). Modeling traffic conflicts for use in road safety analysis: A review of analytic methods and future directions. *Analytic Methods in Accident Research* **29** 100142.

Chen Qian. Department of Statistics, Virginia Tech, U.S.A.
E-mail address: chenqian@vt.edu

Chenyang Tao. Amazon, U.S.A.
E-mail address: chenyang.tao@duke.edu

Jingbin Xu. Department of Statistics, Virginia Tech, U.S.A.
E-mail address: jingbin@vt.edu

Feng Guo. Department of Statistics, Virginia Tech and Virginia Tech Transportation Institute, U.S.A.
E-mail address: feng.guo@vt.edu