PREDICTION IN THE PRESENCE OF RESPONSE-DEPENDENT MISSING LABELS

Hyebin Song[1], *Garvesh Raskutti*[2], *Rebecca Willett*[3]

[1] Department of Statistics, The Pennsylvania State University,
[2]Department of Statistics, University of Wisconsin-Madison,
[3]Department of Statistics, University of Chicago

ABSTRACT

In various settings, limitations of sensing technologies or other sampling mechanisms result in missing labels, where the likelihood of a missing label is an unknown function of the data. For example, satellites used to detect forest fires cannot sense fires below a certain size threshold. In such cases, training datasets consist of positive and pseudo-negative observations (true negatives or undetected positives with small magnitudes). We develop a new methodology and non-convex algorithm which jointly estimates the magnitude and occurrence of events, utilizing prior knowledge of the detection mechanism. We provide conditions under which our model is identifiable. We prove that even though our approach leads to a non-convex objective, any local minimizer has an optimal statistical error (up to a log term) and the projected gradient descent algorithm has geometric convergence rates. We demonstrate on both synthetic data and a California wildfire dataset that our method outperforms existing state-of-the-art approaches.

1. INTRODUCTION

A common challenge in many statistical machine learning problems is *noisy or missing labels.* In such settings, it is often common to assume the labels are missing at random and place a distribution on the missing labels (see e.g. [10, 15]). However, in many applications, labels are missing systematically due to aspects of the technology in the data collection process. Consider, for example, a dataset consisting of wildfire events in California where fire size is measured using satellite imagery. Due to the limited resolution of the satellite optics, fires smaller than a certain threshold will not be observed, complicating the effort of building a predictor of fire size. Similarly, consider forecasting the spread or impact of a virus, where a person's likelihood of being tested and included in a dataset depends on the severity of their symptoms. These are both examples of *response-dependent missing labels* where labels or measurements are missing based on the *magnitude or size* of the measured event. This response-dependent sampling bias poses a significant challenge in terms of (i) predicting event (such as fire) occurrence, since small magnitude events are not recorded and (ii) predicting the magnitude of each event (due to positive bias of the measurements).

In this paper, we develop a statistical framework that addresses *response-dependent missing labels* with a two-level model that (i) models the *true event magnitude* Y as a mixture of 0, indicating no event, and a positive distribution if the event occurs; and (ii) models the *observed event magnitude* Z, which is either the same as Y or 0, depending on the true response Y. More specifically,

$$P(Z = 0|Y = y > 0, \mathbf{X} = \mathbf{x}) = 1 - \Gamma(y),$$

where $\mathbf{X} = \mathbf{x}$ denotes the features or covariates and $\Gamma(y)$ represents a probability depending on y which accounts for the outcome-dependence. Hence $Z = 0$ could either denote a "true" negative where $Y = 0$ or a "false negative" where $Y > 0$ but $Z = 0$.

This flexible framework allows us to model response-dependent missing labels through an occurrence-magnitude mixture distribution for Y and the probability function $\Gamma(y)$ for the observed response Z. This model presents both identifiability and computational challenges that we address in this paper. Since $Z = 0$ could either denote a true 0 or a false 0, we first provide identifiablity conditions on our mixed model. Secondly, two computational challenges arise: (i) the likelihood of the observed data Z involves integration over the function $\Gamma(y)$ and (ii) even if this integration is possible, the objective is non-convex. To address (i), we choose $\Gamma(y)$ to be the CDF of a Gamma distribution which allows a closed-form computation of the integral; to address (ii), we demonstrate that even though the objective is non-convex, using projected gradient descent leads to a local minimizer with desirable statistical properties.

Related Work Our proposed model is in contrast with the Type I Tobit model [21], where excess zeros arise due to the censoring of an underlying continuous variable. Zeros are only proxies for values below a certain threshold, and thus the goal of Tobit analysis is to estimate magnitude only. On the contrary, our framework models a two-part mixture that separately models the probability of event occurrences and magnitude of the events [19, 16].

Our approach is also related to Positive-Unlabeled (PU) learning [11, 6, 5] and non-ignorable missing data (see e.g. [18, 10]), as the responses in our setting consist of positive and pseudo-negative observations, where pseudo-negative observations arise from response-dependent missing labels. However, PU learning focuses exclusively on the occurrence of events (labels), while our framework involves a mixture distribution of Y that simultaneously estimates occurrence and magnitude. Another related literature is missing not at random (MNAR) mechanisms (see e.g. [22**?** , 8, 13]). In the MNAR setting, which observations are missing is known a priori while in our setting true and false negatives are unknown a priori.

Lastly, an active line of work exists in non-convex estimation problems in which various statistical and algorithmic guarantees for a non-convex M-estimator are studied [12**?** , 14, 7]. Our objective turns out to be a non-convex function of parameters, and our work utilizes a number of tools in the non-convex literature to obtain statistical and algorithmic guarantees of the proposed estimator which is a stationary point of the non-convex objective function.

Contributions Our paper makes the following contributions: 1. a general statistical framework for dealing with response-dependent missing labels, leading to a closed-form log-likelihood; 2. identifiability conditions (Theorem 1) for our model; 3. provably optimal statistical error (up to a log term) and efficient algorithm (Theorem 2 and 3); and 4. illustration of the advantages of our approach using simulated data and real data analysis of wildfire prediction in California.

2. MODEL AND ALGORITHM

2.1. Problem Set-up

We consider the following problem set-up for estimation and prediction using contaminated data. We assume that Y has a mixture distribution of a point mass at 0 (denoting no event) and continuous distribution over $\mathbb{R}_+$ (denoting the magnitude of the event), and each component distribution depends on the value of a set of features $\mathbf{x} \in \mathbb{R}^p$. In other words, the pdf of Y given $X = \mathbf{x}$ is as follows[1]:

$$p_Y(t|\mathbf{x};\beta,\theta) = (1 - p_1(\mathbf{x}))\delta_0(t) + p_1(\mathbf{x})g(t|\mathbf{x}) \qquad (1)$$

for some $p_1(\mathbf{x})$ and $g(\cdot|\mathbf{x})$ where p_1 takes a value between 0 and 1 depending on $\mathbf{x}$ and $g(t|\mathbf{x})$ is a pdf of the continuous distribution. Here, each p_1 and g is related to occurrence and magnitude of the mixture distribution for Y.

First, we model $\mathbb{P}(Y > 0|\mathbf{x}) = p_1(\mathbf{x};\theta)$ and $\mathbb{P}(Y = 0|\mathbf{x}) = 1 - p_1(\mathbf{x};\theta)$, where we let $p_1(\mathbf{x};\theta) := \sigma(\mathbf{x}^\top\theta) := (1 + \exp(-\mathbf{x}^\top\theta))^{-1}$. When $Y > 0$, we use an exponential GLM; specifically,

$$p_{Y|Y>0,\mathbf{X}}(y|\mathbf{x}) = g(t|\mathbf{x};\beta) := \lambda_X \exp(-\lambda_X t)$$

where $\lambda_X = \exp(-\mathbf{x}^\top\beta)$. Here, each exponentiated coefficient represents the multiplicative effect of the corresponding feature. The exponential GLM is chosen to reflect that a size of an event is always non-negative. That is, given that an event has occurred, i.e. $Y > 0$, the probability that Y is larger than t is $\mathbb{P}(Y > t|Y > 0, \mathbf{x}) = \int_t^\infty g(s|\mathbf{x};\beta)ds$.

If an i.i.d sample of $(\mathbf{x}_i, y_i)_{i=1}^n$ is available, the mixture modeling approach (e.g. [4, 17]) can be utilized to estimate the parameters θ and β. However, in our setting, not all y_i are observed since events with small magnitude tend to have missing labels. We introduce a random variable Z to denote the *observed* size of an event. If an event has occurred but is unobserved, then $y_i > 0$ but $z_i = 0$. On the other hand, if the event is observed, the recorded size is the same as the true size, i.e. $z_i = y_i$. Since $z_i = 0$ no longer implies that no event has occurred, we cannot simply estimate the parameters using the observed sizes (z_is) instead of the true sizes (y_is).

2.2. Likelihood model and identifiability

We model the likelihood of correctly observing events as

$$\mathbb{P}(Z > 0|Y = y > 0, \mathbf{X} = \mathbf{x}) = \Gamma(y). \qquad (2)$$

In other words, the probability that the magnitude Y is observed depends only on the value of Y itself. In many practical applications, this "self-masking phenomenon" occurs where true value itself determines whether the observation would be hidden or revealed. For example, if we consider fire prediction, the size of fire affects whether the fire event would be detected or not; hence $\Gamma(\cdot)$ is a monotonically increasing function. From here, we combine (1) and (2) and integrate out the unobserved Y to derive $p_{Z|\mathbf{X}}$; the log of this quantity forms our loss function for a collection of samples $(\mathbf{x}_i, z_i)$ for $i = 1, \ldots, n$:

$$\begin{aligned}\mathcal{L}_n(\theta,\beta) = &-\frac{1}{n}\sum_{i;z_i=0}\log\left(1 - \phi(\mathbf{x}_i;\beta,\Gamma)p_1(\mathbf{x}_i;\theta)\right) \\ &-\frac{1}{n}\sum_{i;z_i>0}\log\left\{g(z_i|\mathbf{x};\beta)\Gamma(z_i)p_1(\mathbf{x}_i;\theta)\right\}\end{aligned} \qquad (3)$$

where

$$\phi(\mathbf{x};\beta,\Gamma) = \int_0^\infty \Gamma(y)g(y|\mathbf{x};\beta)dy. \qquad (4)$$

Identifiability. The model is not identifiable if no assumptions about the structure of g in (1) and Γ are made because the likelihood (3) is defined via $\Gamma(y)g(y|\mathbf{x})$. On the other hand, both parameters are identifiable under parametric assumptions on p_1 and g for any given positive Γ, if two parameter vectors are distinct, i.e., $\beta \neq c\theta$ for any $c \neq 0$, and the feature vector $\mathbf{x}$ spans all directions in $\mathbb{R}^p$. More concretely, We have the following result about the identifiability of the model (3) under the following Assumption **A1**:

A1. *Two parameter vectors β and θ in* (1) *are linearly independent. The density of $\mathbb{P}_X$ with respect to the Lebesgue measure is positive everywhere.*

Theorem 1. *For any given positive Γ and under Assumption **A1** , the parameters (β, θ) in the model* (3) *are identifiable.*

The proof is based on constructing a set of observations $(\mathbf{x}_i, z_i)$ that distinguish the likelihoods evaluated at different parameter values, and is deferred to the full version of this paper [20].

Choice of $\Gamma(\cdot)$. The next question is how to choose the label observation probability $\Gamma(y)$. One of the determining factors is that the integral in (4) needs to be computable and $\Gamma(y)$ also needs to be monotonically increasing. If ϕ does not have an analytical form, approximation of the function via a numerical integration is needed, which can be computationally challenging. Hence we choose Γ to be the cumulative distribution function of an exponential function with parameter λ_ϵ. This choice of Γ has several advantages: first, by choosing Γ as a cdf, Γ is a monotonically increasing function in y, which is in accordance with our setting where events with small sizes are more likely to be not included in the dataset. Also, this choice is computationally attractive since ϕ in (4) can be analytically solved.

Our estimation method is defined as the maximizer of the log-likelihood (3) with the choice of $\Gamma(y) := 1 - \exp(-\lambda_\epsilon y)$ and ϕ in (4). We use the name PU-OMM to refer to our method, which stands for *Positive-Unlabeled Occurrence Magnitude Mixture*.

2.3. Algorithm

Given data $(\mathbf{x}_i, z_i)$ for $i = 1, \ldots, n$, the objective function is

$$\widehat{\omega} \in \operatorname*{argmin}_{\omega \in \mathbb{B}_2(r)} \mathcal{L}_n(\omega) := -\frac{1}{n}\sum_{i=1}^n \ell(\omega; (\mathbf{x}_i, z_i)), \qquad (5)$$

where $\omega := (\beta, \theta)$, and $\ell(\omega; (\mathbf{x}_i, z_i))$ is the ith component of the likelihood in (3). We also define the population risk function $\mathcal{R}(\omega) := \mathbb{E}[\mathcal{L}_n(\omega)]$ and define $\omega_0 := (\beta_0, \theta_0)$ as the minimizer of $\mathcal{R}(\omega)$. We let the search space $\mathbb{B}_2(r)$ be an ℓ_2 ball with a radius r, for a sufficiently large $r > 0$ so that ω_0 is feasible.

To optimize (5), we propose to use the standard projected gradient descent (projected to $\mathbb{B}_2(r)$). We will show in Theorem 2 and 3 that it is feasible to obtain $\widehat{\omega}$ in (5) despite $\mathcal{L}_n(\omega)$ being non-convex, and the convergence of iterates $\{\omega^t\}_{t\geq 1}$ in Algorithm 1 is linear given a sufficiently large sample size.

[1] By pdf, we mean a Radon-Nikodym derivative of $P_{Y|\mathbf{X}}$ with respect to the Lebesgue measure plus a point mass at zero.

Algorithm 1: Projected Gradient Descent

```
Input: Data $(\mathbf{x}_i, z_i)_{i=1}^n$, step size $\eta$, initial point $\omega^0$,
       hyperparameter $\lambda_\epsilon$, search space radius $r$
for $t = 1, 2, 3, \ldots$ do
    $\omega^{t+1} = \mathcal{P}_{\mathbb{B}_2(r)}(\omega^t - \eta\nabla\mathcal{L}_n(\omega^t))$;
    if converged then
        STOP
    end
end
```

3. THEORETICAL GUARANTEES

Throughout this section, we assume that $\Gamma(t) = 1 - \exp(-\lambda_\epsilon t)$ is given. We first introduce a set of conditions for the response variable, feature vector, and the degree of missingness, under which we prove algorithmic and statistical convergence.

A2. *(Random design) A random feature vector $\mathbf{x} \in \mathbb{R}^p$ with distribution $\mathbb{P}_X$ is mean-zero sub-Gaussian with parameter K_X for a positive constant $K_X < \infty$. In other words, for any fixed unit vector $v \in \mathbb{R}^p$, we have $\mathbb{E}[\exp(\mathbf{x}^\top v)^2/K_X^2] \le 2$. Moreover, there exists $C_\lambda > 0$ such that $\lambda_{\min}(\mathbb{E}[\mathbf{x}\mathbf{x}^\top]) \ge C_\lambda$.*

A3. *(Boundedness) There exist constants $C_X, C_Y < \infty$ such that for the random feature $\mathbf{x} \in \mathbb{R}^p$ and response variable $y \sim p_Y(\cdot|\mathbf{x};(\beta_0, \omega_0))$, $\|\mathbf{x}\|_2 \le C_X$ and $|y/e^{\mathbf{x}^\top\beta_0}| \le C_Y$ a.s.*

Assumption 2 is a mild assumption on the feature vector $\mathbf{x}$ which states that $\mathbf{x}$ has a light probability tail and the smallest eigenvalue of the population covariance matrix is lower-bounded by a positive constant. The boundedness condition is assumed mainly for the technical convenience and states that both $\mathbf{x}$ and the deviation of y from its mean are absolutely bounded, where we recall that $\mathbb{E}[Y|Y > 0, \mathbf{x}] = e^{\mathbf{x}^\top\beta_0}$.

A4. *We assume the following condition holds:*

$$\max_{1\le i\le n} \frac{1 - \sigma(\mathbf{x}_i^\top\beta + \log\lambda_\epsilon)}{1 - \sigma(\mathbf{x}_i^\top\theta)} \le r_0(\omega_0, C_X, r) \tag{6}$$

where r_0 is a constant depending on model parameters $\omega_0 = (\beta_0, \theta_0)$, C_X, and r.

We give the full expression for $r_0(\omega_0, C_X, r)$ in the proof of Theorem 2 which can be found in the full version of this paper for ease of exposition [20]. We recall that $\lambda_\epsilon = \infty$ corresponds to "no missingness" where all y_i are the same as z_i since $\Gamma(y) = 1, \forall y$. The equation (6) trivially holds in this case. Assumption **A4** essentially states that albeit $\lambda_\epsilon < \infty$, λ_ϵ is sufficiently large so that (6) holds. Assumption **A4** ensures there exists sufficient signal in the data to estimate both parameters β and θ.

The following two theorems provide algorithmic and statistical error bounds.

Theorem 2. *Under Assumptions **A1,-A4**, if $n \ge Cp\log p$, the empirical risk function $\mathcal{L}_n(\omega)$ admits a unique local minimizer in $\mathbb{B}_2(r)$ which coincides with the global optimizer $\widehat{\omega}$. In addition, for any $\delta > 0$, the following inequality holds with probability $1 - \delta$,*

$$\|\widehat{\omega} - \omega_0\|_2 \le \frac{C}{\alpha}\sqrt{\frac{C_Y^2 p\log(n)\log(C_Y/\delta)}{n}} \tag{7}$$

where $\alpha, C, C_Y > 0$ are constants only depending on model parameters (but not on n, p).

Theorem 3. *Assume **A1-A4** hold. If $n \ge Cp\log p$, for any initialization $\omega^0 \in \mathbb{B}_2(r/2)$,*

$$\|\omega^t - \widehat{\omega}\|_2 \le C_1\kappa^t\|\omega^0 - \widehat{\omega}\|_2 \tag{8}$$

for $\kappa < 1$, where $C, C_1 > 0$ are constants depending on model parameters (but not on n, p).

The convergence rate in (7) nearly matches the parametric rate of $\sqrt{p/n}$. Also, running Algorithm 1 efficiently finds the optimum of (5), in the sense that $O(\log(1/\epsilon))$ iterations are needed to find a point within distance ϵ of the global optimum $\widehat{\omega}$ of the objective (5).

Extension to the high-dimensional setting: It is worth noting that the theory we develop here has a direct generalization to the high-dimensional setting where $p \gg n$ and we assume ω_0 is s-sparse, for $s \ll p$. In the proof of Theorem 2, we show that the population risk function $\mathcal{R}(\omega)$, despite of being non-convex, admits a unique stationary point, and utilizing such result and uniform convergence we derive the ℓ_2 error bound. A similar approach can be used to obtain a statistical error bound of an ℓ_1-penalized M-estimator $\widehat{\omega}(\lambda)$, defined as $\widehat{\omega}(\lambda) := \operatorname{argmin} \mathcal{L}_n(\omega) + \lambda\|\omega\|_1$, where we control the difference between $\mathcal{L}_n(\omega)$ and $\mathcal{R}(\omega)$ over a restricted cone including $\mathbb{B}_2(r)$ (see, for instance, [14]).

4. SIMULATION STUDY

We now study the performance of the proposed method and compare with other state-of-the-art approaches in terms of parameter estimation accuracy and prediction using simulated datasets representing a number of scenarios. In particular, we consider the following three settings for generating simulated datasets where in the first setting our model is correctly specified and in the others, different mis-specifications are introduced:

1. **Correct specification:** the size of an event $Y|(Y > 0, \mathbf{x})$ is generated from the exponential distribution with parameter $\lambda_X = \exp(-\mathbf{x}^\top\beta_0)$. Missing in y_is are probabilistic, whose probabilities depend on y_i via $\Gamma(y) = 1 - \exp(-\lambda_\epsilon y)$ for $\lambda_\epsilon = .24$
2. **Misspecification 1:** g is log-Normal instead of exponential, i.e. $Y|(Y > 0, \mathbf{x}) \sim \text{LogNormal}(\mathbf{x}_i^\top\beta_0, I_p)$.
3. **Misspecification 2:** missing in y_i is deterministic and y_i below a certain threshold is recorded to be zero, i.e. $z_i = \mathbb{1}\{y_i \ge \tau\}$ for a threshold $\tau = 3$.

Data Generation. We generate a design matrix $\mathbf{X}$ by drawing each row from a multivariate Gaussian distribution $\mathcal{N}(0, \Sigma)$ where $\Sigma_{ij} = 0.2^{|i-j|}$. The true unobserved responses y_i are sampled from a mixture of zero and a continuous distribution, where zeros are sampled from a Bernoulli distribution with probabilities $\sigma(\mathbf{x}_i^\top\theta_0)$ and continuous responses are sampled from $g(\cdot|\mathbf{x}_i;\beta_0)$. Depending on the setting, g is set to be Exponential (Settings 1 and 3) or Lognormal (Setting 2). Additionally, a binary $r_i \in \{0, 1\}$ is sampled to determine whether each y_i is missing or not. Depending on the setting, we let $\mathbb{P}(r_i = 1|y) = \Gamma(y) = 1 - e^{-\lambda_\epsilon y}$ for $\lambda_\epsilon = .24$ (Settings 1 and 2) or $\Gamma(y) = \mathbb{1}\{y_i \ge \tau\}$ for $\tau = 3$ (Setting 3).

Methods.

1. Oracle: two GLMs (Logistic, Exponential) using $(\mathbf{x}_i, y_i)_{i=1}^n$ where y_i with fully labelled responses.
2. Proposed method (PU-OMM): our proposed method.
3. Logistic-Gamma mixture model (Logistic-Gamma): we fit two separate GLMs using $(\mathbf{x}_i, z_i)_{i=1}^n$, one for the occurrence and the other for the size of the event using logistic and Gamma distributions

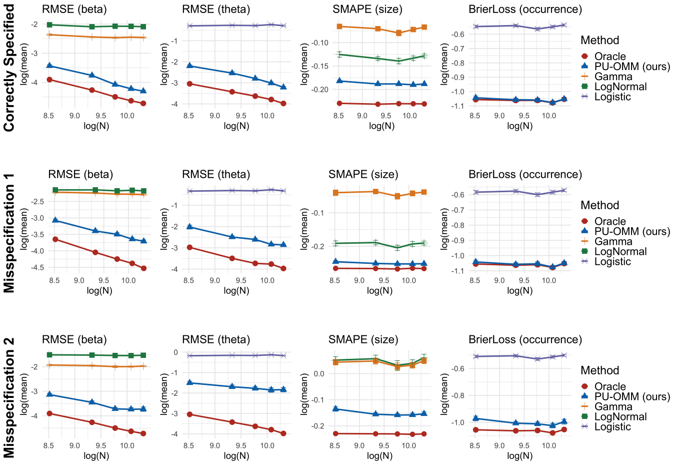

Fig. 1: Parametric estimation and prediction accuracy for each method under Settings 1-3 (Correctly Specified, Misspecification 1, and Misspecification 2). Each row i corresponds to the Setting i, for $i = 1, 2, 3$. For each row, first two panels (RMSE (beta), RMSE (theta)) show parameter estimation accuracy results, and the last two panels (SMAPE (size), BrierLoss (occurrence)) plot the accuracy of each method in predicting the true size and occurrence of each observation in test datasets. Average values from $B = 50$ trials are plotted, together with error bars corresponding to one standard error. Note for the Logistic-Gamma and Logistic-LogNormal Mixture models, the logistic model is used to predict the occurrence of events, and Gamma/LogNormal model is used to predict the magnitudes of events. Therefore, our PU-OMM model is compared with the Gamma/LogNormal models in RMSE (beta) and SMAPE (size) panels, and PU-OMM is compared with the Logistic model in RMSE (theta) and BrierLoss (occurrence) panels.

4. Logistic-LogNormal mixture model (Logistic-LogNormal): Gamma distribution is replaced with log-normal distribution in 3.

Evaluation Metrics. We compute Root Mean Squared Errors (RMSE) for each estimated $(\widehat{\beta}, \widehat{\theta})$ to evaluate parameter estimation accuracy. For prediction accuracy, we evaluate the prediction accuracy of each model in terms of predicting both occurrence and size of the *true* events. For predicting the occurrence of an event, we use BrierLoss$(\mathbf{u}, \widehat{\mathbf{p}})$:= $\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (\widehat{p}_i - u_i)^2$ which is a normalized ℓ_2 loss as a metric. For predicting the magnitude of an event, we use Symmetric Mean Absolute Percentage Error (SMAPE), Mean Absolute Deviation (MAD), and root mean squared error (RMSE) for evaluation metrics. SMAPE is considered to evaluate prediction performance in a relative scale, as results of MAD and RMSE can be affected by a few observations with large errors [2].

Results. Figure 1 presents estimation and prediction accuracy for each method under Settings 1-3. We plot results using SMAPE and BrierLoss in Figure 1 for prediction evaluation and defer the remaining plots to the Supplementary Material in the full version of the paper [20]. Unsurprisingly, the oracle estimator performs the best. Among non-oracle methods, the proposed method appears to perform the best in both correctly specified and misspecified settings, even when the hyperparameter λ_ϵ is chosen based on the data. In fact, the difference between the two PU-OMM models–one based on the true λ_ϵ value and the other based on the choice from data–was quite small. We also include a comparison plot between the two PU-OMM models in the full version of this paper [20].

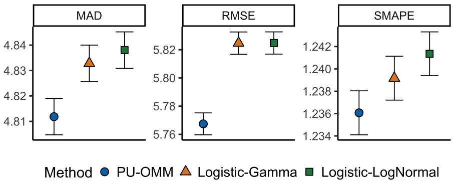

Fig. 2: Prediction performance comparison for PU-OMM, Logistic-Gamma, and Logistic-LogNormal models with the California Wildfire dataset. Average MAD, RMSE, and SMAPE values are plotted for each method. Error bars represent 1 standard error.

5. CALIFORNIA WILDFIRE DATA

California Wildfire Dataset. We use a global wildfire dataset from [1] to obtain observed fire events in California from 2001 to 2018. The database [1] includes fire events–sets of burnt areas that are connected by touching or intersecting–together with fire perimeters and the final dates of the fire events. We obtain fire sizes by computing areas of fire events based on fire perimeters. In the obtained dataset, most of the fires whose sizes are below 1km^2 are not present. Given the lack of small fires in the database, we additionally sampled points from places with no observed fires. We augmented the fire events dataset from [1] by adding these pseudo-negative points where the fire sizes corresponding to these points are set to be zero. We also incorporated information on meteorological, topographical, geographical aspects of each sampled location as covariates[9, 3]. The final dataset has dimensions $(n, p) = (15846, 43)$.

Results. All of the models are trained based on a training dataset and tested on the remaining hold-out set. For each $b = 1, \ldots, B = 100$, we randomly split the dataset into 90/10 subsamples and assigned 90% of the subsamples to a training dataset and the remaining 10% of the subsamples to a testing dataset. Unlike the simulated study, true y_i are unavailable, and thus validation needs to be based on the observed z_i. We compute predicted $\widehat{z}_i$ using fitted models. In particular, MAD$(\mathbf{z}, \widehat{\mathbf{z}})$, RMSE$(\mathbf{z}, \widehat{\mathbf{z}})$, and SMAPE$(\mathbf{z}, \widehat{\mathbf{z}})$ are computed based on the observed z_i and predicted $\widehat{z}_i$. Figure 2 plots computed MAD, RMSE, and SMAPE from various models from B trials. It appears that the proposed PU-OMM method performs the best, followed by Logistic-Gamma, and then followed by Logistic-LogNormal model.

6. DISCUSSION AND CONCLUSION

In this paper, we developed a statistical framework PU-OMM which addresses occurrence and magnitude prediction when we have response-dependent missing labels. We prove that our approach achieves optimal statistical error up to a log factor, even though the likelihood loss is non-convex. Moreover, we also showed that our projected gradient descent algorithm achieves linear convergence to a stationary point of the objective. Further, we demonstrate the benefits of our method compared to existing methods on a California wildfire dataset.

Our flexible framework can be generalized to other response-dependent missing labels settings where the missing mechanism is a stochastic function of the response values but with different models of the occurrence-magnitude mixture response. This extra flexibility comes with statistical and algorithmic challenges such as computing the integral required for the log-likelihood and providing guarantees for the non-convex objective. Adapting this framework to other missing label settings remains an open challenge.

References

[1] Tomàs Artés, Duarte Oom, Daniele de Rigo, Tracy Houston Durrant, Pieralberto Maianti, Giorgio Libertà, and Jesús San-Miguel-Ayanz. A global wildfire dataset for the analysis of fire regimes and fire behaviour. *Sci Data*, 6(1):296, November 2019.

[2] Chao Chen, Jamie Twycross, and Jonathan M Garibaldi. A new accuracy measure based on bounded relative error for time series forecasting. *PLoS One*, 12(3):e0174202, March 2017.

[3] CIESIN. Gridded population of the world, version 4 (gpwv4): Population density, revision 11, 2017.

[4] John G Cragg. Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, 39(5):829–844, 1971.

[5] Marthinus Du Plessis, Gang Niu, and Masashi Sugiyama. Convex formulation for learning from positive and unlabeled data. In *International Conference on Machine Learning*, pages 1386–1394, June 2015.

[6] Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 213–220, New York, NY, USA, 2008. ACM.

[7] A Elsener and S van de Geer. Sharp oracle inequalities for stationary points of nonconvex penalized M-Estimators. *IEEE Trans. Inf. Theory*, 65(3):1452–1472, March 2019.

[8] Alexander M Franks, Edoardo M Airoldi, and Donald B Rubin. Nonstandard conditionally specified models for nonignorable missing data. *Proc. Natl. Acad. Sci. U. S. A.*, 117(32):19045–19053, August 2020.

[9] Andy Jarvis, Hannes I Reuter, Andy Nelson, Edward Guevara, and Others. Hole-filled SRTM for the globe version 4, available from the CGIAR-CSI SRTM 90m database, 2008.

[10] Roderick J A Little and Donald B Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, April 2019.

[11] B Liu, Y Dai, X Li, W S Lee, and P S Yu. Building text classifiers using positive and unlabeled examples. In *Third IEEE International Conference on Data Mining*, pages 179–186, November 2003.

[12] Po-Ling Loh and Martin J Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Ann. Stat.*, 40(3):1637–1664, June 2012.

[13] Wei Ma and George H Chen. Missing not at random in matrix completion: The effectiveness of estimating missingness probabilities under a low nuclear norm assumption. In *Advances in Neural Information Processing Systems 32*, pages 14900–14909. Curran Associates, Inc., 2019.

[14] Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for nonconvex losses. *Ann. Stat.*, 46(6A):2747–2774, December 2018.

[15] Geert Molenberghs, Garrett Fitzmaurice, Michael G Kenward, Anastasios Tsiatis, and Geert Verbeke. *Handbook of missing data methodology*. CRC Press, 2014.

[16] Brian Neelon, A James O'Malley, and Valerie A Smith. Modeling zero-modified count and semicontinuous data in health services research part 1: background and overview. *Stat. Med.*, 35(27):5070–5093, November 2016.

[17] Maren K Olsen and Joseph L Schafer. A Two-Part Random-Effects model for semicontinuous longitudinal data. *J. Am. Stat. Assoc.*, 96(454):730–745, June 2001.

[18] Donald B Rubin. Characterizing the estimation of parameters in Incomplete-Data problems. *J. Am. Stat. Assoc.*, 69(346): 467–474, June 1974.

[19] Valerie A Smith, John S Preisser, Brian Neelon, and Matthew L Maciejewski. A marginalized two-part model for semicontinuous data. *Stat. Med.*, 33(28):4891–4903, December 2014.

[20] Hyebin Song, Garvesh Raskutti, and Rebecca Willett. Prediction in the presence of response-dependent missing labels. *ArXiv e-prints*, March 2021.

[21] James Tobin. Estimation of relationships for limited dependent variables. *Econometrica*, 26(1):24–36, 1958.

[22] Jiwei Zhao and Jun Shao. Semiparametric Pseudo-Likelihoods in generalized linear models with nonignorable missing data. *J. Am. Stat. Assoc.*, 110(512):1577–1590, October 2015.