ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10447272

A GRAPH-PREDICTION-BASED APPROACH FOR DEBIASING UNDERREPORTED DATA

Hanyang Jiang, Yao Xie

Georgia Institute of Technology
H. Milton Stewart School of Industrial and Systems Engineering
Atlanta, GA, USA

ABSTRACT

We present a novel Graph-based debiasing Algorithm for Underreported Data (GRAUD) aiming at an efficient joint estimation of event counts and discovery probabilities across spatial or graphical structures. This innovative method provides a solution to problems seen in fields such as policing data and COVID-19 data analysis. Our approach avoids the need for strong priors typically associated with Bayesian frameworks. By leveraging the graph structures on unknown variables n and p, our method debiases the under-report data and estimates the discovery probability at the same time. We validate the effectiveness of our method through simulation experiments and illustrate its practicality in one real-world application: police 911 calls-to-service data.

Index Terms— Graph signal separation, Data debiasing, Alternating minimization

1. INTRODUCTION

Bias in data collection is a prevalent issue in many real-world applications due to a variety of reasons. One common scenario is under-reporting, as elucidated by [1]. For example, police 911 calls-to-service reports, as [2] illustrates, potentially omitting a significant number of unrecorded incidents. Similarly, during the COVID-19 pandemic, data collection, as indicated by [3], only accounted for those individuals who tested positive, which overlooked asymptomatic individuals and those who hadn't undergone testing, further perpetuating data bias.

The challenge in addressing underreporting data is that a substantial identifiability issue exists. For instance, while the count of observed cases, denoted by y, is known, the count of unobserved instances is not uniquely determined. This is due to the fact that there are infinitely many solutions for the equation $y = np$ when only y is known.

The problem of estimating the probability p in a binomial Bin(n, p) distribution when the number of trials n is known has been thoroughly addressed in the classic statistical literature. However, the circumstance where both n and p are unknown is much harder and more interesting. This gives rise to the binomial n problem [4]. In the realm of statistics, this problem is a well-known issue regarding one-dimensional cases. Traditionally, the approach to resolving it involves utilizing Bayesian methodologies, as detailed in works like [5, 6, 7, 8, 9]. However, this leads to the secondary challenge of selecting an appropriate prior.

Following the setting mentioned before, where there are n possible events, each being observed with a probability p. We are then confronted with a count, y, which is modeled as a Binomial random variable, Binomial(n, p). Our primary focus is on estimating the parameters n and p while relying exclusively on the observations of y. A significant complication arises from the fact that the expected value of y is np. Although we mentioned the identifiability issue, we may utilize additional information to help circumvent the identifiability issue. This paper presents a graph-prediction-based approach for debiasing underreported data called GRAUD.

1.1. Related work

The early studies on simultaneously estimating the parameters n and p were spearheaded by Whitaker [10], Fisher [11], and Haldane [12]. They introduced the Method of Moments Estimators (MMEs) and Maximum Likelihood Estimates (MLEs). While Fisher argued that an adequately extensive dataset would make n discernible, this becomes impractical for smaller p values as the required dataset would be excessively large.

Recently, DasGupta and Rubin [4] introduced two innovative, more efficient estimators. The first is a novel moment estimator that utilizes the sample maximum, mean, and variance, while the second introduces a bias correction for the sample maximum. These estimators have shown superior performance in various scenarios, and their asymptotic properties have been thoroughly studied.

Several prior works have also considered the binomial n problem from a Bayesian viewpoint. For example, Draper and Guttman [7] proposed a Bayes point estimate that presumes a discrete uniform distribution for n over a set $1, 2, \ldots, N$. Other researchers have proposed Bayes estimators based on various prior distributions for n [13, 14, 15]. While Bayesian approaches have successfully mitigated some difficulties associated with classical approaches, they lack grounding in

asymptotic theory, thus better suiting "small" practical problems.

In the scenario where n is to be estimated with p known, Feldman and Fox [8] have provided estimates based on MLE, MVUE, and MME and explored their asymptotic properties.

Despite these efforts, the binomial n problem remains fundamentally challenging when p is unknown. The problem is characterized by intrinsic instability, and both n and p parameters have been proven not to be unbiasedly estimable [4], resulting in difficulties in obtaining reliable estimates. The most common issue across estimators is the severe underestimation of n, particularly when n is large or p is small. Without replication, drawing inferences about n becomes impossible.

2. PROBLEM SETTING

Consider the following scenario. Let's assume that we are able to observe a collection of counts at the vertices of a graph consisting of M nodes with index set V and edge set E. At every individual node, the observed number of incidents is denoted as y_i, where $i \in V$. We further assume that at each node, the probability of observing an incident is p_i, and the true number of incidents, though unknown, is n_i. This problem can be expressed as a binomial model in the form of:

$$y_i \sim \text{Binomial}(n_i, p_i) \tag{1}$$

and the expected value of y_i is $\mathbb{E}[y_i] = n_i p_i$. Our target is to jointly estimate the set p_i, n_i given the observed data y_i and the graph structure, which can be denoted by the adjacency matrix A. It's worth noting that there is an identifiability issue [4] associated with this problem, and hence, we must impose additional structure and regularization to make this problem meaningful and solvable.

In numerous practical applications, such as the analysis of policing data, spatial information forms an inherent part of the data [16]. Consider a scenario where we partition a state into various regions and represent each region by a node, where each node corresponds to a count y_i. The aim is to recover the true number of incidents n_i and the discovery probability p_i in each region. Inspired by this setup, we put forth two reasonable assumptions to address the identifiability issue inherent to this problem.

Our first assumption is that the discovery probabilities p_i are spatially smooth, which means that the probability across neighboring regions should not vary significantly. The graph Laplacian quadratic form [17] is often used to represent such smoothness, we posit that the quantity $p^T L p$ should be small. Here, $p = [p_1, \ldots, p_M]^T$ represents the vector of discovery probabilities, and $L = D - A$ signifies the graph Laplacian, where D stands for the degree matrix. Based on the equation

$$p^T L p = \frac{1}{2} \sum_{(i,j) \in E} (p_i - p_j)^2, \tag{2}$$

a small value of $p^T L p$ indicates that the absolute difference $|p_i - p_j|$ is small for all edges $(i, j) \in E$. This aligns with our assumption that the discovery probability remains fairly uniform across adjacent regions.

Our second assumption is that the true counts of incidents, represented as n_i, are determined by an underlying model. Socioeconomic factors and characteristics of each region, such as population density, average income, education level, and other pertinent demographic or geographic factors, influence this model. We posit that n_i follows a log-linear model [18], a common choice for count data. This leads us to the following equation:

$$\log n = X\beta + \epsilon, \tag{3}$$

where $n = (n_1, \cdots, n_M)^T \in \mathbb{N}^{M \times 1}$ is the vector of true counts, $\epsilon \sim N(0, \sigma_n^2 I_M)$ is the error term, $\beta \in \mathbb{R}^{K \times 1}$ is the vector of parameters, and $X \in \mathbb{R}^{M \times K}$ is the known matrix representing the influence of the features on the incident counts.

3. PROPOSED DEBIASING ALGORITHM: GRAUD

This section introduces an optimization problem as a part of a novel debiasing algorithm, GRAUD. The optimization problem revolves around two variables: n and p under certain constraints. The formulation originates from the fact that $\mathbb{E}[y] = np$ combined with graph smoothness and the underlying model of n. Then, a series of transformations and manipulations are conducted, leading to an alternative but conceptually equivalent optimization problem. To begin with, we can consider the following estimation problem:

$$\min_{n,\beta,0 \le p \le 1} \ \|y - n \odot p\|^2 + \lambda_1 p^T L p + \lambda_2 \|\log n - X\beta\|^2, \tag{4}$$

where $\odot$ is elementwise (Hadamard) product, $y = [y_1, \ldots, y_M]^T$, $n = [y_1 \ldots, y_M]^T$, X is a $M \times K$ matrix and the regularization parameters are $\lambda_1 > 0$ and $\lambda_2 > 0$. The two regularization parameters control the trade-off between the data-fitting term and the regularization terms

Given that both n and y represent count data in this case, a log transformation might be beneficial as it can make the data more normally distributed and reduce the variability [19]. Due to these advantages, applying a log transformation to count data is a common practice in statistical analysis. Furthermore, using log transformations can simplify the formulation of the problem.

In this formulation, optimal β can be directly computed through least-squares regression:

$$\beta^* = \text{argmin}_\beta \|\log n - X\beta\|^2 = (X^T X)^{-1} X^T \log n. \tag{5}$$

Upon substituting the optimal β^*, the optimization problem transforms into:

$$\begin{aligned} \min_{n \ge 1, 0 < p \le 1} & \|\log y - \log n - \log p\|^2 \\ & + \lambda_1 p^T L p + \lambda_2 \log n^T H \log n, \end{aligned} \tag{6}$$

Algorithm 1 Graph-based debiasing Algorithm for Underreported Data (GRAUD)

Require: Initial u_1, v_1, $\tilde{y}$, λ_1, λ_2, L, H, iteration T_{in}, T_{out}, threshold ϵ, stepsize η

Ensure: $n^* = \exp(u), p^* = \exp(v)$

for $k = 1, \cdots, T_{\text{out}}$ **do**:

  for $t = 1, \cdots, T_{\text{in}}$ **do**:

    $du = u_k + v_k - \tilde{y} + \lambda_2 H u_k$

    $u_k = u_k - \eta du$

  for $t = 1, \cdots, T_{\text{in}}$ **do**:

    $dv = u_k + v_k - \tilde{y} + \lambda_2 L v_k$

    $v_k = v_k - \eta dv$

  $u_{k+1} = u_k, v_{k+1} = v_k$

where $H = I - X(X^T X)^{-1} X^T$ is the projection matrix.

With the new variables $\tilde{y} = \log y$, $u = \log n$ and $v = \log p$, the optimization problem can be expressed as:

$$\min_{u \geq 0, v \leq 0} \|\tilde{y} - u - v\|^2 + \lambda_1 v^T L v + \lambda_2 u^T H u. \quad (7)$$

An alternating minimization algorithm can be utilized to address this optimization problem, and this proposed method is outlined in Algorithm 1.

4. THEORETICAL ANALYSIS

4.1. Assumptions

First, let's enumerate the assumptions vital to our approach. These assumptions direct the algorithm design and lead to theoretical guarantees. Here we denote $u_0 = \log n_0$ and $v_0 = \log p_0$ as ground truth.

Assumption 1. *Let $\epsilon_u = u_0^T H u_0$ and $\epsilon_v = v_0^T L v_0$, we assume that the quantity ϵ_u and ϵ_v are small.*

Assumption 2. *Assume $\{x_1, \cdots, x_{r_X}\}$ form an orthonormal basis of the column space of X, $\{l_1, \cdots, l_{r_L}\}$ form an orthonormal basis of the null space of L, where r_X is the dimension of the column space of X (null space of H), and r_L is the dimension of the null space of L. There exists a $\delta_1 > 0$ so that $\min_{\|\alpha\|=1} \|\alpha^T (x_1, \cdots, x_K, l_1, \cdots, l_t)\|^2 = \delta_1$.*

The first assumption, as discussed in the preceding section, plays a crucial role in the accuracy of GRAUD. The second assumption essentially states that the zero vector is the only common element between the null spaces of H and L. This assumption is important for resolving the identifiability problem.

4.2. Recovery Guarantee

In this section, we dissect the properties of our proposed problem. We aim to showcase the applicability of this method in debiasing the under-count data. We clarify how this optimization problem aligns with our goal.

Recall that $u = \log n \geq 0$, $v = \log p \leq 0$ and $\tilde{y} = \log y$, where $y \sim \text{Binomial}(n, p)$. All the proof of theorems can be found in the Appendix.

Proposition 4.1. *Under assumption 2, problem* (7) *is convex and has a unique solution.*

Let $\epsilon_y = \log y - \log n_0 - \log p_0$, and ϵ_u, ϵ_v defined in Assumption 1. The following is our main theorem.

Theorem 4.2. *Under assumption 1 and 2, the solution u^* and v^* of the optimization problem* (7) *satisfies*

$$\begin{aligned} \|u^* - u_0\| &\leq \tilde{c_1} \|\epsilon_y\|^2 + \|\epsilon_u\|^2 \\ \|v^* - v_0\| &\leq \tilde{c_2} \|\epsilon_y\|^2 + \|\epsilon_v\|^2, \end{aligned} \quad (8)$$

for some constant $\tilde{c_1}, \tilde{c_2} > 0$.

The term $\|\epsilon_y\|$ diminishes towards zero with a high probability as n increases. Additionally, based on our initial assumptions, the terms $\|\epsilon_u\|$ and $\|\epsilon_v\|$ are expected to be very small. Given these factors, we can infer that the upper bound delineated on the right side of the equation will be considerably small.

Besides, we have a global convergence result for our Algorithm 1.

Proposition 4.3. *The output (u_k, v_k) generated by Algorithm 1 converges to a critical point of Problem* (6)*, which is the unique global minimum.*

5. NUMERICAL EXPERIMENTS

5.1. Simulated Examples

We proceed to evaluate the efficacy of GRAUD through two simulated examples. In the initial experiment, we arbitrarily select $X - 2 \in \mathbb{R}^{M \times K}$ from a standard normal distribution, setting $\beta \in \mathbb{R}^{K \times 1}$ as a vector with all elements equal to one. We adopt $M = 10$ and $K = 3$ for this experiment. Then, we create $n = [\exp(X\beta)]$, ensuring that Assumption (1) is satisfied by maintaining $n^T H n$ relatively small. For p, we compute it using $p = 0.7 + 0.1\epsilon$, where $\epsilon \in N(0, I_M)$ is derived from a standard normal distribution. We subsequently constrain p to fall within the $[0.05, 0.95]$ interval to circumvent extreme scenarios. Furthermore, we set $p^T L p \leq 0.02$ to satisfy Assumption 1. In this context, λ_1 is set at 0.01, and λ_2 at 0.9. We select the regularization parameters through 5-fold cross-validation. As for the initial values, we assign $n_1 = y$ and $p_1 = y/n_1$. This results in $u_1 = \log n_1$ and $v_1 = \log p_1$.

As demonstrated in Figure 1, the debiased solution generated by GRAUD is closely aligned with the ground truth, indicating high accuracy in our approach. The proximity of GRAUD's output to the ground truth underscores its reliability in providing accurate results, thus justifying its application in this context.

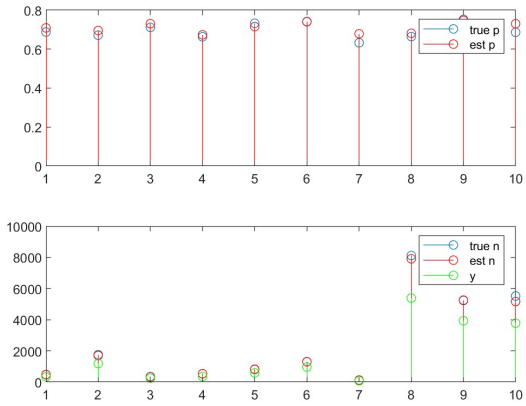

Fig. 1. The plot presents a comparison between the true value of p and its estimated counterpart, as well as a comparison between the true value of n, the estimated value of n, and y. The relative ℓ_1 error for p stands at 0.035, while the relative ℓ_1 error for n is 0.029.

5.2. Real Data Experiment

In the real-world experiment, we direct our attention towards emergency (911) call data originating from Atlanta, specifically from the year 2019, which comprises approximately 580,000 instances. Notably, the actual number of emergency situations is likely higher than represented by these calls, as they tend to underestimate the true magnitude of emergencies. We use this data to establish the variable y_i for every individual beat, with a beat referring to the distinct geographical area assigned to a police officer for patrolling. These beats subdivide Atlanta into 78 distinct sections, which offers a naturally discrete geographical division for our research.

To enhance our understanding, we create a graphical model in which each beat is symbolized as a node, and edges are formed between nodes corresponding to neighboring beats. This graph-based representation allows us to visualize and comprehend the spatial connections and proximity among the different beats more intuitively.

To supplement our dataset further, we include the census data from 2019, factoring in aspects such as population size, income, and level of education (quantified as the fraction of the population that has achieved at least a high school diploma). These factors constitute our x_{ij} variables, thereby incorporating socioeconomic factors into our analysis.

To begin our analysis, we set the initial p_i vector as a vector of all 0.8s and $n_i = y_i/p_i$. The localized solution we achieve from this starting point is represented in Figure 3. The yellow areas represent a higher discovery probability, and those areas with higher discovery rates are mainly located in the downtown, midtown, or other prosperous areas in Atlanta. This makes sense because those flourishing areas usually have better public security, thus resulting in higher discovery rates.

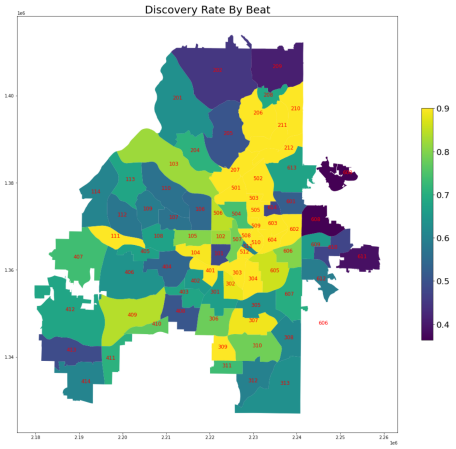

Fig. 2. The estimated p_i in each beat when initializing with the discovery probability of all 0.8.

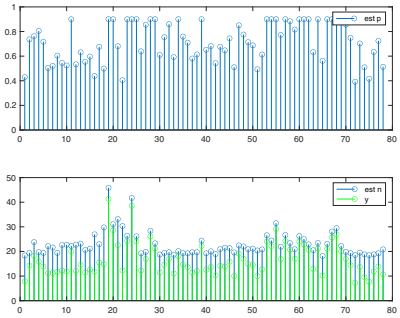

Fig. 3. The estimated $\hat{p}_i$ from data (when converges to locally). The p_i are initialized to be constant 0.1. Clearly, the different regions may have different levels of crime discovery.

6. CONCLUSION

In this paper, we proposed a novel graph prediction method for debiasing under-count data. We utilize the intrinsic graph structure in the problem and overcome the identifiability issue. We reformulate the problem as an optimization problem and establish the connection between the binomial n problem and the graph signal separation problem. We provide an alternating minimization optimization algorithm for efficiently recovering data. We establish recovery bounds and convergence results for our proposed method and conduct several experiments on both synthetic data and real data, demonstrating the accuracy and efficiency of our proposed method.

Acknowledgement

We want to thank Sarah Huestis for her help with the real data experiment. This work is partially supported by an NSF CAREER CCF-1650913, NSF DMS-2134037, CMMI-2015787, CMMI-2112533, DMS-1938106, and DMS-1830210, and a Coca-Cola Foundation fund.

7. REFERENCES

[1] Lorna Hazell and Saad AW Shakir, "Under-reporting of adverse drug reactions," *Drug safety*, vol. 29, no. 5, pp. 385–396, 2006.

[2] Angela Watson, Barry Watson, and Kirsten Vallmuur, "Estimating under-reporting of road crash injuries to police using multiple linked data collections," *Accident Analysis & Prevention*, vol. 83, pp. 18–25, 2015.

[3] Junaid Shuja, Eisa Alanazi, Waleed Alasmary, and Abdulaziz Alashaikh, "Covid-19 open source data sets: a comprehensive survey," *Applied Intelligence*, vol. 51, pp. 1296–1325, 2021.

[4] A DasGupta and Herman Rubin, "Estimation of binomial parameters when both n, p are unknown," *Journal of Statistical Planning and Inference*, vol. 130, no. 1-2, pp. 391–404, 2005.

[5] Sanjib Basu, "Ch. 31. bayesian inference for the number of undetected errors," *Handbook of Statistics*, vol. 22, pp. 1131–1150, 2003.

[6] Sanjib Basu and Nader Ebrahimi, "Bayesian capture-recapture methods for error detection and estimation of population size: Heterogeneity and dependence," *Biometrika*, vol. 88, no. 1, pp. 269–279, 2001.

[7] Norman Draper and Irwin Guttman, "Bayesian estimation of the binomial parameter," *Technometrics*, vol. 13, no. 3, pp. 667–673, 1971.

[8] Dorian Feldman and Martin Fox, "Estimation of the parameter n in the binomial distribution," *Journal of the American Statistical Association*, vol. 63, no. 321, pp. 150–158, 1968.

[9] Adrian E Raftery, "Inference for the binomial n parameter: A bayes empirical bayes approach. revision.," Tech. Rep., WASHINGTON UNIV SEATTLE DEPT OF STATISTICS, 1987.

[10] Lucy Whitaker, "On the poisson law of small numbers," *Biometrika*, vol. 10, no. 1, pp. 36–71, 1914.

[11] P Fisher et al., "Negative binomial distribution.," *Annals of Eugenics*, vol. 11, pp. 182–787, 1941.

[12] John Burdon Sanderson Haldane, "The fitting of binomial distributions," *Annals of Eugenics*, vol. 11, no. 1, pp. 179–181, 1941.

[13] William D Kahn, "A cautionary note for bayesian estimation of the binomial parameter n," *The American Statistician*, vol. 41, no. 1, pp. 38–40, 1987.

[14] GG Hamedani and GG Walter, "Bayes estimation of the binomial parameter n," *Communications in Statistics-Theory and Methods*, vol. 17, no. 6, pp. 1829–1843, 1988.

[15] Erdogan Günel and Daniel Chilko, "Estimation of parameter n of the binomial distribution," *Communications in Statistics-Simulation and Computation*, vol. 18, no. 2, pp. 537–551, 1989.

[16] Shixiang Zhu and Yao Xie, "Generalized hypercube queuing models with overlapping service regions," *arXiv preprint arXiv:2304.02824*, 2023.

[17] David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE signal processing magazine*, vol. 30, no. 3, pp. 83–98, 2013.

[18] Alexander Von Eye, Eun-Young Mun, and Patrick Mair, "Log-linear modeling," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 4, no. 2, pp. 218–223, 2012.

[19] FENG Changyong, WANG Hongyue, LU Naiji, CHEN Tian, HE Hua, LU Ying, et al., "Log-transformation and its implications for data analysis," *Shanghai archives of psychiatry*, vol. 26, no. 2, pp. 105, 2014.