Generalized Linear Models for Aggregated Data

Avradeep Bhowmik
University of Texas at Austin
avradeep.1@utexas.edu

Joydeep Ghosh
University of Texas at Austin
ghosh@ece.utexas.edu

Oluwasanmi Koyejo
Stanford University
sanmi@stanford.edu

Abstract

Databases in domains such as healthcare are routinely released to the public in aggregated form. Unfortunately, naïve modeling with aggregated data may significantly diminish the accuracy of inferences at the individual level. This paper addresses the scenario where features are provided at the individual level, but the target variables are only available as histogram aggregates or order statistics. We consider a limiting case of generalized linear modeling when the target variables are only known up to permutation, and explore how this relates to permutation testing; a standard technique for assessing statistical dependency. Based on this relationship, we propose a simple algorithm to estimate the model parameters and individual level inferences via alternating imputation and standard generalized linear model fitting. Our results suggest the effectiveness of the proposed approach when, in the original data, permutation testing accurately ascertains the veracity of the linear relationship. The framework is extended to general histogram data with larger bins - with order statistics such as the median as a limiting case. Our experimental results on simulated data and aggregated healthcare data suggest a diminishing returns property with respect to the granularity of the histogram - when a linear relationship holds in the original data, the targets can be predicted accurately given relatively coarse histograms.

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.

1 Introduction

Modern life is highly data driven. Datasets with records at the individual level are generated every day in large volumes. This creates an opportunity for researchers and policy-makers to analyze the data and examine individual level inferences. However, in many domains, individual records are difficult to obtain. This particularly true in the healthcare industry where protecting the privacy of patients restricts public access to much of the sensitive data. Therefore, in many cases, multiple Statistical Disclosure Limitation (SDL) techniques are applied [9]. Of these, data aggregation is the most widely used technique [5].

It is common for agencies to report both individual level information for non-sensitive attributes together with the aggregated information in the form of sample statistics. Care must be taken in the analysis of such data, as naïve modeling with aggregated data may significantly diminish the accuracy of inferences at the individual level. In particular, inferences drawn from aggregated data may lead to the problem of ecological fallacy [23], hence the resulting conclusions at the group level may be misleading to researchers and policy makers interested in individual level inferences. An example that has been cited [21] is the high correlation between per capita consumption of dietary fat and breast cancer in different countries, which may lead to the incorrect conclusion that dietary fat causes breast cancer [7].

Aggregated data in the form of histograms and other sample statistics are becoming more and more common. Further, most of the data that is collected relates to questions for which the respondents have only a few discrete options from which to select their answer. For example, data available from the Generalized Social Survey (GSS) [1] are often in this form. This paper addresses the scenario where features are provided at the individual level, but the target variables are only available as histogram aggregates or order statistics. Despite the prevalence of order-statistic and histogram

aggregated data, to the best of our knowledge, this problem has not been addressed in the literature.

We consider a limiting case of generalized linear modeling when the target variables are only known up to permutation, and explore how this relates to permutation testing [12]; a standard technique for assessing statistical dependency. Based on this relationship, we propose a simple algorithm to estimate the model parameters and individual level inferences via alternating imputation and standard generalized linear model fitting. Our results suggest the effectiveness of the proposed approach when, in the original data, permutation testing accurately ascertains the veracity of the linear relationship. The framework is extended to general histogram data with larger bins - with order statistics such as the median as a limiting case. Our experimental results suggest a diminishing returns property - when a linear relationship holds in the original data, the targets can be predicted accurately given relatively coarse histograms. Our results also suggest caution in in the widespread use of aggregation for ensuring the privacy of sensitive data.

In summary, the main contributions of this manuscript are as follows:

(i) we propose a framework for estimating the response variables of a generalized linear model given only a histogram aggregate summary by formulating it as an optimization problem that alternates between imputation and generalized linear model fitting.

(ii) we examine a limiting case of the framework when all the data is known up to permutation. Our examination suggests the effectiveness of the proposed approach when, in the original data, permutation testing accurately ascertains the veracity of the linear relationship.

(iii) we examine a second limiting case where only a few order statistics are provided. Our experimental results suggest a diminishing returns property - when a linear relationship holds in the original data, the targets can be predicted accurately given relatively coarse histograms.

The proposed approach is applied to the analysis of simulated datasets. In addition, we examine the Texas Inpatient Discharge dataset from the Texas Department of State Health Services [3] and a subset of the 2008-2010 SynPUF dataset [2].

Notation

Matrices are denoted by boldface capital letters, vectors by boldface lower case letters and individual elements of the vector by the same lowercase letter with the boldface removed and the index added as a superscript. $\mathbf{v}^\top$ refers to the transpose of the column vector $\mathbf{v}$. We denote column partitions using semicolons, that is, $\mathbf{M} = [\mathbf{X}; \mathbf{Y}]$ implies that the columns of the submatrices $\mathbf{X}$ and $\mathbf{Y}$ are, in order, the columns of the full matrix $\mathbf{M}$. We use $\|\cdot\|$ to denote the L_2 norm for vectors and Frobenius norm for matrices. The vector $\mathbf{v}$ is said to be in increasing order if $v^{(i)} \leq v^{(j)}$ whenever $i \leq j$, and the set of all such vectors in $\mathbb{R}^n$ is denoted with a subscripted downward pointing arrow as $\mathbb{R}^n_\downarrow$. Two vectors $\mathbf{v}$ and $\mathbf{w}$ are said to be isotonic, $\mathbf{v} \sim_\downarrow \mathbf{w}$, if $v^{(i)} \geq v^{(j)}$ if and only if $w^{(i)} \geq w^{(j)}$ for all i, j.

1.1 Preliminaries and Related Work

Aggregated data is often summarized using a sample statistic, which provides a succinct descriptive summary [26]. Examples of sample statistics include the average, median and various other quantiles. While the mean is still the most common choice, the best choice for summarizing a sample generally depends on the distribution the sample has been generated from. In many cases, the use of histograms [24] or order statistic summaries is much more "natural" e.g. for categorical data, binary data, count valued data, etc.

The problem of imputing individual level records from the sample mean has been studied in [20] and [21] among others. In particular, the paper [21] attempts to reconstruct the individual level matrix by assuming a low rank structure and compares their framework with other approaches which include an extension of the neighborhood model [10] and a variation of ecological regression [13] for the task of imputing individual level records of the response variable. However, these approaches exploit the fact that sample mean is a linear function of the sample values. Hence, none of these approaches are extendable to non-linear functions such as order statistics and histogram aggregates. To the best of our knowledge, the special case of individual inferences based on sample statistic aggregate data has not been addressed before. Our work proposes a first solution to address this open problem.

Order Statistics: Given a sample of n real valued datapoints, the τ^{th} order statistic of the sample is the τ^{th} smallest value in the sample. For example, the first order statistic is the minimum value of the sample, the $\frac{n}{2}^{th}$ order statistic is the median and the n^{th} order statistic is the maximum value of the sample. We specifically design a framework which makes it relatively straightforward to work with order statistics.

Histograms : Given a finite sample of n items from a set $\mathcal{C}$, a histogram is a partition of the set $\mathcal{C}$ into disjoint bins $C_i : \cup_i C_i = \mathcal{C}$ and the respective count or

percentage of elements from the sample in each bin. Seen this way, for any sample from $\mathcal{C} \subseteq \mathbb{R}$, a histogram is essentially a set of order statistics for that sample. Histograms can sometimes be specified without their boundary values (eg. "$x < 30$" as opposed to "$0 < x < 30$")- this is equivalent to leaving out the first and the n^{th} order statistic. Further, a set of sample statistic summaries are easily converted to (and from) a discrete cumulative distribution by identifying the quantile value as the cumulative histogram boundary, and the quantile identity as the height. This cumulative histogram is easily converted to a standard histogram by differencing of adjacent bins. A similar strategy is also applicable to unbounded domains using abstract *max* and *min* boundaries of $\pm\infty$. Based on this bijection, we we will refer to a histogram as a generalization of the order statistic for the remainder of this manuscript.

Bregman Divergence : Let $\boldsymbol{\phi} : \Theta \mapsto \mathbb{R}$ be a strictly convex, closed function on the domain $\Theta \subseteq \mathbb{R}^m$ which is differentiable on int(Θ). Then, the Bregman divergence $D_{\boldsymbol{\phi}}(\cdot\|\cdot)$ corresponding to the function $\boldsymbol{\phi}$ is defined as

$$D_{\boldsymbol{\phi}}(\mathbf{y}\|\mathbf{x}) \triangleq \boldsymbol{\phi}(\mathbf{y}) - \boldsymbol{\phi}(\mathbf{x}) - \langle \nabla\boldsymbol{\phi}(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle$$

From strict convexity, it follows that $D_{\boldsymbol{\phi}}(\mathbf{y}\|\mathbf{x}) \geq 0$ and $D_{\boldsymbol{\phi}}(\mathbf{y}\|\mathbf{x}) = 0$ if and only if $\mathbf{y} = \mathbf{x}$. Bregman divergences are strictly convex in their first argument but not necessarily in their second argument. In this paper we only consider convex functions of the form $\boldsymbol{\phi}(\cdot) : \mathbf{R}^m \ni \mathbf{x} \mapsto \sum_i \phi(x^{(i)})$ that are sums of identical scalar convex functions applied to each component of the vector $\mathbf{x}$. We refer to this class as *identically separable* (**IS**). Square loss, Kullback-Leibler (KL) divergence and generalized I-Divergence (GI) are members of this family (Table 1).

Table 1: Examples of Bregman Divergences

$\boldsymbol{\phi}(\mathbf{x})$	$D_{\boldsymbol{\phi}}(\mathbf{y}\|\mathbf{x})$
$\frac{1}{2}\|\mathbf{x}\|^2$	$\frac{1}{2}\|\mathbf{y} - \mathbf{x}\|^2$
$\sum_i (x^{(i)} \log x^{(i)})$ $\mathbf{x} \in$ Prob. Simplex	$\mathrm{KL}(\mathbf{y}\|\mathbf{x}) =$ $\sum_i \left(y^{(i)} \log(\frac{y^{(i)}}{x^{(i)}}) \right)$
$\sum_i x^{(i)} \log x^{(i)} - x^{(i)}$ $\mathbf{x} \in \mathbb{R}^n_+$	$\mathrm{GI}(\mathbf{y}\|\mathbf{x}) =$ $\sum_i y^{(i)} \log(\frac{y^{(i)}}{x^{(i)}}) - y^{(i)} + x^{(i)}$

Generalized Linear Models: While least squares regression is useful for modeling continuous real valued data generated from a Gaussian distribution. This is not always a valid assumption. In many cases, the data of interest may be binary valued or count valued. A generalized linear model (GLM) [18] is a generalization of linear regression that subsumes various models like Poisson regression, logistic regression, etc. as special cases[1]. A generalized linear model assumes that the response variables, y are generated from a distribution in the exponential family with the mean parameter related via a link function to a linear function of the predictor $\mathbf{x}$. The model therefore is specified completely by a distribution $P_{\boldsymbol{\phi}}(\cdot \mid \boldsymbol{\beta})$ from the exponential family, a linear predictor $\eta = \mathbf{x}\boldsymbol{\beta}$, and a link function $(\nabla\phi)^{-1}(\cdot)$ which connects the expectation parameter of the response variable to the predictor variables as $E(y) = (\nabla\phi)^{-1}(\mathbf{x}\boldsymbol{\beta})$.

As explored in great detail in [6], Bregman Divergences have a very close relationship with generalized linear models. In particular, maximum likelihood parameter estimation for a generalized linear model is equivalent to minimizing a corresponding Bregman divergence. For example, maximum likelihood for a Gaussian corresponds to squares loss, for Poisson the corresponding divergence is generalized I-divergence and for Binomial, the corresponding divergence is the KL divergence (see [6] for details). GLMs have been successfully applied in a wide variety of fields including machine learning , biological surveys [19], image segmentation and reconstruction [22], analysis of medical trials [8], studying species-environment relationships in ecological sciences [15], virology [11] and estimating mortality from infectious diseases [14], among many others, and are widely prized for the interpretability of their results and the extendability of their methods in a plethora of domain specific variations [25]. They are easy to use and implement and many off-the-shelf software packages are available for most major programming platforms.

2 Problem Description

Consider a set of fully observed covariates $\mathbf{X} = [\mathbf{x_1}; \mathbf{x_2}; \cdots; \mathbf{x_{d-p}}] \in \mathbb{R}^{n\times(d-p)}$, and columns of response variables, $\mathbf{Z} = [\mathbf{z_1}; \mathbf{z_2}; \cdots \mathbf{z_p}] \in \mathbb{R}^{n\times p}$, which are only known only up to the respective histograms of their values (i.e., up to order statistics).

We assume that each element of $\mathbf{z_i}$ has been generated from covariates $\mathbf{X}$ according to some generalized linear model with parameters $\boldsymbol{\beta}_i$. The objective is to estimate the $\boldsymbol{\beta}_i$ together with $\mathbf{Z} = [\mathbf{z_1}; \mathbf{z_2}; \cdots \mathbf{z_p}]$ subject to the given order statistic constraints. Since maximum likelihood estimation in a generalized linear model is equivalent to minimizing a corresponding Bregman divergence, we choose the loss function $\mathcal{L}(\mathbf{Z}, \boldsymbol{\beta}) = D_{\boldsymbol{\phi}}\left(\mathbf{Z}\|(\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta})\right)$ to be minimized over

[1] see [17] or [16] for a detailed discussion on GLMs

the variables $\mathbf{Z}, \boldsymbol{\beta}$ while satisfying order statistics constraints on $\mathbf{Z}$.

Without additional structure, the regression problem for each column can be solved independently, therefore without loss of generality we assume $\mathbf{Z}$ is a single column $\mathbf{z}$. We denote the ${\tau_i}^{th}$ order statistic of $\mathbf{z}$ as s_{τ_i}, with $\tau_i \in \{\tau_1, \tau_2, \cdots \tau_h\} \subseteq [n]$, which is the set of h order statistics specified via the histogram. For simplicity, in the following section we consider estimation under a single order statistic which has been computed over the entire column. We extend it subsequently to the more general case of multiple order statistics computed over disjoint partitions.

Therefore, with Frobenius regularization terms $\mathcal{R}(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|^2$, the overall problem statement boils down to the following optimization problem:

$$\begin{aligned} \min_{\mathbf{z},\boldsymbol{\beta}} \quad & D_{\boldsymbol{\phi}}\left(\mathbf{z}\|(\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta})\right) + \lambda\|\boldsymbol{\beta}\|^2 \\ \text{s.t.} \quad & {\tau_i}^{th} \text{ order statistic of } \mathbf{z_i} = s_{\tau_i} \end{aligned} \tag{1}$$

2.1 Estimation under a Single Order Statistic constraint

Estimating under order statistics constraints is in general a highly non-trivial problem. It is easy to see that the set of vectors with a given order statistic is not a convex set. Therefore, the above optimization problem looks especially difficult to even represent in a concise manner in terms of $\mathbf{z}$. However, it turns out that with the following reformulation, the analysis of the problem becomes much more manageable.

We rewrite $\mathbf{z} = \mathbf{P}\mathbf{y}$ where $\mathbf{P} \in \mathbb{P}$ is a permutation matrix and $\mathbf{y}$ is a vector sorted in increasing order. Note the following-

(i) For a $\mathbf{y} \in \mathbb{R}^n_{\downarrow}$, if $\mathbf{e}^{\boldsymbol{\tau_i}}$ is a row vector with 1 in the ${\tau_i}^{th}$ index and 0 everywhere else, then $\mathbf{e}^{\boldsymbol{\tau_i}}\mathbf{y}$ represents the ${\tau_i}^{th}$ order statistic of $\mathbf{y}$. Since permutation does not change the value of order statistics, this is also the ${\tau_i}^{th}$ order statistic of $\mathbf{z}$

(ii) If $\boldsymbol{\Lambda}$ is the matrix with $\boldsymbol{\Lambda}_{j,j+1} = -1, \boldsymbol{\Lambda}_{j,j} = 1$ and $\boldsymbol{\Lambda}_{j,k} = 0$ for all other $j, k: \ (k-j) \neq 0, \pm 1$, the condition that $\mathbf{y}$ is sorted in increasing order is equivalent to the linear constraint $\boldsymbol{\Lambda}\mathbf{y} \leq 0$.

Putting all this together, the optimization problem (1) becomes the following

$$\begin{aligned} \min_{\mathbf{P},\mathbf{y},\boldsymbol{\beta}} \quad & D_{\boldsymbol{\phi}}\left(\mathbf{P}\mathbf{y}\|(\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta})\right) + \mathcal{R}(\boldsymbol{\beta}) \\ \text{s.t.} \quad & \mathbf{e}^{\boldsymbol{\tau_i}}\mathbf{y} = s_{\tau_i}, \ \boldsymbol{\Lambda}\mathbf{y} \geq 0, \ \mathbf{P} \in \mathbb{P} \end{aligned} \tag{2}$$

The above optimization problem is jointly convex in $\mathbf{y}$ and $\boldsymbol{\beta}$ for a fixed $\mathbf{P}$, but the presence of $\mathbf{P}$ as a variable makes the problem much more complicated. Therefore, we attempt to solve it iteratively for each variable in an alternating minimization framework. The update steps consist of the following for each timestep:

(i) $\boldsymbol{\beta}_t = \underset{\boldsymbol{\beta}}{\text{argmin}}\ D_{\boldsymbol{\phi}}\left(\mathbf{P_{t-1}}\mathbf{y_{t-1}}\|(\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta})\right) + \mathcal{R}(\boldsymbol{\beta})$

(ii) $\mathbf{y_t} = \underset{\mathbf{y}}{\text{argmin}} D_{\boldsymbol{\phi}}\left(\mathbf{P_{t-1}}\mathbf{y}\|(\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta}_t)\right)$ such that $\boldsymbol{\Lambda}\mathbf{y} \leq 0$ and $\mathbf{e}^{\boldsymbol{\tau_i}}\mathbf{y} = s_{\tau_i}$

(iii) $\mathbf{P_t} = \underset{\mathbf{P}\in\mathbb{P}}{\text{argmin}}\ D_{\boldsymbol{\phi}}\left(\mathbf{P}\mathbf{y_t}\|(\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta}_t)\right)$

Step (i) is a standard generalized linear model parameter estimation problem. This problem has been studied in great detail in literature and a variety of off-the-shelf GLM solvers can be used for this. We focus instead on steps (ii) and (iii) which are much more interesting.

For (ii), note that since we assumed that $\boldsymbol{\phi}$ is identically separable, the same permutation applied to both arguments of the corresponding Bregman divergence $D_\phi(\cdot\|\cdot)$ does not change its value. For any constraint set $\mathcal{C}$, we have $\underset{\mathbf{y}\in\mathcal{C}}{\text{argmin}}\ D_{\boldsymbol{\phi}}\left(\mathbf{P}\mathbf{y}\|(\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta}_t)\right) = \underset{\mathbf{y}\in\mathcal{C}}{\text{argmin}}\ D_{\boldsymbol{\phi}}\left(\mathbf{y}\|\mathbf{P}^{-1}(\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta}_t)\right)$ given[2] a fixed $\mathbf{P}, \mathbf{X}, \boldsymbol{\beta}$. Following this fact, step (ii) is a convex optimization problem in $\mathbf{y}$ and can be solved very easily.

Step (iii) is a non-convex optimization problem in general. However, for an identically separable Bregman divergence it turns out that the solution to this is remarkably simple.

Lemma 1. *The (set of) optimal permutation(s) in step (iv) above is given by-*

$$\underset{\mathbf{P}\in\mathbb{P}}{\text{argmin}}\ D_{\boldsymbol{\phi}}\left(\mathbf{P}\mathbf{y_t}\|(\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta}_t)\right) = \hat{\mathbf{P}} : \hat{\mathbf{P}}\mathbf{y_t} \sim_{\downarrow} (\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta}_t)$$

In other words, the optimal permutation is the one which makes $\mathbf{y_{i,t}}$ isotonic with $(\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta}_t)$. Note that the optimal permutation is not unique if $(\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta}_t)$ is not totally ordered. This is a direct application of the following result which appeared as Lemma 3 in the paper [4].

Lemma 2. *If $x_1 \geq x_2$ and $y_1 \geq y_2$ and $\boldsymbol{\phi}(\cdot)$ is identically separable, then*

$$\begin{aligned} D_{\boldsymbol{\phi}}(\left[\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right]\|\left[\begin{smallmatrix} y_1 \\ y_2 \end{smallmatrix}\right]) &\leq D_{\boldsymbol{\phi}}(\left[\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right]\|\left[\begin{smallmatrix} y_2 \\ y_1 \end{smallmatrix}\right]), \text{ and} \\ D_{\boldsymbol{\phi}}(\left[\begin{smallmatrix} y_1 \\ y_2 \end{smallmatrix}\right]\|\left[\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right]) &\leq D_{\boldsymbol{\phi}}(\left[\begin{smallmatrix} y_2 \\ y_1 \end{smallmatrix}\right]\|\left[\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right]) \end{aligned}$$

2.1.1 Solution in terms of z

Lemmata 1 and 2 suggest that we can optimize jointly over $\mathbf{P}$ and $\mathbf{y}$ instead of separately, since for any $\mathbf{y}$ we

[2] note that for a permutation matrix $\mathbf{P}$, $\mathbf{P^{-1}} = \mathbf{P}^\top$

already know the optimal $\mathbf{P}$. Combining the optimization steps (ii) and (iii) in terms of $\mathbf{P}$ and $\mathbf{y}$, our update step for $\mathbf{z}$ in the original optimization problem is the following

$$\begin{aligned} \hat{\mathbf{z}}_\mathbf{t} &= \underset{\mathbf{z}}{\operatorname{argmin}}\ D_{\boldsymbol{\phi}}\left(\mathbf{z} \| (\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta}_t)\right) \\ &\text{s.t.}\quad \tau_i^{th} \text{ order statistic of } \mathbf{z} = s_{\tau_i} \end{aligned} \tag{3}$$

It is not immediately obvious how to approach the solution to this since the constraint set for $\mathbf{z}$ is not convex. However, note that as a result of Lemma 2 it is clear that given a fixed $\mathbf{X}$ and $\boldsymbol{\beta}_\mathbf{t}$ if $\hat{\mathbf{z}}_\mathbf{t}$ is a solution to the subproblem (3), we must have $\hat{\mathbf{z}}_\mathbf{t} \sim_\downarrow (\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta}_t)$.

Therefore, instead of searching over the set of all vectors in $\mathbb{R}^n$, it is sufficient to search only in the subset of vectors that are isotonic with $(\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta}_t)$. It turns out that not only is this set convex given a fixed $\mathbf{X}, \boldsymbol{\beta}_t$, the solution for $\mathbf{z}_t$ is readily available in closed form.

Let $\boldsymbol{\Gamma}_\mathbf{t} = (\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta}_t)$. Since the Bregman Divergence is IS, without loss of generality we can assume that $\boldsymbol{\Gamma}_\mathbf{t}$ is in increasing order, therefore the constraint set for $\mathbf{z}$ becomes $\mathbf{z} \in \mathbb{R}^n_\downarrow$ and order statistics constraints for $\mathbf{z}$ becomes the linear constraint $\mathbf{e}^{\boldsymbol{\tau_i}}\mathbf{z} = s_{\tau_i}$.

Therefore, the optimization problem (3) over $\mathbf{z}$ is equivalent, up to a simple re-permutation step, to the following

$$\begin{aligned} &\min_{\mathbf{z}}\ D_\phi(\mathbf{z} \| \boldsymbol{\Gamma}_\mathbf{t}) \\ &\text{s.t. } \mathbf{z} \in \mathbb{R}^n_\downarrow,\ \mathbf{e}^{\boldsymbol{\tau_i}}\mathbf{z} = s_{\tau_i} \end{aligned} \tag{4}$$

Lemma 3. *Let $\hat{\mathbf{z}}$ be the solution to the optimization problem (4). Then, $\hat{\mathbf{z}}$ is given by-*

$$\hat{z}_t^{(j)} = \begin{cases} s_{\tau_i} & j = \tau_i \\ \max(\Gamma_t^{(j)}, s_{\tau_i}) & j > \tau_i \\ \min(\Gamma_t^{(j)}, s_{\tau_i}) & j < \tau_i \end{cases} \tag{5}$$

Sketch of Proof In the space of all $\mathbf{z}$ ordered in increasing order, the τ_i^{th} order statistic constraint simply becomes $\hat{z}_t^{(j)} < s_{\tau_i}$ for $j < \tau_i$ and vice versa for $j > \tau_i$. Suppose we were to optimize over all space instead of $\mathbb{R}^n_\downarrow$- because the Bregman divergence is identically separable, the optimization problem separates out over different coordinates j as $\hat{z}_t^{(j)} = \arg\min_z D_\phi(z \| \Gamma_t^{(j)})$ such that $z < (>) s_{\tau_i}$ for $j < (>)\tau_i$. This is a unidimensional convex optimization problem the solution to which is given by equation (5) above.

Finally we note that $\hat{\mathbf{z}}_\mathbf{t}$, automatically lies in $\mathbb{R}^n_\downarrow$ since $\boldsymbol{\Gamma}_\mathbf{t} \in \mathbb{R}^n_\downarrow$, and hence, is also the solution to the optimisation problem (4). □

Now, note that since we are performing iterative minimization, the cost function is non-increasing at every step. As the cost function is bounded below by 0, the algorithm converges to a stationary point. We now extend the framework to include histogram constraints and blockwise partitioning.

2.2 Histogram Constraints

In case there are multiple order statistics constraints (histogram), the solution can be obtained by repeated application of equation (5).

Suppose for the column $\mathbf{z}$ we have constraints as ${\tau_i}^{th}$ order statistic of $\mathbf{z} = s_{\tau_i}$ for $\tau_i \in \{\tau_1, \tau_2, \cdots \tau_h\} \subseteq \{1, 2, \cdots n\}$, the solution is given by the following-

1. For all $j < \tau_1$, $\hat{z}^{(j)} = \min(\Gamma_i^{(j)}, s_{\tau_1})$; similarly, for all $j > \tau_h$, $\hat{z}^{(j)} = \max(\Gamma_i^{(j)}, s_{\tau_h})$
2. For all $1 \leq k < h$, and $j\ : \tau_k \leq j \leq \tau_{k+1}$,

$$\hat{z}^{(j)} = \begin{cases} s_{\tau_k} & j = \tau_k \\ s_{\tau_{k+1}} & j = \tau_{k+1} \\ \min\left(s_{\tau_{k+1}}, \max(\Gamma_i^{(j)}, s_{\tau_k})\right) & \tau_k \leq j \leq \tau_{k+1} \end{cases}$$

The proof for this follows in an identical manner to the proof for the non-partitioned case earlier. As above, the updated $\mathbf{z}_\mathbf{t}$ can be obtained by re-permuting $\hat{\mathbf{z}}$ to preserve isotonicity with $(\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta})$. For a fully observed histogram, the update for $\mathbf{z}$ only involves a permutation at each step.

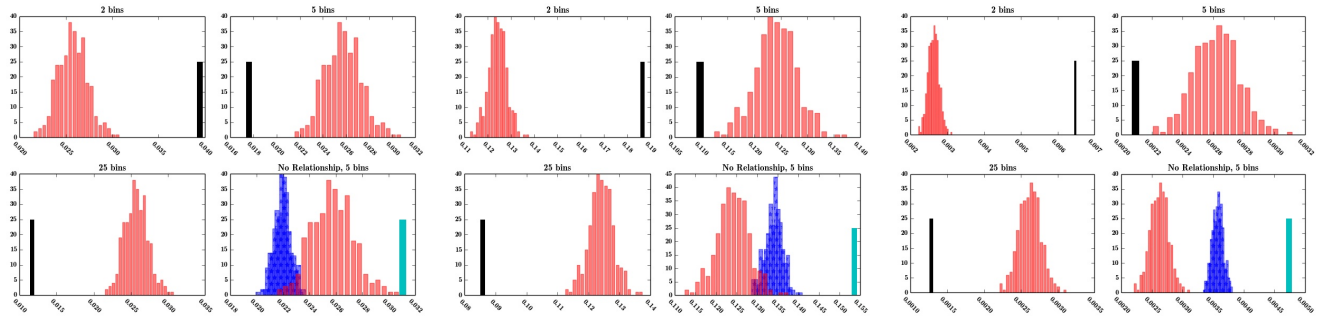
2.3 Blockwise Order Statistic Constraints

In the setup where the order statistics (or histograms) are computed over blockwise partitions of the sample, the permutation matrix is a blockwise permutation matrix and the isotonicity constraint is a blockwise isotonicity costraint.

Since the Bregman Divergence is identically separable, the update for $\mathbf{z}$ separates out into independent updates for every block which can be done in a manner identical to that given by Lemma 3. The update step for $\boldsymbol{\beta}$ remains unchanged.

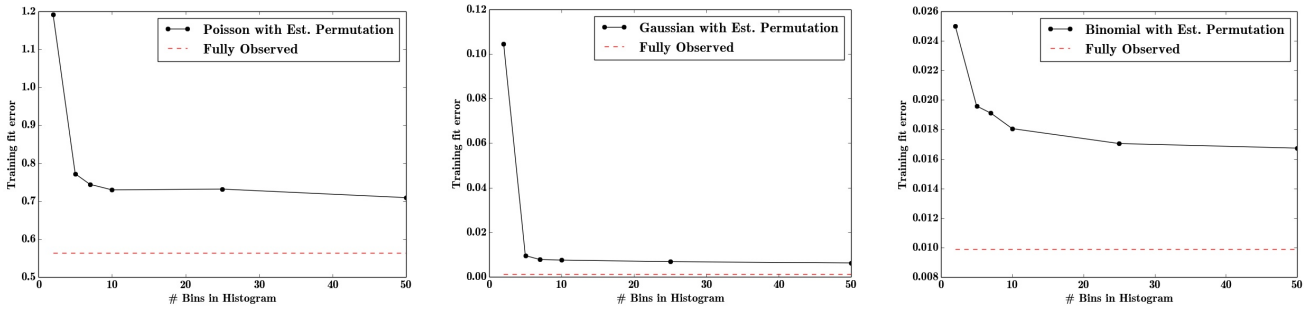
3 Experiments

We provide experimental results using both simulated data and real data. Error for each generalized linear model is defined as the corresponding Bregman divergence (square loss for Gaussian, generalized I-divergence for Poisson, etc. see [6]) between the true and recovered targets. The average errors for each model is shown separately.



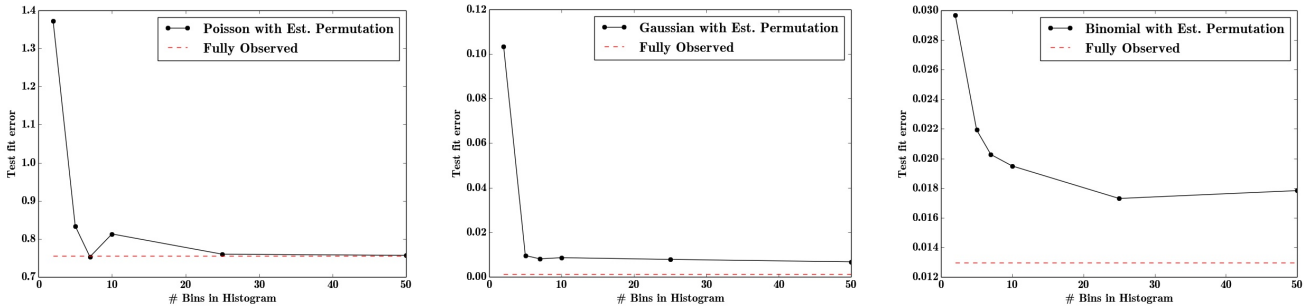
(a) Poisson Fit Error (b) Gaussian Fit Error (c) Binomial Fit Error

Figure 1: Permutation tests under Poisson, Gaussian and Binomial Estimation for 2, 5, 25 bins (top left, top right, bottom left) and "No Relationship" (bottom right)



(a) Poisson Training Error (b) Gaussian Training Error (c) Binomial Training Error

Figure 2: Training Error under Poisson, Gaussian and Binomial Estimation



(a) Poisson Test Error (b) Gaussian Test Error (c) Binomial Test Error

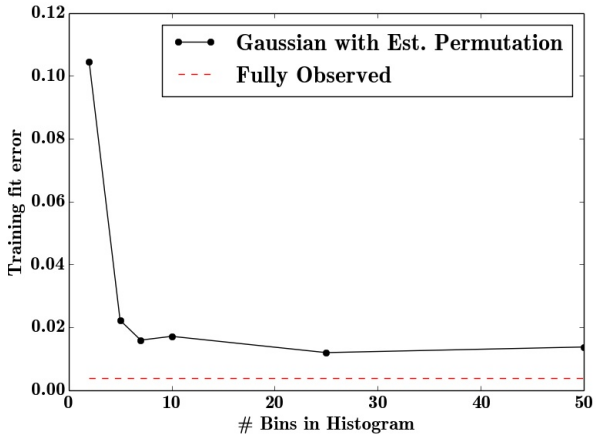
Figure 3: Test Set Error under Poisson, Gaussian and Binomial Estimation

3.1 Simulated Data

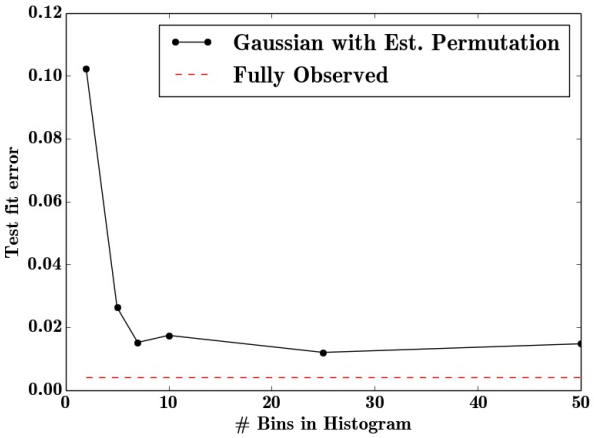
We randomly generate different sets of real valued predictor variables and parameters, and use the corresponding exponential family to generate their respective response variables. We compute histograms for the response variables thus generated with varying number of bins and test our algorithm for each case. We perform the experiments for three different models - Gaussian, Poisson and Binomial.

We perform a basic permutation test[3] to show how our algorithm performs with respect to the fit by a generalized linear model which knows the values of the target variables but permutes the target variables randomly for estimation. We perform the randomized permutations multiple times and plot a histogram of the fitting errors thus obtained and see how the results from our algorithm compares to the histogram (Figure 1). The black bar is the error obtained by our framework, the red histogram is the histogram of errors obtained by fitting after randomly permuting the targets. The blue histogram is the histogram of errors obtained by fitting a model where there is no relationship between the target variable and the covariate, the cyan bar is

[3]Refer to [12] for more details on permutation tests

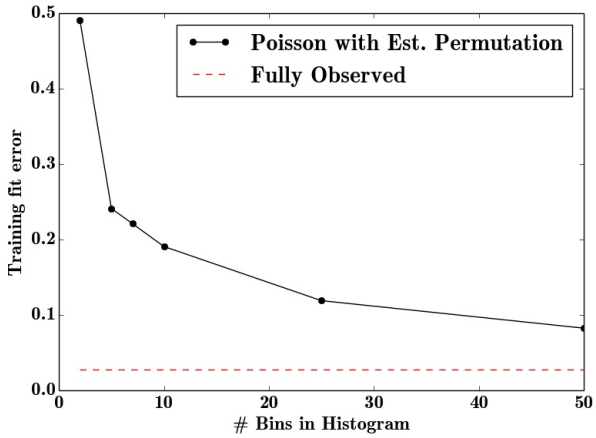


(a) Training Set Error

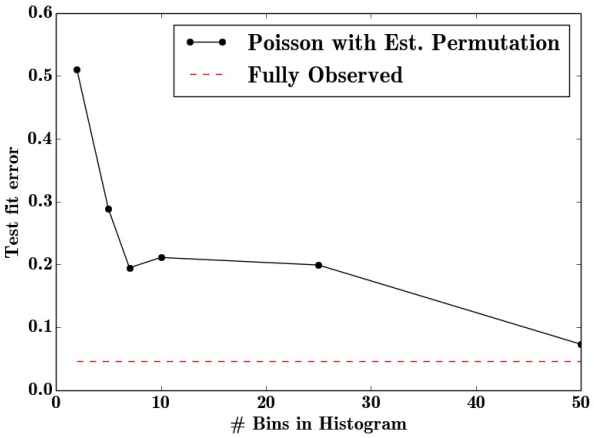


(b) Test Set Error

Figure 4: Performance on SynPUF dataset

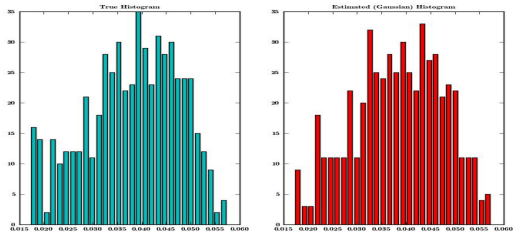


(a) Training Set Error

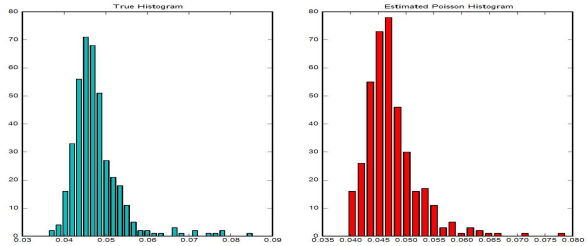


(b) Test Set Error

Figure 5: Performance on Texas Inpatient Discharge dataset



(a) Recovered Histogram of DE-SynPUF Data



(b) Recovered Histogram of Texas Discharged Data

Figure 6: Recovered Histograms of both datasets (true histograms on the left)

the result of our framework applied to this data with a histogram of 5 bins (histograms of other granularities perform similarly). Our test successfully rejects the null hypothesis of "no relationship when the the black bar is to the left of the red histogram. Figure 1 shows that as histogram becomes finer (i.e number of bins increase) error is lower i.e. black bar shifts towards left.

We plot the average fitting and predictive performance of our algorithm with increasing number of bins over five fold cross validation. We compare our results with the results obtained with the best possible GLM esti-

mator which observes the full dataset (Figures 2 and 3)[4]. It can be seen in each case that as the histogram of targets becomes finer (i.e., more bins) the error decreases but with a diminishing returns property with respect to the coarseness of the histogram.

3.2 DE-SynPUF dataset

The CMS Beneficiary Summary DE-SynPUF dataset is a public use dataset created by the Centers for Medicare and Medicaid Services by applying different statistical disclosure limitation techniques to real beneficiary claims data in a way so as to very closely resemble real Medicare data. It is often used for testing different data mining or statistical inferential methods before getting access to real Medicare data. We use a subset of the DE-SynPUF dataset for a single state from the year 2008. With some trimming of datapoints (eg, we do not take into account deceased beneficiaries) we model outpatient institutional annual primary payer reimbursement amount (*PPPYMT-OP*) with a number of available predictor variables including age, race, sex, duration of coverage, presence/absence of a variety of chronic conditions, etc.

We perform a log transform and compute histograms of varying granularity on the target variables. We use a Gaussian model for our estimation and evaluate the average performance of our algorithm over five fold cross validation in fitting both the training and test data sample points, comparing with the best possible Gaussian estimator which performs the estimation by observing the full dataset (Figures 4). As seen in the plot, the performance of our framework improves as the histogram of targets becomes finer in granularity and approaches the performance of the best Gaussian estimator. We also compare the histogram of target variables as recovered by our framework with the true histogram (Figure 6a).

3.3 Texas Inpatient Discharge dataset

We then test our algorithm on the Texas Inpatient Discharge dataset from the Texas Department of State Health Services [3] used in [21]. As with the simulated data, we use histograms of varying granularity on the respective response variables and evaluate the average performance in fitting both the training and test data sample points over five fold cross validation. We use hospital billing records from the fourth quarter of 2006 in the Texas Inpatient Discharge dataset and regress it on the available individual level predictor variables including binary variables race and sex, categorical variables county and zipcode, and real valued variables like length of stay.

Following [21], we perform a log transform on the hospital charges and length of stay before applying a Poisson regression model. We compare the performance of our algorithm over five-fold cross-validation with the best possible Poisson estimator which estimates in a fully observed scenario with an uncensored dataset (Figure 5). The plot shows that the performance of our framework improves with increasingly finer granularity of histograms and approaches the performance of the best Poisson estimator. Finally, we compare the histogram recovered by our framework with the true histogram for the dataset (Figure 6b).

[4] training/test error in figures 2b and 3b for the Gaussian estimator for the fully observed case is ≈ 0

4 Conclusion and Future Work

This paper addresses the scenario where features are provided at the individual level, but the target variables are only available as histogram aggregates or order statistics. We proposed a simple algorithm to estimate the model parameters and individual level inferences via alternating imputation and standard generalized linear model fitting. We considered two limiting cases. In the first, the target variables are only known up to permutation. Our results suggest the effectiveness of the proposed approach when, in the original data, permutation testing accurately ascertains the veracity of the linear relationship. The framework was then extended to general histogram data with larger bins - with order statistics such as the median as a second limiting case. Experimental results on simulated data and real healthcare data show the effectiveness of the proposed approach which may have implications on using aggregation as a means of preserving privacy. For future work, we plan a more detailed analysis to better understand the properties and limits of the framework given binned histogram data. We also plan to extend the approach to non-linear modeling.

Acknowledgements

Authors acknowledge support from NSF grant IIS 1421729.

References

[1] *General Social Survey, NORC.* `http://www3.norc.org/GSS+Website/`.

[2] *Medicare Claims Synthetic Public Use Files (SynPUFs), Centers for Medicare and Medicaid Services.* `http://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/index.html`.

[3] *Texas Department of State Health Services. Texas Inpatient Public Use Data File*, 2014. `https://www.dshs.state.tx.us/thcic/hospitals/Inpatientpudf.shtm`.

[4] S. Acharyya, O. Koyejo, and J. Ghosh. Learning to rank with Bregman divergences and monotone retargeting. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2012.

[5] M. P. Armstrong, G. Rushton, and D. L. Zimmerman. Geographically masking health data to preserve confidentiality. *Statistics in Medicine*, 18 (5):497–525, 1999.

[6] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *The Journal of Machine Learning Research*, 6: 1705–1749, 2005.

[7] K. K. Carroll. Experimental evidence of dietary factors and hormone-dependent cancers. *Cancer Research*, 35(11 Part 2):3374–3383, 1975.

[8] S. Dias, A. J. Sutton, A. Ades, and N. J. Welton. Evidence synthesis for decision making 2 a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. *Medical Decision Making*, 33(5):607–617, 2013.

[9] G. Duncan, M. Elliot, and J. Salazar-González. Statistical confidentiality: Principles and practice 2011.

[10] D. A. Freedman, S. P. Klein, J. Sacks, C. A. Smyth, and C. G. Everett. Ecological regression and voting rights. *Evaluation Review*, 15(6):673–711, 1991.

[11] J. J. Gart. The analysis of poisson regression with an application in virology. *Biometrika*, 51(3/4): pp. 517–521, 1964. ISSN 00063444.

[12] P. I. Good. *Permutation, parametric and bootstrap tests of hypotheses*, volume 3. Springer, 2005.

[13] L. A. Goodman. Ecological regressions and behavior of individuals. *American Sociological Review*, 1953.

[14] P. Hardelid, R. Pebody, and N. Andrews. Mortality caused by influenza and respiratory syncytial virus by age group in England and Wales 1999–2010. *Influenza and other respiratory viruses*, 7 (1):35–45, 2013.

[15] T. Jamil, W. A. Ozinga, M. Kleyer, and C. J. ter Braak. Selecting traits that explain species–environment relationships: a generalized linear mixed model approach. *Journal of Vegetation Science*, 24(6):988–1000, 2013.

[16] P. McCullagh and J. A. Nelder. Generalized linear models. 1989.

[17] J. A. Nelder and R. Baker. *Generalized linear models*. Wiley Online Library, 1972.

[18] J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):pp. 370–384, 1972. ISSN 00359238.

[19] A. Nicholls. How to make biological surveys go further with generalised linear models. *Biological Conservation*, 50(1):51–75, 1989.

[20] Y. Park and J. Ghosh. A probabilistic imputation framework for predictive analysis using variably aggregated, multi-source healthcare data. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 445–454. ACM, 2012.

[21] Y. Park and J. Ghosh. Ludia: an aggregate-constrained low-rank reconstruction algorithm to leverage publicly released health data. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 55–64. ACM, 2014.

[22] G. Paul, J. Cardinale, and I. F. Sbalzarini. Coupling image restoration and segmentation: a generalized linear model/bregman perspective. *International Journal of Computer Vision*, 104(1): 69–93, 2013.

[23] W. S. Robinson. Ecological correlations and the behavior of individuals. *International journal of epidemiology*, 38(2):337–341, 2009.

[24] D. W. Scott. On optimal and data-based histograms. *Biometrika*, 66(3):605–610, 1979.

[25] L. Song, P. Langfelder, and S. Horvath. Random generalized linear model: a highly accurate and interpretable ensemble predictor. *BMC bioinformatics*, 14(1):5, 2013.

[26] S. S. Wilks. Mathematical statistics. *New York*, page 644, 1962.