Sparse Parameter Recovery from Aggregated Data

Avradeep Bhowmik AVRADEEP.1@UTEXAS.EDU
Joydeep Ghosh GHOSH@ECE.UTEXAS.EDU
The University of Texas at Austin, TX, USA

Oluwasanmi Koyejo SANMI@ILLINOIS.EDU
Stanford University, CA & University of Illinois at Urbana Champaign, IL, USA

Abstract

Data aggregation is becoming an increasingly common technique for sharing sensitive information, and for reducing data size when storage and/or communication costs are high. Aggregate quantities such as group-average are a form of semi-supervision as they do not directly provide information of individual values, but despite their wide-spread use, prior literature on learning individual-level models from aggregated data is extremely limited. This paper investigates the effect of data aggregation on parameter recovery for a sparse linear model, when known results are no longer applicable. In particular, we consider a scenario where the data are collected into groups e.g. aggregated patient records, and first-order empirical moments are available only at the group level. Despite this obfuscation of individual data values, we can show that the true parameter is recoverable with high probability using these aggregates when the collection of true group moments is an incoherent matrix, and the empirical moment estimates have been computed from a sufficiently large number of samples. To the best of our knowledge, ours are the first results on structured parameter recovery using only aggregated data. Experimental results on synthetic data are provided in support of these theoretical claims. We also show that parameter estimation from aggregated data approaches the accuracy of parameter estimation obtainable from non-aggregated or "individual" samples, when applied to two real world healthcare applications- predictive modeling of CMS Medicare reimbursement claims, and modeling of Texas State healthcare charges.

1. Introduction

As the scale and scope of data collection continues to grow, data aggregation has become increasingly popular in such varied domains as healthcare and sensor networks. Aggregation is a common technique for sharing of privacy-sensitive healthcare data, where sensitive patient information is subject to various Statistical Disclosure Limitation (SDL) techniques [Armstrong et al. 1999] before public release. Similarly, large scale data collection programs like the General Social Survey (GSS) report data in aggregated form[1]. Data from IoTs and other distributed sensor networks are often collected in aggregated form to mitigate communication costs, and improve robustness to noise and malicious interference [Wagner 2004; Zhao et al. 2003].

Building individual-level models given aggregates in the form of means, sample statistics, etc., constitutes a relatively unexplored semi-supervision framework. We note that even standard problems like regression and parameter recovery become very challenging in the context of aggregated data. Specifically, naïve application of standard techniques in the aggregated context is vulnerable to the ecological fallacy [Robinson 2009; Goodman 1953], wherein conclusions drawn from aggregated data can differ significantly from inferences at individual level, and are misleading to researchers/policy makers using the data.

As a first work on parameter recovery from aggregated data, we investigate the problem for regression in the case of linear models, where the mapping between input features and the output variable is defined by a vector parameter. We consider the scenario, very common in domains like healthcare, sociological studies, etc., where data is collected and aggregated within groups, e.g., patient records aggregated at county or hospital level, and empirical estimates of true group level moments for features and targets are the only available information.

Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48. Copyright 2016 by the author(s).

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

While this problem is relatively easy to handle in the non-aggregated setup, parameter recovery becomes highly challenging when only aggregated data is available and the resulting linear systems are under-determined. Well known works on compressed sensing [Donoho & Elad 2003; Candes & Tao 2005] have shown that recovery is still possible from such systems when the parameter is sparse (common in many applications of interest, e.g. in healthcare where interpretability is part of the desiderata), but existing analyses do not apply directly to the aggregated case.

Our work is motivated by the question: "Is it possible to infer the individual-level parameter of a linear model given aggregated data?" Surprisingly, we answer this question in the affirmative, and to our knowledge, ours is the first such work. We use techniques that exploit structural properties of the data aggregation procedure and show that under standard incoherence conditions on the matrix of true group level moments, the true parameter is recoverable with high probability.

The key contributions of this paper are summarised below:

1. to our knowledge we are the first to investigate the problem of recovery of the sparse population parameter of a linear model when both target variables as well as features are aggregated as sample moments. We provide a theoretical analysis showing that under standard conditions, the parameter can be recovered exactly with high probability.
2. we extend the analysis to capture approximation effects such as sample estimates of the population moment, additive noise, and histogram aggregated targets, showing that the population parameter is recoverable in these scenarios.
3. in the bigger picture, our work extends existing results in the compressed sensing literature by providing guarantees for exact and approximate parameter recovery for the case when the noise in the sensing matrix and measurement vector are linearly correlated, which may be of independent interest.

Experimental results on synthetic data are provided in support of these theoretical claims. We also show that the estimated parameter approaches the predictive accuracy of parameter estimation from non-aggregated or "individual-level" samples when applied to two real world healthcare applications - predictive modeling of reimbursement on CMS Medicare data, and estimation of healthcare charges using Texas State hospital billing records.

2. Parameter Recovery from Exact Means

Let $\mathbf{x} \in \mathbb{R}^d$ represent features and $y \in \mathbb{R}$ represent the target variables, drawn independently from a joint distribution $(\mathbf{x}, \mathbf{y}) \sim \mathcal{P}$. We assume a linear model where each feature is related to the target y via some parameter $\boldsymbol{\beta}_0 \in \mathbb{R}^d$ with noise ϵ as

$$y = \mathbf{x}^\top \boldsymbol{\beta}_0 + \epsilon \tag{1}$$

where ϵ represents observation noise assumed zero mean $E[\epsilon] = 0$ without loss of generality. In the standard regression setting, data is observed at the individual level in the form of n pairs of targets and their corresponding features as $\mathbb{D}_{(x,y)} = \{(\mathbf{x}_i, y_i) : i = 1, 2, \ldots n\}$, so $\boldsymbol{\beta}_0$ may be estimated using standard techniques. Instead, we assume that the inputs $\mathbb{D}_x = \{\mathbf{x}_i : i = 1, 2, \cdots n\}$ and the targets $\mathbb{D}_y = \{y_l : l = 1, 2, \cdots n\}$ are subject to an aggregation process (not controlled by the learner) that produces summaries. In particular, we focus on an aggregation procedure that produces means or first order moments of the data[2].

We consider the case when this aggregation procedure is applied separately to k subgroups of the population. This is common in many domains, e.g., in healthcare, such groups may refer to patient data aggregated by ward, or by hospitals, or based on administrative units like HRR's or HSA's. Similarly, the natural grouping could be demographic information for GSS data and topological clustering for sensor networks.

We assume that the grouping is fixed, and data associated with each group $j \in \{1, 2, \cdots k\}$ is drawn independently from a possibly group-dependent distribution $(\mathbf{x}, y)_j \sim \mathcal{P}_j$ with their own corresponding group-dependent means for covariates/features $\{\boldsymbol{\mu}_j = E_{\mathcal{P}_j}[\mathbf{x}],\ j = 1, \cdots, k\}$ and targets $\{\nu_j = E_{\mathcal{P}_j}[y],\ j = 1, \cdots, k\}$.

We also assume that the model parameter of interest $\boldsymbol{\beta}_0$ is shared by the entire population. By the distributive property of inner products and linearity of the expectation operator, any $\boldsymbol{\beta}_0$ consistent with the data satisfies the set of equations $\boldsymbol{\mu}_j^\top \boldsymbol{\beta}_0 = \nu_j \ \forall\ j = 1, 2, \cdots, k$. Let $\mathbf{M} = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots \boldsymbol{\mu}_k]^\top \in \mathbb{R}^{k \times d}$ be the matrix of feature means, and $\mathbf{y} = [\nu_1, \nu_2, \cdots \nu_k]^\top \in \mathbb{R}^k$ is the vector of target means, it follows from eq. (1) that $\boldsymbol{\beta}_0$ satisfies

$$\mathbf{M}\boldsymbol{\beta}_0 = \mathbf{y}. \tag{2}$$

Clearly, if $k \geq d$ and the rank of $\mathbf{M}$ is greater than d, then (2) is sufficient to characterize $\boldsymbol{\beta}_0$. The more interesting case, and a more practical scenario, is when $k \ll d$, that is, the dimensionality of the problem is much larger than the number of subgroups. We defer to compressed sensing approaches to estimate $\boldsymbol{\beta}_0$ from such systems.

[2] a discussion on higher order moments is presented in the supplementary material

2.1. Estimation from True Means using Compressed Sensing

The compressed sensing literature includes several theoretical and empirical results on the recovery of model parameters in under-determined systems. A line of work including [Candes & Tao 2006; Donoho 2006], among others, have shown that subject to certain sparsity conditions on β_0 and *restricted isometry* constraints on the matrix $\mathbf{M}$, the parameter β_0 can be recovered.

Definition 2.1. *For a $k \times d$ matrix $\mathbf{M}$ and a set $T \subseteq \{1, 2, \cdots, d\}$, suppose $\mathbf{M}_T$ is the $k \times |T|$ matrix consisting of the columns of $\mathbf{M}$ corresponding to T. Then, the* s-restricted isometry constant δ_s *of the matrix $\mathbf{M}$ is defined as the smallest quantity δ_s such that the matrix $\mathbf{M}_T$ obeys*

$$(1 - \delta_s)\|c\|_2^2 \leq \|\mathbf{M}_T c\|_2^2 \leq (1 + \delta_s)\|c\|_2^2$$

for every subset $T \subset \{1, 2, \cdots, d\}$ of size $|T| < s$ and all real $c \in \mathbb{R}^{|T|}$

Restricted isometry is a common and standard assumption in the sparse parameter recovery literature. Intuitively, this property means that when $\mathbf{M}$ satisfies Definition 2.1 with a small δ_s, every sub-matrix of small enough size constructed out of the columns of the matrix behaves approximately like an orthonormal system. In fact, a number of random matrices satisfy this property including the Gaussian ensemble and the Bernoulli ensemble [Donoho 2006; Candès et al. 2006].

For the rest of the paper we assume that the matrix of true means $\mathbf{M}$ satisfies the restricted isometry property. This is quite general as it is a direct corollary for many kinds of common and standard assumptions on the true mean matrix, for example the assumption that the true mean matrix is generated from a Gaussian distribution. Evidence from health care literature [Armstrong et al. 1999; Robinson 2009] suggests that indeed, there is a significant geographical variation in demographics and health outcomes (due to variations in demographic make-up, average economic status, prevalent industries, etc.) which is often used as a predictive feature for healthcare models [Park & Ghosh 2014; Bhowmik et al. 2015]. All of this, together with our experiments on real datasets, suggest that there is sufficient inhomogeneity in mean healthcare attributes across groups to justify the matrix incoherence assumption for $\mathbf{M}$.

Suppose we had access to the true mean matrices $(\mathbf{M}, \mathbf{y})$. First, we consider the case when observations are noise-free, i.e. $\epsilon = 0$. Suppose β_0 is known to be κ_0-sparse and $\mathbf{M}$ satisfies the restricted isometry hypothesis, then the following result applies:

Theorem 2.1 (Exact Recovery [Foucart 2010]). *Let $\Theta_0 = \frac{3}{4+\sqrt{6}} \approx 0.465$. If there exists an s_0 such that $\delta_{2s_0} < \Theta_0$ for $\mathbf{M}$, then as long as $\kappa_0 < s_0$, the constraint $\mathbf{M}\beta_0 = \mathbf{y}$ is sufficient to uniquely recover any κ_0-sparse β_0 exactly as the solution of the following optimization problem:*

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1 \quad \text{s.t. } \mathbf{M}\boldsymbol{\beta} = \mathbf{y}. \tag{3}$$

A similar result for approximate recovery holds for the case when the observations are corrupted with noise $\boldsymbol{\epsilon}$, i.e., instead of $\mathbf{y} = \mathbf{M}\beta_0$, we are given $\mathbf{y}_\epsilon = \mathbf{M}\beta_0 + \boldsymbol{\epsilon}$.

Theorem 2.2 (Approximate Recovery [Candes 2008]). *Let $\Theta_1 = \sqrt{2} - 1 \approx 0.414$. If there exists an s_0 for $\mathbf{M}$ such that $\delta_{2s_0} < \Theta_1$, then as long as $\kappa_0 < s_0$ and the noise $\boldsymbol{\epsilon}$ in observations $\mathbf{y}_\epsilon = \mathbf{M}\beta_0 + \boldsymbol{\epsilon}$ is bounded as $\|\boldsymbol{\epsilon}\|_2 < \xi$, any κ_0-sparse β_0 can be recovered within an ℓ_2 distance of $C_{s_0}\xi$ from the true parameter β_0 using the noisy measurements $(\mathbf{M}, \mathbf{y}_\epsilon)$. That is, the solution $\hat{\boldsymbol{\beta}}$ to the following optimization problem:*

$$\min_{\boldsymbol{\beta}_0} \|\boldsymbol{\beta}\|_1 \text{ s.t. } \|\mathbf{M}\boldsymbol{\beta} - \mathbf{y}_\epsilon\|_2 < \xi \tag{4}$$

satisfies $\|\hat{\boldsymbol{\beta}} - \beta_0\|_2 < C_{s_0}\xi$ where the constant C_{s_0} depends only on δ_{2s_0} and is well-behaved (for example when $\delta_{2s_0} = 0.2$, the constant is less than 8.5).

2.2. Empirical Mean Estimates and Aggregation Error

Clearly, if the matrix of true means $\mathbf{M}$ satisfies the restricted isometry hypothesis, and β_0 is sufficiently sparse, Theorems 2.1 and 2.2 apply. Therefore, given the true population means $\mathbf{M}$ and $\mathbf{y}$, the parameter β_0 can be recovered exactly from noiseless data $\mathbf{y}$ by solving (3) and approximately from noisy observations by solving (4).

Unfortunately, in many practical scenarios we do not have access to the true $\mathbf{M}$ or $\mathbf{y}$, but only to group level empirical estimates computed from a finite number of samples. Assume n samples for each group to simplify the analysis. Denote the corresponding empirically estimated means for the j^{th} group by $\hat{\boldsymbol{\mu}}_{j,n}$ and $\hat{\nu}_{j,n}$ for each $j = 1, \cdots k$. The corresponding sample mean matrices are given by $\widehat{\mathbf{M}}_n = [\hat{\boldsymbol{\mu}}_{1,n}, \cdots \hat{\boldsymbol{\mu}}_{k,n}]^\top$ and $\hat{\boldsymbol{\upsilon}}_n = [\hat{\nu}_{1,n}, \cdots \hat{\nu}_{k,n}]^\top$.

The empirical mean estimation procedure introduces aggregation errors $\mathbf{E}_n$ and $\mathbf{s}_n$ to the setup. That is instead of the true group means $(\mathbf{M}, \mathbf{y})$, the data available for estimating β_0 are restricted to empirical estimates $(\widehat{\mathbf{M}}_n, \hat{\boldsymbol{\upsilon}}_n)$ where $\widehat{\mathbf{M}}_n = \mathbf{M} + \mathbf{E}_n$ and $\hat{\boldsymbol{\upsilon}}_n = \mathbf{y} + \mathbf{s}_n$, and the results from section 2.1 no longer apply directly. For the rest of the manuscript, we investigate parameter recovery for this scenario.

3. Parameter Recovery from Approximate Means

As mentioned earlier, the aggregation procedure for the estimation of true means introduces additive error terms $\mathbf{E}_n$ and $\mathbf{s}_n$ to the matrices $\mathbf{M}$ and $\mathbf{y}$. Note that for the models we study in this work, these two noise terms are not independent but are linearly correlated. Existing compressed sensing literature is restricted to the analysis of models where the additive error terms $\mathbf{E}_n$ and $\mathbf{s}_n$ are independent. Furthermore, any such existing analysis that deals with additive error terms are severely limited in the sense that they can only provide guarantees for approximate recovery rather than exact recovery (e.g. see [Zhao & Yu 2006; Rosenbaum et al. 2013; Rudelson & Zhou 2015]).

Remarkably, as we show in the subsequent sections the true parameter is still exactly recoverable with high probability, even in the presence of linearly correlated aggregation error. This is because the aggregation procedure applied to linear models generates additional structure, which can then be exploited by the estimation procedure to get exact parameter recovery even from empirical estimates of the data means from a finite number of samples.

We first analyse the case where the aggregation procedure has been applied to noise-free samples and then extend the analysis to the noisy case, and to the special case of data collected as histogram aggregates.

Throughout this manuscript we shall make the standard assumption [Georgiou & Kyriakakis 2006; Hsu et al. 2012] that the marginal distribution of each coordinate of the covariates is sub-Gaussian with parameter σ^2. Thus, for each covariate $x_j^{(i)} \in \mathbf{x}_j = [x_j^{(1)}, x_j^{(2)} \cdots x_j^{(d)}]$ and each group $j \in \{1, 2, \cdots k\}$, and for every $t \in \mathbb{R}$, the logarithm of the moment generating function is quadratically bounded

$$\ln E[e^{t(x_j^{(i)} - \mu_j^{(i)})}] < \frac{t^2\sigma^2}{2}.$$

Similarly, we assume that the marginal distribution for each noise term is zero-mean and sub-Gaussian with parameter ρ. Note that the assumptions on the covariates and the noise terms are only on the marginal distributions. In particular, we do not require either independence or identical distribution across groups or even across individual coordinates. As discussed in section 5.1, the analysis for alternative distributional assumptions follows along very similar lines by using other standard concentration inequalities. Proofs for all subsequent results are presented in the supplement.

3.1. Noise-Free Observations

First we consider empirical means computed from noiseless observations. As mentioned earlier, the true parameter $\boldsymbol{\beta}_0$ can still be recovered exactly from empirical estimates of group means $(\widehat{\mathbf{M}}_n, \hat{\boldsymbol{\upsilon}}_n)$ despite the presence of linearly correlated aggregation error $(\mathbf{E}_n, \mathbf{s}_n)$.

Key observation: For a linear model, the relationship satisfied by the true group means $E[y] = E[\mathbf{x}]^\top \boldsymbol{\beta}_0$ is also exactly satisfied by the empirically estimated means $\frac{\sum y}{n} = \left(\frac{\sum \mathbf{x}}{n}\right)^\top \boldsymbol{\beta}_0$. Therefore, for aggregated noise-free observations, the equation

$$\widehat{\mathbf{M}}_n \boldsymbol{\beta}_0 = \hat{\boldsymbol{\upsilon}}_n \tag{5}$$

still holds exactly. As long as the empirical moment matrix $\widehat{\mathbf{M}}_n$ satisfies the restricted isometry constraints, we may still guarantee exact recovery by solving the optimization problem:

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & \|\boldsymbol{\beta}\|_1 \\ \text{s.t.} \quad & \widehat{\mathbf{M}}_n \boldsymbol{\beta} = \hat{\boldsymbol{\upsilon}}_n. \end{aligned} \tag{6}$$

Our first main result is to show that this is indeed the case, and the true parameter $\boldsymbol{\beta}_0$ can be recovered with high probability if the number of samples n used to compute empirical moment estimates in each subgroup is sufficiently large.

Theorem 3.1 (Main result 1). *Let $\Theta_0 = \frac{3}{4+\sqrt{6}} \approx 0.465$. Suppose there exists an s_0 such that the isometry constant δ_{2s_0} for the true mean matrix $\mathbf{M}$ satisfies $\delta_{2s_0} < \Theta_0$. Also suppose that the marginal distribution of the coordinates of each feature is sub-Gaussian with parameter σ^2. Then, given $(\widehat{\mathbf{M}}_n, \hat{\boldsymbol{\upsilon}}_n)$ any κ_0-sparse $\boldsymbol{\beta}_0$ with $\kappa_0 < s_0$ can be recovered exactly with probability at least $1 - e^{-C_0 n}$ by solving (6). Here, the constant C_0 in the expression is such that $C_0 \sim O\left(\frac{(\Theta_0 - \delta_{2s_0})^2}{kd\sigma^2(1+\delta_{2s_0})}\right)$.*

We can unpack the result with respect to the constant C_0 which depends on the isometry parameter δ_{2s_0}, the size of the mean matrix (k, d) and the sub-Gaussian parameter of the feature terms σ. The robustness of the isometry property of $\widehat{\mathbf{M}}_n$ depends on the strength of the isometry property in the true moment matrix $\mathbf{M}$. Fewer samples are required for estimating $\widehat{\mathbf{M}}_n$ if $\mathbf{M}$ satisfies the isometry hypothesis more robustly (that is, δ_{2s_0} small) and consequently, a larger value of $\frac{(\Theta_0 - \delta_{2s_0})^2}{1+\delta_{2s_0}}$. Similarly, if the feature distributions have a thinner tail i.e. a smaller value of the sub-Gaussian parameter σ^2, empirically estimated means are more accurate with fewer samples.

3.2. Observations with Noise

We now consider the case when the observations are noisy and the equation (5) no longer holds exactly. In particular, we assume that the data used to compute the sample moments is observed with zero mean additive noise as $y_{i,j}^{\epsilon} = \mathbf{x}_{i,j}^\top \boldsymbol{\beta}_0 + \epsilon_{i,j}$ for each datapoint $i \in \{1, \cdots, n\}$ in population subgroup $j \in \{1, \cdots, k\}$. This leads to an error in the empirical target means over and above the aggregation error.

Let $\hat{\boldsymbol{\upsilon}}_{n,\epsilon} = \hat{\boldsymbol{\upsilon}}_n + \boldsymbol{\epsilon}_n$ where $\hat{\boldsymbol{\upsilon}}_{n,\epsilon}$ (henceforth denoted $\hat{\boldsymbol{\upsilon}}_\epsilon$) is the empirical target mean estimated from noisy samples and $\boldsymbol{\epsilon}_n$ is the cumulative estimation error due to noise in

n samples. With the feature sample mean $\widehat{\mathbf{M}}_n$, eq. (5) becomes

$$\widehat{\mathbf{M}}_n\boldsymbol{\beta} = \hat{\boldsymbol{v}}_n = \hat{\boldsymbol{v}}_\epsilon - \boldsymbol{\epsilon}_n. \tag{7}$$

Similar to the results of Theorem 2.2, it can be expected that if the sample mean matrix $\widehat{\mathbf{M}}_n$ satisfies the isometry hypothesis for noisy measurements, and if the error term $\boldsymbol{\epsilon}_n$ is bounded as $\|\boldsymbol{\epsilon}_n\|_2 < \xi$ for some $\xi > 0$, then $\boldsymbol{\beta}_0$ can be recovered to within an ℓ_2 distance of $O(\xi)$ by solving the following optimization problem

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & \|\boldsymbol{\beta}\|_1 \\ \text{s.t.} \quad & \|\widehat{\mathbf{M}}_n\boldsymbol{\beta} - \hat{\boldsymbol{v}}_\epsilon\|_2 < \xi. \end{aligned} \tag{8}$$

In fact, in our case we can show that the aggregation procedure smooths out the destabilising effects of noise in observations to allow arbitrarily accurate parameter recovery within any small degree ξ of ℓ_2 estimation error.

Theorem 3.2 (Main Result 2). *Let $\Theta_1 = \sqrt{2} - 1 \approx 0.414$. Suppose there exists an s_0 such that the isometry constant δ_{2s_0} for the true mean matrix $\mathbf{M}$ satisfies $\delta_{2s_0} < \Theta_1$. Also suppose that the marginal distribution of the coordinates of each feature is sub-Gaussian with parameter σ^2, and noise in each observation is zero-mean and sub-Gaussian with parameter ρ^2. Let $\xi > 0$ be any small positive real value. Then, any κ_0-sparse $\boldsymbol{\beta}_0$ with $\kappa_0 < s_0$ can be recovered within an ℓ_2 distance of $O(\xi)$ with probability at least $1 - e^{-C_1 n} - e^{-C_2 n}$ by solving (8). Here, the constant C_1 is such that $C_1 \sim O\left(\frac{(\Theta_1 - \delta_{2s_0})^2}{kd\sigma^2(1+\delta_{2s_0})}\right)$ and the constant C_2 is such that $C_2 \sim O\left(\frac{\xi^2}{\rho^2 k}\right)$.*

The constant term in $O(\xi)$ is the same as that in Theorem 2.2 and it depends only on δ_{2s_0} and is well-behaved for small values of δ_{2s_0}. Note the similarity of the constant C_1 in the noisy case and the constant C_0 in the exact case. As for exact recovery, the probability of recovery depends on the tail properties of the feature distribution as well as the robustness of the isometry property for the true mean matrix $\mathbf{M}$. The constant $\frac{\xi^2}{\rho^2 k}$ in the additional term accounts for observational noise. As expected, more samples are required if the noise has heavy tails ρ^2 or if the degree of approximation ξ is small. In addition, the constant for $O(\xi)$ in the approximation factor may depend only δ_{2s_0} in a manner similar to Theorem 2.2.

3.3. Extension to Histogram Aggregation

For the preceding analysis, we have assumed that errors in the target moments is a result of the empirical aggregation or observational noise. It is worth noting that this analysis can be extended to cover any additional source of error which can be bounded deterministically or with high probability. An example of this is when the targets are available as histogram aggregates with bin size Δ and the mean is estimated from the histogram. Suppose h_Δ is the error in estimation of target mean from the histogram such that the estimated sample mean $\hat{\boldsymbol{v}}_\Delta$ is related to the true sample mean for the targets as $\hat{\boldsymbol{v}}_\Delta = \hat{\boldsymbol{v}}_n + h_\Delta$.

Then, we can use the exact same procedure as for noisy observations to bound the ℓ_2 error in estimation of $\boldsymbol{\beta}_0$ to $O(\xi_\Delta)$ by solving the optimisation problem

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & \|\boldsymbol{\beta}\|_1 \\ \text{s.t.} \quad & \|\widehat{\mathbf{M}}_n\boldsymbol{\beta} - \hat{\boldsymbol{v}}_\Delta\|_2 < \xi_\Delta \end{aligned} \tag{9}$$

for some positive $\xi_\Delta > 0$.

The value of ξ_Δ and theoretical guarantees arising therefrom will depend on the manner in which the target mean in estimated from the histogram. Here, we analyse one such standard moment estimation approach.

Consider a single population subgroup. Suppose the range of the targets is bounded by some R, that is, $y_{\max} - y_{\min} < R$. We have a set of bins $\mathcal{B} = \{B_\tau = (b_\tau, b_{\tau+1}) : \tau = 1, 2, \cdots, \lfloor \frac{R}{\Delta} \rfloor\}$ such that $b_{\tau+1} - b_\tau = \Delta$ for each bin. We also have for each bin an integer n_τ which is the number of targets for that subgroup that fall in that particular bin. Suppose $\bar{b}_\tau = \frac{(b_\tau + b_{\tau+1})}{2}$ is the mid point of each bin. Then, the target mean for that group is estimated as

$$\hat{\nu}_\Delta = \frac{\sum_\tau n_\tau \bar{b}_\tau}{\sum_\tau n_\tau} = \frac{\sum_\tau n_\tau \bar{b}_\tau}{n}.$$

For this mean imputation procedure, we get a very similar result to Theorem 3.2 for aggregated data that bounds the probability of recovery in terms of the isometry constants of the true mean matrix and the granularity of the histogram.

Theorem 3.3 (Main Result 3). *Let $\Theta_1 = \sqrt{2} - 1 \approx 0.414$. Suppose there exists an s_0 such that the isometry constant δ_{2s_0} for the true mean matrix $\mathbf{M}$ satisfies $\delta_{2s_0} < \Theta_1$.Also suppose that each covariate has a sub-Gaussian distribution with parameter σ^2. Let the targets for each group be available as histogram aggregates with bin size bounded below by Δ. Then, any κ_0-sparse $\boldsymbol{\beta}_0$ with $\kappa_0 < s_0$ can be recovered within an ℓ_2 distance of $O(\sqrt{k}\Delta)$ with probability at least $1 - e^{-C_1 n}$ by solving (9) with $\xi_\Delta = \sqrt{k}\frac{\Delta}{2}$. Here, the constant C_1 is such that $C_1 \sim O\left(\frac{(\Theta_1 - \delta_{2s_0})^2}{kd\sigma^2(1+\delta_{2s_0})}\right)$.*

Note that the constants on $O(\sqrt{k}\Delta)$ are the same as in the case of noisy observations. Also, in the case of exact estimation, bin size $\Delta \to 0$, therefore $\boldsymbol{\beta}_0$ can be recovered exactly. Furthermore, the bin size does not have any effect on the sample complexity of recovery probability, only on the accuracy of estimation.

In particular, the recovery error is small for a histogram of fine enough granularity. In most cases of binned data, the bin size used for reporting the histogram decreases as a function of n. In fact for many real world scenarios (see [Scott 1979]) the bin size decreases at least as fast as $\Delta = O(\frac{1}{n^c})$ for some $0 < c < 1$. In any case, the worst case error in parameter estimation is limited solely by the bin size, and tighter bounds can be obtained by making reasonable assumptions on the target distribution. Note that if instead of supplying a coarse histogram the data is released in full (without specifying the relationship between $\mathbf{x}$ and $\mathbf{y}$ in each group), the effective bin size is 0 and the parameter can be estimated exactly by Theorem 3.3.

Related Work

While there is a rich literature on sparse parameter recovery and predictive modeling in general, the aggregated data case is much more limited. To our knowledge, ours is the first analysis of sparse parameter recovery for aggregated data of *any* kind, and our main results have not been shown in more than 60 years of ecological data analysis dating at least to Goodman [Goodman 1953], with parallel work in the compressed sensing literature, and renewed interest in machine learning [Park & Ghosh 2014; Bhowmik et al. 2015]. We now briefly outline the relevant literature.

Data aggregation was studied in the context of privacy preservation by [Park & Ghosh 2014] which used clustering and low rank models for data reconstruction from averaged targets. In the classification literature, learning from label proportions (LLP) [Quadrianto et al. 2009; Patrini et al. 2014] involves estimation of classifiers given the proportion of discrete valued labels in groups or bags of labeled targets. Regression involving histogram aggregated targets was introduced by [Bhowmik et al. 2015] which introduced an estimation algorithm and evaluated it empirically, but did not provide a theoretical analysis.

There are several differences between our work and the works described above. First and most importantly, all three of the aforementioned lines of work assumed aggregation only in targets and studied a setup where features are known unaggregated at individual level. In our work, both targets and features are aggregated. Unlike our work, [Park & Ghosh 2014] was focused on data reconstruction rather than predictive modeling. Next, the LLP literature looks at classification given discrete-values targets, while we look at regression where targets can take arbitrary values. Furthermore, unlike [Bhowmik et al. 2015], our work provides a rigorous theoretical analysis with recovery guarantees. Finally, all existing lines of work are concerned with accurate prediction, and to our knowledge there have been no studies of sparse parameter recovery.

The techniques used in our work follows a long line of research on compressed sensing as discussed in Section 2.1, where related analyses fall mainly under three categories:

1. error in the design matrix $\widehat{\mathbf{M}} = \mathbf{M} + \mathbf{E}$, without any error or noise in observation vector $\mathbf{y}$
2. noise in observations $\hat{\boldsymbol{\upsilon}} = \mathbf{y} + \mathbf{s}$, with a fixed design matrix $\mathbf{M}$ without error
3. design matrix error $\mathbf{E}$ and observation noise $\mathbf{s}$, where $\mathbf{E}$ and $\mathbf{s}$ are independent

Prior work, eg.[Herman & Strohmer 2010; Zhao & Yu 2006; Rudelson & Zhou 2015], deals only with case 1, or with cases 2 and 3 in a way to only provide *approximate* parameter recovery guarantees. We focus our investigation on the aggregated data case 4: where $\mathbf{E}$ and $\mathbf{s}$ are linearly correlated. Even ignoring the linear correlation in the noise model, the best existing analyses are still limited to using a naive error bounding technique to analyse the stability of the LASSO resulting in weak guarantees for only approximate parameter recovery.

In contrast, we propose non-trivial modifications to the analysis, and are able to exploit the additional structure generated by the data aggregation procedure to recover the sparse parameter *exactly* even with aggregation error, as in Theorem 3.1, and upto arbitrarily accurate degree of estimation from noisy data as we see in Theorems 3.2 and 3.3.

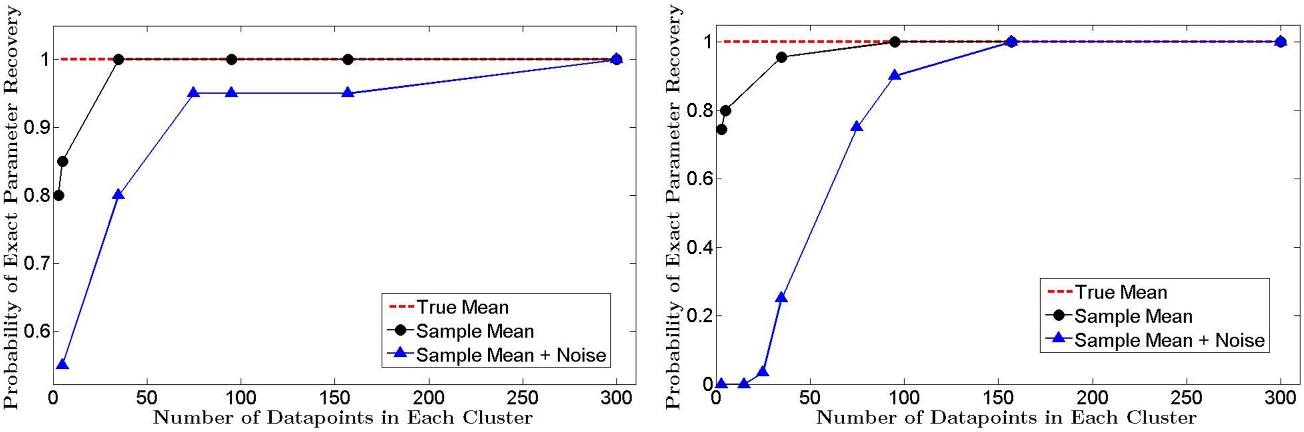
4. Experiments

We corroborate our theoretical results with experiments on synthetic data to show that probability of exact parameter recovery follows a pattern just as predicted by our main results. We also demonstrate the efficacy of our technique in two real world applications by applying it to predictive modeling of outpatient reimbursement claims in CMS Medicare data (DE-SynPUF), and to modeling healthcare costs using Texas Inpatient Discharge dataset (TxID) from the Texas Department of State Health Services.

4.1. Synthetic Data

We first generate the true covariate mean matrix $\mathbf{M}$ using a Gaussian and a Bernoulli ensemble, and compute the respective true target means using a sparse $\boldsymbol{\beta}_0$. We then generate random covariates centred around the true mean matrix and compute the corresponding empirical mean matrix $\widehat{\mathbf{M}}_n$ from the covariates. The targets are then generated using the parameter $\boldsymbol{\beta}_0$. We consider two cases separately- noiseless targets $\mathbf{y}$ and targets $\mathbf{y}_\epsilon$ to which noise has been added. The corresponding empirical target means $\hat{\boldsymbol{\upsilon}}_n$ and $\hat{\boldsymbol{\upsilon}}_\epsilon$ are computed for both sets of targets and used together with the sample covariate means $\widehat{\mathbf{M}}_n$ to estimate $\boldsymbol{\beta}_0$.

This entire procedure is repeated multiple times and the proportion of instances in which the true parameter $\boldsymbol{\beta}_0$ is recovered exactly, both in magnitude and support, is plotted against the number of datapoints used to compute the



(a) Probability of recovering exact parameter

(b) Probability of recovering exact parameter

Figure 1: Probably of exact parameter recovery (both magnitude and signed support) on Gaussian (fig 1a) and Bernoulli (fig 1b) models for noise-free (black) and noisy (blue) observations with increasing number of datapoints in each group

empirical sample means. Figures 1a and 1b show the results for Gaussian and Bernoulli ensembles respectively. As can be seen in the figures, the probability of recovering the exact parameter increases as the number of data points used to compute the empirical sample means increases, in a manner exactly as predicted by our theoretical results.

4.2. Real datasets - DE-SynPUF and TxID

We now apply our methods to two real datasets. Since ours is the first work on sparse recovery from aggregated data, we do not know of any competing algorithmic baselines. We evaluate our methods by comparing the parameter estimated from aggregated data to the performance upper bound of the “true” parameter that is estimated from the full non-aggregated dataset.

Our first dataset is the CMS Beneficiary Summary (DE-SynPUF) dataset [DESynPUF 2008] which is a public use dataset created by the Centers for Medicare and Medicaid Services and is often used for testing different data mining or statistical inferential methods before getting access to full Medicare data. We use a subset of the DE-SynPUF dataset for Louisiana state from the year 2008 and model outpatient institutional annual primary payer reimbursement (*PPPYMT-OP*) with all the available predictor variables that include age, race, sex, duration of coverage, presence/absence of a variety of chronic conditions, etc.

Our second dataset is the Texas Inpatient Discharge dataset (TxID) from the Texas Department of State Health Services ([TxID 2014], see also [Park & Ghosh 2014]). We model healthcare charges using hospital billing records from the fourth quarter of 2006 in the TxID dataset, and use all the available individual level predictor variables, which include demographic information like race, and real valued variables like length of hospital stay for each datapoint.

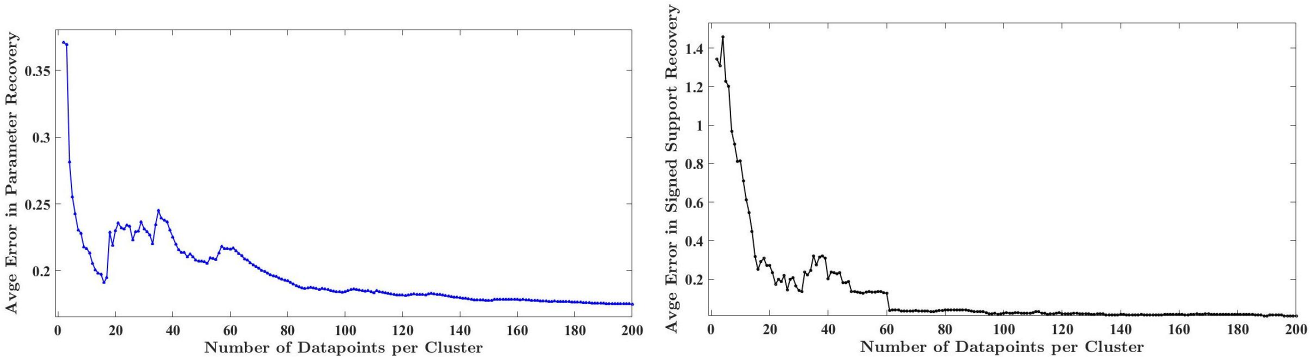
In both these datasets, we first use a LASSO estimator (with parameter chosen via cross-validation) on the full dataset to obtain a sparse regression parameter $\boldsymbol{\beta}_{full}$. We use a k-means algorithm to cluster the datapoints into groups and compute the sample means for each group with increasing number of datapoints. We then use only these empirical sample means to obtain an estimate $\boldsymbol{\beta}_{agg}$ for the parameter, and compare $\boldsymbol{\beta}_{agg}$ to the parameter $\boldsymbol{\beta}_{full}$ obtained from full non-aggregated dataset. Results averaged across multiple clusterings are shown in figures 2 and 3.

Figures 2a and 3a show the ℓ_2 norm of the distance between the parameter estimated from the full dataset $\boldsymbol{\beta}_{full}$ and the parameter estimated from the aggregated version $\boldsymbol{\beta}_{agg}$, for the DE-SynPUF dataset and TxID dataset respectively, plotted against the number of datapoints used to estimate the means. Figure 2b and 3b show the number of conflicts or discrepancies between the support (non-zero coordinates) of $\boldsymbol{\beta}_{agg}$ estimated from aggregated data and support of $\boldsymbol{\beta}_{full}$ estimated from the non-aggregated dataset, for the DE-SynPUF dataset and TxID dataset respectively. As the number of datapoints used to compute the sample means increases, the parameter recovered using aggregated data exactly identifies the support of the “true” parameter estimated from the full dataset, and also closely matches it in magnitude.

5. Discussion

5.1. Extensions

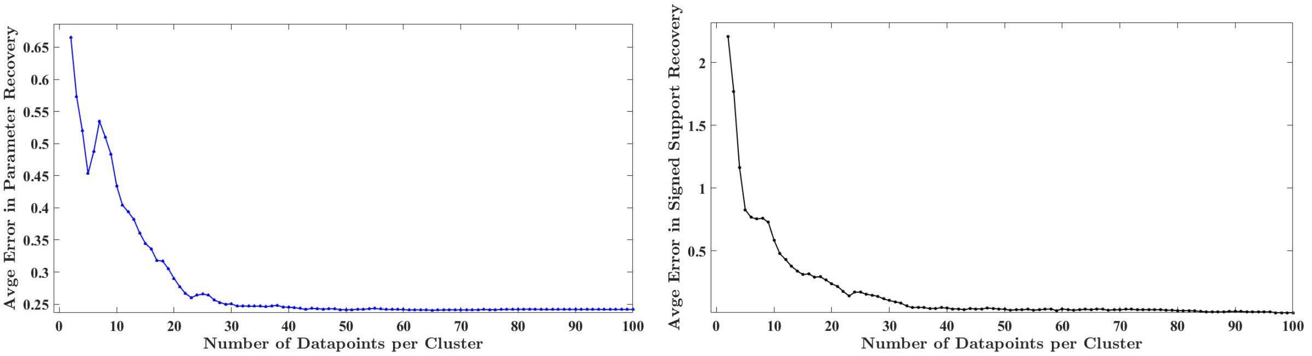
The techniques presented in this work can be applied to the parameter recovery problem in a much wider class of cases of interest by building on and extending existing re-



(a) Avg. Error in Parameter Recovery $||\boldsymbol{\beta}_{full} - \boldsymbol{\beta}_{agg}||$

(b) Avg. Error in Support Recovery

Figure 2: Performance on DESynPUF dataset with increasing number of datapoints in each group



(a) Avg. Error in Parameter Recovery $||\boldsymbol{\beta}_{full} - \boldsymbol{\beta}_{agg}||$

(b) Avg. Error in Support Recovery

Figure 3: Performance on TxID dataset with increasing number of datapoints in each group

sults in the compressed sensing literature (see [Candes et al. 2006; Candes & Tao 2007; Cai et al. 2010, 2009], etc). In particular, we note that various alternative frameworks like non-sparse $\boldsymbol{\beta}_0$, alternative estimators to LASSO, beyond sub-gaussian assumptions on different marginals, etc. can be analysed in an identical manner, and our main results on parameter recovery would still continue to hold, albeit with slightly different sample complexity.

5.2. Higher Order Moments

The results in this paper focused on estimation from first order moments. It may seem like including higher order moments might make estimation in this framework easier but it turns out that this is not the case in general. We include a discussion in the supplement on the difficulties of using higher order moments for estimation. In particular, we prove a surprising and counter-intuitive negative result which shows that even with second order moments, in the general case the estimation cannot be guaranteed to be easier or more accurate than when we use only first order moments. Similar results may also hold for other higher order moments.

6. Conclusion and Future Work

In this paper we study the problem of parameter recovery for sparse linear models from data which has been aggregated in the form of empirical means computed from different subgroups of the population. We show that when the collection of true group moments is an incoherent matrix, the parameter can be recovered with high probability from the empirical moments alone provided the empirical moments are computed from a sufficiently large number of samples. We extend the framework to the case of moments computed from noisy or histogram aggregated data and show that the parameter can still be recovered within an arbitrarily small degree of error. We corroborate our theoretical results with experiments on synthetic data and also show results on two real world healthcare applications- predictive modeling of reimbursement claims from CMS Medicare data, and modeling healthcare charges using hospital billing records from the Texas Department of State Health Services. For future work, we plan to extend the framework to more general models including GLM's and non-linear models, and to design techniques to incorporate higher order moments in the procedure.

Acknowledgements

Authors acknowledge support from NSF under grant IIS-1421729.

References

Armstrong, Marc P, Rushton, Gerard, and Zimmerman, Dale L. Geographically masking health data to preserve confidentiality. *Statistics in Medicine*, 18(5):497–525, 1999.

Bhowmik, Avradeep, Ghosh, Joydeep, and Koyejo, Oluwasanmi. Generalized Linear Models for Aggregated Data. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pp. 93–101, 2015.

Cai, T Tony, Xu, Guangwu, and Zhang, Jun. On recovery of sparse signals via ℓ_1 minimization. *Information Theory, IEEE Transactions on*, 55(7):3388–3397, 2009.

Cai, T Tony, Wang, Lie, and Xu, Guangwu. Shifting inequality and recovery of sparse signals. *Signal Processing, IEEE Transactions on*, 58(3):1300–1308, 2010.

Candes, Emmanuel and Tao, Terence. The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, pp. 2313–2351, 2007.

Candes, Emmanuel J. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9):589–592, 2008.

Candes, Emmanuel J and Tao, Terence. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12):4203–4215, 2005.

Candes, Emmanuel J and Tao, Terence. Near-optimal signal recovery from random projections: Universal encoding strategies? *Information Theory, IEEE Transactions on*, 52(12):5406–5425, 2006.

Candès, Emmanuel J, Romberg, Justin, and Tao, Terence. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *Information Theory, IEEE Transactions on*, 52(2):489–509, 2006.

Candes, Emmanuel J, Romberg, Justin K, and Tao, Terence. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006.

DESynPUF. Medicare Claims Synthetic Public Use Files (SynPUFs). *Centers for Medicare and Medicaid Services*, 2008. `http://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/index.html`.

Donoho, David L. For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Communications on pure and applied mathematics*, 59(6):797–829, 2006.

Donoho, David L and Elad, Michael. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

Foucart, Simon. A note on guaranteed sparse recovery via ℓ_1-minimization. *Applied and Computational Harmonic Analysis*, 29(1):97–103, 2010.

Georgiou, Panayiotis G and Kyriakakis, Chris. Robust maximum likelihood source localization: the case for sub-gaussian versus gaussian. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(4):1470–1480, 2006.

Goodman, Leo A. Ecological regressions and behavior of individuals. *American Sociological Review*, 1953.

Herman, Matthew A and Strohmer, Thomas. General deviants: An analysis of perturbations in compressed sensing. *Selected Topics in Signal Processing, IEEE Journal of*, 4(2):342–349, 2010.

Hsu, Daniel, Kakade, Sham M, and Zhang, Tong. A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab*, 17(52):1–6, 2012.

Park, Yubin and Ghosh, Joydeep. 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*, pp. 55–64. ACM, 2014.

Patrini, Giorgio, Nock, Richard, Caetano, Tiberio, and Rivera, Paul. (almost) no label no cry. In *Advances in Neural Information Processing Systems*, pp. 190–198, 2014.

Quadrianto, Novi, Smola, Alex J, Caetano, Tiberio S, and Le, Quoc V. Estimating labels from label proportions. *The Journal of Machine Learning Research*, 10:2349–2374, 2009.

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

Rosenbaum, Mathieu, Tsybakov, Alexandre B, et al. Improved matrix uncertainty selector. In *From Probability to Statistics and Back: High-Dimensional Models and*

Processes–A Festschrift in Honor of Jon A. Wellner, pp. 276–290. Institute of Mathematical Statistics, 2013.

Rudelson, Mark and Zhou, Shuheng. High dimensional errors-in-variables models with dependent measurements. *arXiv preprint arXiv:1502.02355*, 2015.

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

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

Wagner, David. Resilient aggregation in sensor networks. In *Proceedings of the 2nd ACM workshop on Security of ad hoc and sensor networks*, pp. 78–87. ACM, 2004.

Zhao, Jerry, Govindan, Ramesh, and Estrin, Deborah. Computing aggregates for monitoring wireless sensor networks. In *Sensor Network Protocols and Applications, 2003*, pp. 139–148. IEEE, 2003.

Zhao, Peng and Yu, Bin. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7: 2541–2563, 2006.