A Probabilistic Imputation Framework for Predictive Analysis using Variably Aggregated, Multi-source Healthcare Data

Yubin Park
Department of Electrical and Computer Engg.
University of Texas at Austin
Austin, TX, USA
yubin.park@utexas.edu

Joydeep Ghosh
Department of Electrical and Computer Engg.
University of Texas at Austin
Austin, TX, USA
ghosh@ece.utexas.edu

ABSTRACT

In healthcare-related studies, individual patient or hospital data are not often publicly available due to privacy restrictions, legal issues or reporting norms. However, such measures may be provided at a higher or more aggregated level, such as state-level, county-level summaries or averages over health zones (HRR[1] or HSA[2]). Such levels constitute partitions of the underlying individual level data, which may not match the data segments that would have been obtained if one clustered individual-level data. Treating these aggregated values as representatives for the individuals can result in the ecological fallacy. How can one run data mining procedures on such data where different variables are available at different levels of aggregation or granularity? We examine this problem in a clustering setting given a mix of individual-level and (arbitrarily) aggregated-level data. For this setting, a generative process of such data is constructed using a Bayesian directed graphical model. This model is further developed to capture the properties of the aggregated-level data using the Central Limit theorem.. The model provides reasonable cluster centroids under certain conditions, and is extended to estimate the masked individual values for the aggregated data. The model parameters are learned using an approximated Gibbs sampling method, which employs the Metropolis-Hastings algorithm efficiently. A deterministic approximation algorithm is derived from the model, which scales up to massive datasets. Furthermore, the imputed features can help to improve the performance in subsequent predictive modeling tasks. Experimental results using data from the Dartmouth Health Atlas, CDC, and the U.S. Census Bureau are provided to illustrate the generality and capabilities of the proposed framework.

[1]Hospital Referral Region
[2]Hospital Service Area

IHI'12, January 28–30, 2012, Miami, Florida, USA.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Probabilistic algorithms; J.3 [**Life and Medical Sciences**]: Health

General Terms

Algorithms

Keywords

Clustering, Privacy Preserving Data Mining, Dartmouth Health Atlas, Multi-source Health Metrics, Ecological Fallacy

1. INTRODUCTION

Despite the tremendous information explosion and availability of public-domain medical and healthcare data recently (for example see www.data.gov/health), many of the health-related features or indicators are only available at a highly aggregated level, due to privacy concerns, reporting norms or legal issues [**?**]. In particular, routinely collected administrative data sets, such as national registers, aim to collect information on a limited number of variables for the whole population, while survey and cohort studies contain more detailed data from a sample of the population [**?**]. Even if the individual records are available, some features may be suppressed to protect identities of data holders. For example, Texas Department of State Health Services provides 'Texas Inpatient Public Use Data File (PUDF)', which contains data on discharges from Texas hospitals [**?**], but the ZIP code information in PUDF is suppressed or eliminated depending on the number of patients in a given region. As data mining algorithms should ideally be applied to individual-level data to discover valuable information, limited access to the raw entries introduces conflict of interests between data miners, patients and providers [**?**]. Several privacy preserving data mining algorithms have been suggested to overcome this conflict [**?**, **?**, **?**, **?**]. However, requirements of privacy preservation are difficult to achieve for several types of analyses, and these algorithms are typically more complex and less capable compared to privacy-agnostic techniques.

Many health or healthcare indicators are available at different aggregated levels, rather than providing an entry for each individual. For example, average income by state, average death ratio by city, or average smoking rate by country are available through a variety of easily accessible public

reports. Although these aggregated statistics cannot reconstruct the underlying individual-level data, these aggregated data can be combined with individual data to produce more informative models. In epidemiology, it has been observed that ecological bias from aggregate administrative data can be alleviated by incorporating surveys of individual exposures or case-control data, leading to recent attempts at integrating data at multiple levels of summarization [**?**, **?**].

In this paper, we seek a better utilization of such aggregated information for augmenting the individual-level data. Assuming that the dataset of interest is generated by a mixture model, and that the partitions that form aggregation units (such as states or counties) contain different ratios of the mixture components, we introduce a novel generative process, which captures the underlying distributions using a Bayesian directed graphical model and the Central Limit Theorem. Despite the limited nature of given aggregated information, our clustering algorithm provides not only reasonable cluster centroids, but also imputes the unobserved individual features. These imputed features reflect the underlying distribution of the data, thus a predictive model using these extended information shows improved performances. As many datasets in the healthcare domain are divided into multiple tables containing different levels of aggregation (sometimes obtained from different sources), the suggested methodology in this paper can be useful in maximizing the utility of such available information. Furthermore, our approach can easily be extended to situations where different features are aggregated over various partitions of the raw data records.

2. RELATED WORK

In this section, we outline three bodies of related work, starting from traditional imputation techniques in statistics. This is followed by ecological study techniques, where aggregated and individual information are both available. Finally, we briefly discuss various approaches that are used to make inferences in Bayesian graphical models.

In statistics, imputation techniques are mainly used to substitute missing values in data [**?**]. A once-common method is cold-deck imputation, where a missing value is imputed from randomly selected similar records from another dataset. More sophisticated techniques, such as the nearest neighbor imputation and the approximate Bayesian bootstrap, have been also developed to supersede this original method. As a special case, when geographical information is missing in data, geo-imputation techniques are widely used, where the imputation is taken from approximate locations derived from associate data [**?**]. However, these traditional techniques are based on individual-level data, and some of them are limited in their applicabilities.

On the other hand, in ecological studies, aggregated information is usually the unit of analysis, as individual information is usually not available due to expensive acquisition costs or legal issues [**?**, **?**, **?**]. Although ecological studies have been used frequently across multiple domains such as social science and healthcare analysis, the validity of the studies is still controversial because of the difference between ecological correlation and individual correlation [**?**], which is also known as the 'ecological fallacy'. Fortuitously, in recent years, it has been reported that auxiliary individual level information can help to reduce the ecological fallacy [**?**]. In the *hierarchical related regression* (HRR) framework, auxiliary individual information represents a small fraction of the individual samples that constitute the aggregate information [**?**, **?**]. This setting is useful when acquisition costs of getting individual data is expensive, so that the available information covers only a small portion of the entire population. The HRR model relates the regression coefficients from both aggregate and individual data, compensating their disadvantages. This analysis has been shown to reduce the ecological bias, but the type of the auxiliary information used in HRR is different from our setting in this paper. The model we present in this paper assumes auxiliary individual information, which contains a different set of features from provided aggregate data. We first focus on a generative process of such data, then derive an inference mechanism to get estimated individual values for the aggregated features. From the generative process, heterogeneity of ecological groups is naturally captured by suitable mixture distributions, resulting in better imputation.

In Bayesian graphical models such as the model presented in this paper, inferential problems pose key challenges in most cases. The Expectation Maximization (EM) algorithm is the most popular approach when latent variables are present in models. However, many sophisticated models such as Latent Dirichlet Allocation (LDA) [**?**] have intractable posterior distributions for latent variables. To approximate the posterior distributions, other techniques such as variational EM algorithm, Gibbs sampling and collapsed Gibbs sampling are employed. Although their computational complexities and assumptions are slightly different, their performances are marginally the same [**?**]. In this paper, we demonstrate an approximated Gibbs sampling approach, which is specialized for our setting. Then we further introduce a deterministic version, which is not only much faster but also scalable to massive datasets.

3. CLUSTERING MODEL

We denote the set of features that are available at the individual level, where "individual" refers to entities at the highest resolution available, by $\vec{x}_o$. The features that are observed only at an aggregated level are denoted by $\vec{x}_u$, where u denotes 'unobserved' at the individual level. Thus there is an underlying "complete" dataset ($\mathcal{D}_x = \{(\vec{x}_o, \vec{x}_u)_1, (\vec{x}_o, \vec{x}_u)_2, ..., (\vec{x}_o, \vec{x}_u)_N\}$), which has all features observed. The data provider only provides the values of observed variables though. In addition, it specifies a set of partitions: $\mathcal{P} = \{\mathcal{D}_x^1, \mathcal{D}_x^2, ..., \mathcal{D}_x^P\}$ where $\bigcup_{p=1}^{P} \mathcal{D}_x^p = \mathcal{D}_x$ and $\mathcal{D}_x^p \bigcap \mathcal{D}_x^q = \emptyset$ for any p, q. These partitions specify the aggregated values provided on the unobserved features ($\vec{x}_u$), $\mathcal{D}_s = \{\vec{s}_1, \vec{s}_2, ..., \vec{s}_P\}$, where $\vec{s}_p$ is derived from $\mathcal{D}_x$ as $\vec{s}_p = \frac{1}{N_p} \sum_{i=1}^{N} \vec{x}_{ui} \mathbf{1}_{(\vec{x}_{ui} \in \mathcal{D}_x^p)}$ (sample mean within $\mathcal{D}_x^p$) and $N_p = |\mathcal{D}_x^p|$. Note that in general, different partitions (and hence levels of aggregation) may apply to different unobserved variables. Though our approach can be readily extended to cover such situations, in this paper we consider a common partitioning to keep the notation and exposition simple.

Suppose we want to find K clusters in the complete data, denoted by $\{\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_K\}$. To cater to the unobserved data, an assumption of conditional independence is made: $p(\vec{x}_o, \vec{x}_u|\mathcal{C}_k) = p(\vec{x}_o|\mathcal{C}_k)p(\vec{x}_u|\mathcal{C}_k)$. Let $\vec{\pi}_p = (p(\mathcal{C}_1|\mathcal{D}_x^p), p(\mathcal{C}_2|\mathcal{D}_x^p), ..., p(\mathcal{C}_K|\mathcal{D}_x^p))^T = (\pi_{p1}, \pi_{p2}, ..., \pi_{pK})^T$, which represents the mixing coefficients of the partition p. Then, to avoid a pathological symmetry case, we assume that $\vec{\pi}_p \neq \vec{\pi}_q$ for

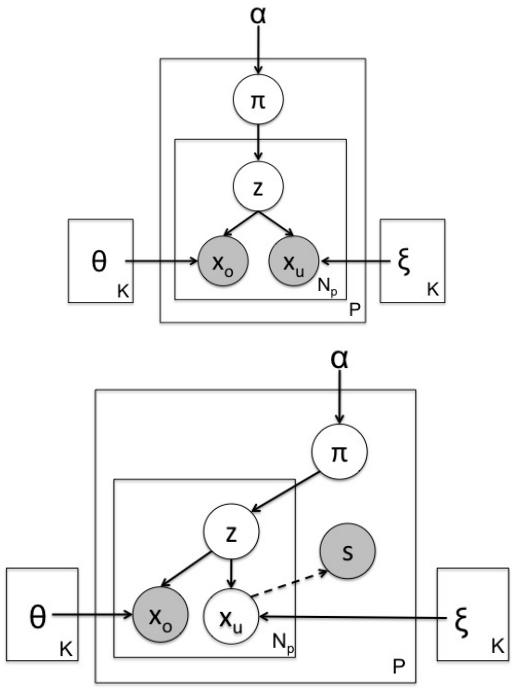

Figure 1: (a) Clustering models when complete data is available (top) and (b) when only aggregates $\vec{s}$ are observed instead of $\vec{x}_u$ (bottom).

any p, q with probability one. Let $\vec{\xi}_k$ and $\vec{\theta}_k$ be the sufficient statistics for the distributions $p(\vec{x}_u|\mathcal{C}_k)$ and $p(\vec{x}_o|\mathcal{C}_k)$ respectively. If all data features are observed at the individual level, a LDA-like clustering model can be built based on the conditional independence assumption as in Figure 1 (a), where $\vec{\pi}$ is sampled from a Dirichlet distribution parametrized by $\vec{\alpha}$. Figure 1 (b) shows a modified clustering model that accommodates the aggregated nature of the unobserved variables. As $\vec{x}_u$ and $\vec{x}_o$ are independent given $\mathcal{C}_k$, they can be separated using different nodes. In the model, $\vec{x}_u$ is not observed; rather the derived (aggregated) features $\vec{s}$ are observed.

Even though the model of Figure 1(b) captures the problem characteristics, it is highly inefficient and contains redundant nodes. Fortunately, the complexity of the model can be reduced by removing the unobserved nodes $\vec{x}_u$'s if N_p is large enough. Let $\vec{\eta}_k$ and $\mathbf{T}_k^2$ be the mean and variance of the distribution, $p(\vec{x}_u|\mathcal{C}_k)$. Using the **linearity** of mean statistics and the **Central Limit Theorem** (CLT), $\vec{s}_p$ can be approximated as being generated from a normal distribution as follows:

$$\vec{s}_p \sim \mathcal{N}(\vec{\mu}_p, \mathbf{\Sigma}_p^2) \tag{1}$$

$$\vec{\mu}_p = \sum_{k=1}^{K} \pi_{pk}\vec{\eta}_k \tag{2}$$

$$\mathbf{\Sigma}_p^2 = \sum_{k=1}^{K} \frac{\pi_{pk}(\vec{\eta}\cdot\vec{\eta}^T + \mathbf{T}_k^2) - \vec{\mu}\cdot\vec{\mu}^T}{N_p} \tag{3}$$

$$\vec{\eta}_k = E[\vec{x}_u|\mathcal{C}_k],\ \mathbf{T}_k^2 = Var[\vec{x}_u|\mathcal{C}_k]. \tag{4}$$

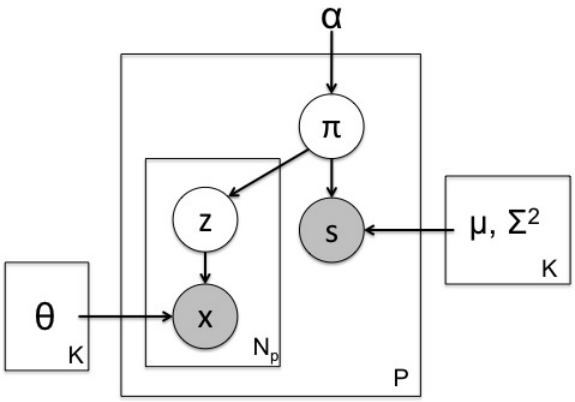

Figure 2: Graphical Model of CUDIA.

Essentially, $\vec{\eta}_k$ and $\mathbf{T}_k^2$ are the sufficient statistics of $\vec{s}_p$'s, since the CLT only requires the mean and variance of the samples. As the actual values of $\vec{x}_u$'s don't contribute to the likelihood of this process, $\vec{x}_u$ can actually be removed, resulting in the efficient **C**lustering **U**sing features with **DI**fferent levels of **A**ggregation (CUDIA) model as shown in Figure 2. The full generative process for CUDIA is as follows:

For $\vec{s}_p$ in $\mathcal{D}_s$,

- Sample $\vec{\pi}_p \sim Dirichlet(\vec{\alpha})$.
- Sample $\vec{s}_p \sim \mathcal{N}(\vec{\mu}_p, \mathbf{\Sigma}_p^2)$,
  where $\vec{\mu}_p = \sum_{k=1}^{K} \pi_{pk}\vec{\eta}_k$
  and $\mathbf{\Sigma}_p^2 = \sum_{k=1}^{K} \frac{\pi_{pk}(\vec{\eta}\cdot\vec{\eta}^T + \mathbf{T}_k^2) - \vec{\mu}\cdot\vec{\mu}^T}{N_p}$.
- For $\vec{x}_{oi}$ in $\mathcal{D}_x^p$,
  Sample $\vec{z}_i \sim Multinomial(\vec{\pi}_p)$.
  Sample $\vec{x}_{oi} \sim \prod_{k=1}^{K} p(\vec{x}_o|\vec{\theta}_k)^{z_{ik}}$.

$\vec{\pi}$ is sampled from a Dirichlet distribution parametrized by $\vec{\alpha}$, and observed sample mean statistics $\vec{s}$ is generated from a Normal distribution parametrized by a mixture of true means $\vec{\eta}$s and a covariance $\mathbf{\Sigma}^2$. $\vec{z}$'s in each partition are sampled from a Multinomial distribution parametrized by $\vec{\pi}$, which is specific to the partition, and corresponding $\vec{x}_o$s are sampled from a distribution $\prod_{k=1}^{K} p(\vec{x}_o|\vec{\theta}_k)^{z_k}$, where the suitable form of $p(\vec{x}_o|\vec{\theta}_k)$ depends on the properties of the variable $\vec{x}_o$'s. For conciseness, the remaining sections of this paper will denote $\vec{x}_o$ as $\vec{x}$.

4. INFERENCE

From the generative process, the likelihood function of the CUDIA model is given by:

$$\begin{aligned} &p(\mathbf{x}, \mathbf{s}|\vec{\eta}, \mathbf{T}^2, \vec{\theta}, \vec{\alpha}) \\ &= \sum_{\mathbf{z}} \int_{\boldsymbol{\pi}} \prod_{p=1}^{P} p(\vec{s}_p|\vec{\pi}_p, \vec{\eta}, \mathbf{T}^2) p(\vec{\pi}_p|\vec{\alpha}) \prod_{i=1}^{N_p} \prod_{k=1}^{K} p(\vec{x}_i|\vec{\theta}_k)^{z_{ik}} p(\vec{z}_i|\vec{\pi}_p) d\boldsymbol{\pi} \\ &= \int_{\boldsymbol{\pi}} \prod_{p=1}^{P} p(\vec{s}_p|\vec{\pi}_p, \vec{\eta}, \mathbf{T}^2) p(\vec{\pi}_p|\vec{\alpha}) \prod_{i=1}^{N_p} \prod_{k=1}^{K} \sum_{\mathbf{z}} p(\vec{x}_i|\vec{\theta}_k)^{z_{ik}} p(\vec{z}_i|\vec{\pi}_p) d\boldsymbol{\pi}. \end{aligned}$$

The posterior distribution of the hidden variables, $\vec{\pi}$'s and $\vec{z}$'s, is as follows:

$$p(\boldsymbol{\pi}, \mathbf{z}|\vec{\eta}, \mathbf{T}^2, \vec{\theta}, \vec{\alpha}, \mathbf{x}, \mathbf{s}) = \frac{p(\mathbf{x}, \mathbf{s}, \boldsymbol{\pi}, \mathbf{z}|\vec{\eta}, \mathbf{T}^2, \vec{\theta}, \vec{\alpha})}{p(\mathbf{x}, \mathbf{s}|\vec{\eta}, \mathbf{T}^2, \vec{\theta}, \vec{\alpha})}. \quad (5)$$

The key inferential problem is how to calculate this posterior distribution. A generic EM algorithm [**?**] cannot be applied, since the normalization constant of its posterior distribution in Equation (5) is intractable. Collapsed Gibbs sampling [**?**] also cannot be applied because $\vec{\pi}$ cannot be integrated out due to non-conjugacy between $\vec{s}$ and $\vec{\pi}$ in $p(\mathbf{x}, \mathbf{s}, \boldsymbol{\pi}, \mathbf{z}|\vec{\eta}, \vec{\theta}, \vec{\alpha})$. In this case, the model can be learned using either variational methods or Gibbs sampling approaches, and this paper follows the latter alternative. Nevertheless, naïve Gibbs sampling approaches are computationally inefficient, thus this paper employs an approximated Gibbs sampling approach, which can be applied when the dimension of $\vec{x}$ is small. The model parameter estimation follows the MCEM algorithm [**?**] using this approximation technique.

4.1 E-step: Gibbs Sampling

In CUDIA, the latent variables are $\vec{\pi}$ and z. So we have:

$$p(\mathbf{x},\mathbf{s}, \boldsymbol{\pi}, \mathbf{z}|\vec{\eta}, \vec{\theta}, \vec{\alpha}) = \prod_{p=1}^{P} p(\vec{s}_p|\vec{\pi}_p, \vec{\eta}) p(\vec{\pi}_p|\vec{\alpha}) \prod_{i=1}^{N_p} \prod_{k=1}^{K} p(\vec{x}_i|\vec{\theta}_k)^{z_{ik}} p(\vec{z}_i|\vec{\pi}_p).$$

For each partition p, the Gibbs sampling is performed as follows:

$$\vec{\pi}_p^{(j+1)} \sim p(\vec{\pi}|\vec{z}_1^{(j)}, \vec{z}_2^{(j)}, ..., \vec{z}_{N_p}^{(j)}, \vec{s}_p, \vec{\eta}, \vec{\alpha}) \quad (6)$$

$$\vec{z}_i^{(j+1)} \sim p(\vec{z}|\vec{\pi}_p^{(j+1)}, \vec{x}_i, \vec{\theta}). \quad (7)$$

However, sampling $\vec{\pi}$ is problematic as Eq. (6) is not a trivial distribution. Instead of sampling directly from Eq. (6), Metropolis-Hastings (MH) algorithm can be used with a proposal density $Dirichlet(\vec{\alpha})$:

$\vec{\pi}_p^{(new)} \sim Dir(\vec{\alpha})$ and $\zeta \sim Uniform(0, 1)$.

$\vec{\pi}_p^{(j+1)} \leftarrow \vec{\pi}_p^{(new)}$ if $\zeta < g(\vec{\pi}_p^{(new)}, \vec{\pi}_p^{(j)}) \prod_k^K (\frac{\pi_{pk}^{(new)}}{\pi_{pk}^{(j)}})^{n(z_{\cdot k}^{(j)})}$,

where $g(\vec{\pi}_p^{(new)}, \vec{\pi}_p^{(j)}) = \frac{p(\vec{s}_p|\vec{\pi}_p^{(new)}, \vec{\eta}) p(\vec{\pi}_p^{(new)}|\vec{\alpha})^2}{p(\vec{s}_p|\vec{\pi}_p^{(j)}, \vec{\eta}) p(\vec{\pi}_p^{(j)}|\vec{\alpha})^2}$ and $n(z_{\cdot k}^{(j)})$ is the count of $z_{\cdot k}^{(j)} = 1$.

Even though this MH algorithm inside the Gibbs sampling becomes inefficient when dealing with large datasets, the sampling step of $\vec{z}$'s can be removed assuming a large enough data size of N_p and a small dimension of $\vec{x}$.

The overall idea of this approximation is as follows: If $\vec{x}$ is generated from an exponential family distribution, $p(z_k|\vec{x}, \pi)$ is continuous with respect to $\vec{x}$, so that $p(\vec{z}|\vec{x}, \vec{\pi}) \approx p(\vec{z}|\vec{x} + d\vec{x}, \vec{\pi})$. Consider a ball of radius $r > 0$ centered at $\vec{x}^c$, $B_r(\vec{x}^c)$, such that $p(\vec{z}|\vec{x}^c, \vec{\pi}) \approx p(\vec{z}|\vec{x}, \vec{\pi})$, where $\vec{x}$ is in the ball. If the number of $\vec{x}$'s that are in the ball is large enough, then $n(z_{\cdot k})$ in the ball can be approximated as $n(z_{\cdot k}) \approx |B_r(\vec{x}^c)| E[z_k|\pi_p, \vec{x}^c] \approx \sum_{\vec{x} \in B_r(\vec{x}^c)} E[z_k|\pi_p, \vec{x}]$. This idea can be effectively applied when N_p is large and the dimension of $\vec{x}$ is small, even better when $\vec{x}$ is a discrete variable. Assuming partitional balls over $\mathcal{D}_x^p$, $n(z_{\cdot k})$ in the partition p can be approximated as $\sum_{i=1}^{N_p} E[z_k|\pi_p, \vec{x}_i]$. Letting the number of Gibbs samples be N_{Gibbs}, the algorithm works as follows:

For $j = 1$ to N_{Gibbs},

- Sample $\pi_p^{(j+1)}$ using MH algorithm, where $n(z_{\cdot k}^{(j)}) \leftarrow \sum_{i=1}^{N_p} E[z_k|\pi_p^{(j)}, \vec{x}_i]$
- Set $E[z_k|\pi_p^{(j+1)}, \vec{x}_i] = \frac{p(\vec{x}_i|\vec{\theta}_k)\pi_{pk}^{(j+1)}}{\sum_{k=1}^K p(\vec{x}_i|\vec{\theta}_k)\pi_{pk}^{(j+1)}}$.

$E[z_k|\vec{x}] \propto \sum_{j=1}^{N_{Gibbs}} E[z_k^{(j)}|\pi_p^{(j)}, \vec{x}]$.

The last line of the algorithm is derived by using the Partition Theorem of conditional expectation [**?**]. As a result, the actual sampling process occurs only in the MH sampling. In this paper, we used a burning period of 10 samples, and $N_{Gibbs} \approx 50$ to 100 [**?**]. Experimental results show that with this small number of samples, the algorithm converges with reasonable speed.

4.2 M-step: Parameter Estimation

Model parameters are $\vec{\alpha}$, $\vec{\theta}$ and $\vec{\eta}$. Maximization on $\vec{\alpha}$ and $\vec{\theta}$ can be easily performed and won't be discussed in this paper. $\vec{\eta}^*$ and $\mathbf{T}^*$ can be obtained by alternating the maximization steps on $\vec{\eta}$ and $\mathbf{T}$ respectively. However, if we assume $\mathbf{T}_k^2 = \delta_k^2 \mathbf{I}$, the maximization step on $\vec{\eta}$ can be simplified. To simplify the notation, the following matrices are defined [**?**] :

$$\mathbf{S}_i = [s_{1i}, s_{2i}, ..., s_{Pi}]^T \quad (8)$$

$$\hat{\mathbf{\Pi}} = [\hat{\vec{\pi}}_1, \hat{\vec{\pi}}_2, ..., \hat{\vec{\pi}}_P]^T, \text{ where } \hat{\vec{\pi}}_p = \frac{\sum_{i=1}^{N_{Gibbs}} \vec{\pi}_p^{(i)}}{N_{Gibbs}} \quad (9)$$

$$\mathbf{W} = diag(N_1, N_2, ..., N_P) \quad (10)$$

$$\mathbf{H} = [\vec{\eta}_1, \vec{\eta}_2, ..., \vec{\eta}_K]^T \quad (11)$$

To help understanding, their expressive forms are given by:

$$\mathbf{S}_i = \begin{pmatrix} s_{1i} \\ s_{2i} \\ ... \\ s_{Pi} \end{pmatrix}, \ \mathbf{H}_{\cdot i} = \begin{pmatrix} \eta_{1i} \\ \eta_{2i} \\ ... \\ \eta_{Ki} \end{pmatrix} \quad (12)$$

$$\hat{\mathbf{\Pi}} = \begin{pmatrix} \hat{\pi}_{11} & \hat{\pi}_{12} & ... & \hat{\pi}_{1K} \\ \hat{\pi}_{21} & \hat{\pi}_{22} & ... & \hat{\pi}_{2K} \\ ... & ... & ... & ... \\ \hat{\pi}_{P1} & \hat{\pi}_{P2} & ... & \hat{\pi}_{PK} \end{pmatrix}. \quad (13)$$

As $\vec{s}$ is normally distributed in CUDIA, the relationship between $\mathbf{S}_i$ and $\mathbf{H}_{\cdot i}$ in the CUDIA model can be described as:

$$\mathbf{S}_i \approx \hat{\mathbf{\Pi}} \cdot \mathbf{H}_{\cdot i} \quad (14)$$

However, each $\vec{s}_p$ has a different variance, thus the solution of 'weighted linear regression' can be applied to get the optimal $\mathbf{H}_{\cdot i}^*$:

$$\mathbf{H}_{\cdot i}^* = (\hat{\mathbf{\Pi}}^T \mathbf{W} \hat{\mathbf{\Pi}})^{-1} \hat{\mathbf{\Pi}}^T \mathbf{W} \mathbf{S}_i. \quad (15)$$

Note that $rank(\hat{\mathbf{\Pi}}^T \mathbf{W} \hat{\mathbf{\Pi}}) = rank(\hat{\mathbf{\Pi}}) = K$ w.p. 1 if $P > K$. However, mean values ($\hat{\mathbf{\Pi}}$) are susceptible to outliers from the Gibbs sampling. To ensure the invertibility, regularization techniques can be incorporated. For example, if a Ridge penalty is used, then $\mathbf{H}$ becomes:

$$\mathbf{H}_{\cdot i}^* = (\hat{\mathbf{\Pi}}^T \mathbf{W} \hat{\mathbf{\Pi}} + \lambda_M \mathbf{I})^{-1} \hat{\mathbf{\Pi}}^T \mathbf{W} \mathbf{S}_i. \quad (16)$$

Furthermore, the regularizer term, λ_M, can be utilized when $P < K$, which makes CUDIA under-determined. But we leave this to the future work.

5. DETERMINISTIC HARD CLUSTERING

The CUDIA model provides an intuitive deterministic hard clustering algorithm. From the log-likelihood of CUDIA, the objective function becomes:

$$\min_{\mathbf{z},\vec{\mu},\vec{\eta}} \sum_p \{\sum_{k,n_p} z_{n_pk} \parallel \vec{x}_{n_p} - \vec{\mu}_k \parallel^2\} + \beta \parallel \vec{s}_p - \sum_k \frac{\sum_{n_p} z_{n_pk}}{N_p} \vec{\eta}_k \parallel^2 \tag{17}$$

$$= \min_{\mathbf{z},\vec{\mu},\vec{\eta}} \sum_{p,k,n_p} z_{n_pk} \parallel \vec{x}_{n_p} - \vec{\mu}_k \parallel^2 + \frac{\beta}{KN_p} \parallel \vec{s}_p - \sum_k \hat{\pi}_{pk} \vec{\eta}_k \parallel^2 \tag{18}$$

where $\hat{\pi}_{pk} = \frac{\sum_{n_p} z_{n_pk}}{N_p}$ and β is a parameter that determines weights to mean statistics. Local minima of this objective function can be found by alternating minimization steps between $\mathbf{z}$ and $(\vec{\mu}, \vec{\eta})$:

- **Assignment Step**

  $z_{n_pk^*} \leftarrow 1$,
  if $k^* = \arg\min_k \parallel \vec{x}_{n_p} - \vec{\mu}_k \parallel^2 - 2(\vec{s}_p - \mathbf{H}^T \hat{\vec{\pi}}_p)^T \vec{\eta}_k (\frac{\beta}{KN_p})$

  $z_{n_pk^*} \leftarrow 0$, otherwise.

- **Update Step**

  $\vec{\mu}_k \leftarrow \sum_n (z_{nk}\vec{x}_n)/N_k$, $\vec{\pi}_p \leftarrow \sum_{n_p} \vec{z}_{n_p}/N_p$

  $\mathbf{H}_{\cdot i} \leftarrow (\hat{\mathbf{\Pi}}^T \mathbf{W} \hat{\mathbf{\Pi}} + \lambda_M \mathbf{I})^{-1} \hat{\mathbf{\Pi}}^T \mathbf{W} \mathbf{S}_i$

One iteration of this algorithm costs $\Theta(KN)$. For a fixed number of iterations I, the overall complexity is therefore $\Theta(KNI)$, which is linear in all relevant factors. The complexity of this algorithm is the same as *k-means* promising its scalability to massive datasets. Moreover, this algorithm can be used as an initialization step for the probabilistic algorithm, which in turn will reduce the total running time.

The squared loss function in the deterministic algorithm is appropriate for an additive Gaussian model. Our approach can however be generalized to any exponential family distribution (of which the Gaussian is a specific example) by exploiting the bijection between this family and the family of loss functions represented by Bregman divergences [**?**]. Given two vectors $\vec{x}$ and $\vec{\mu}$, the Bregman divergence is defined as:

$$d_\phi(\vec{x}, \vec{\mu}) = \phi(\vec{x}) - \phi(\vec{\mu}) - \langle \vec{x} - \vec{\mu}, \nabla\phi(\vec{\mu}) \rangle \tag{19}$$

where $\phi(\cdot)$ is a differentiable convex function and $\nabla\phi(\vec{\mu})$ represents the gradient vector of ϕ evaluated at $\vec{\mu}$. Although the Bregman divergence possesses many other interesting properties, this paper focuses on its bijective relationship to the Exponential family distribution.

This bijective relation can be exploited when clustering data points that cannot be appropriately modeled using the Gaussian distribution, as in the Bregman Hard/Soft Clustering algorithms [**?**]. Table 1 shows the relationship between Bregman divergences and their corresponding Exponential family distributions. Using this bijection, the deterministic algorithm of CUDIA can be extended as follows:

Table 1: Bregman divergence and Exponential family.

Distribution	$\phi(\vec{\mu})$	$d_\phi(\vec{x}, \vec{\mu})$
1-D Gaussian	$\frac{1}{2\sigma^2}\mu^2$	$\frac{1}{2\sigma^2}(x-\mu)^2$
1-D Exponential	$\mu \log \mu - \mu$	$x \log(\frac{x}{\mu}) - (x - \mu)$
d-D Gaussian	$\frac{1}{2\sigma^2} \parallel \vec{\mu} \parallel^2$	$\frac{1}{2\sigma^2} \parallel \vec{x} - \vec{\mu} \parallel^2$
d-D Multinomial	$\sum_{j=1}^d \mu_j \log \frac{\mu_j}{M}$	$\sum_{j=1}^d x_j \log \frac{x_j}{\mu_j}$

- **Assignment Step**

  $z_{n_pk^*} \leftarrow 1$,
  if $k^* = \arg\min_k d_\phi(\vec{x}_{n_p}, \vec{\mu}_k) - 2(\vec{s}_p - \mathbf{H}^T \hat{\vec{\pi}}_p)^T \vec{\eta}_k (\frac{\beta}{KN_p})$

  $z_{n_pk^*} \leftarrow 0$, otherwise.

where ϕ can be chosen based on the distribution of $\vec{x}$ and the update step remains the same.

This extended algorithm captures various distributions while maintaining the original complexity. Furthermore, the linkage between the Bregman divergence and the Exponential family distributions enables probabilistic interpretations on the resultant clustering assignments as in the Bregman Soft Clustering algorithm. Perhaps the most useful case is when the vectors represent probability distributions, in which case the KL-divergence (another special case of Bregman divergences), is the appropriate loss function to use.

6. IMPUTATION

After all the parameters of the CUDIA model are learned, the model allows us to impute the unobserved features $\vec{x}_u$'s at the individual level. Given the observed features and learned parameters, the imputation is as follows:

$$p(\vec{x}_u|\vec{x}_o, \vec{\pi}_p) = \sum_k p(\vec{x}_u, z_k|\vec{x}_o, \vec{\pi}_p) \tag{20}$$

$$= \sum_k \frac{p(\vec{x}_u, z_k, \vec{x}_o, \vec{\pi}_p)}{p(\vec{x}_o)} \tag{21}$$

$$= \sum_k \frac{p(\vec{x}_u|z_k, \vec{x}_o, \vec{\pi}_p) p(z_k|\vec{x}_o, \vec{\pi}_p) p(\vec{x}_o, \vec{\pi}_p)}{p(\vec{x}_o, \vec{\pi}_p)} \tag{22}$$

$$= \sum_k p(\vec{x}_u|z_k) p(z_k|\vec{x}_o, \vec{\pi}_p). \tag{23}$$

The exact imputation formula depends on the pdf of the unobserved features ($p(\vec{x}_u|z_k)$). For example, if $\vec{x}_u$ is generated from a Gaussian distribution with mean $\vec{\eta}_k$, the imputation formula obtained is:

$$\hat{\vec{x}}_u \leftarrow \sum_{k=1}^K \vec{\eta}_k E[z_k|\vec{x}_o, \vec{\pi}_p] \tag{24}$$

where the covariance of $\vec{x}_u$ is assumed to be $\delta^2\mathbf{I}$. This imputation method also can be applied to the deterministic algorithm. The bijective relationship between Bregman divergence and Exponential family yields a soft cluster assignment as follows:

$$E[z_k|\vec{x}_o, \vec{\pi}_p] \propto \frac{exp(-d_\phi(\vec{x}_o, \vec{\mu}_k))}{\sum_l exp(-d_\phi(\vec{x}_o, \vec{\mu}_l))} \pi_{pk}. \tag{25}$$

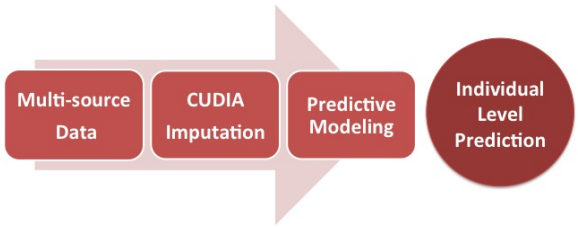

Figure 3: CUDIA utilizes the aggregated information, and generates the imputed individual value for the aggregated features. The resultant information can be used in various predictive modeling settings.

Thus, the deterministic algorithm provides not only the cluster centroids/assignments, but also the basic imputation framework on the unobserved features, which in turn can be used for preliminary tests for the model's applicability.

7. EXPERIMENTAL RESULTS

In this section, we provide two kinds of experimental results. (a) First, imputation quality of the CUDIA model is assessed using Old Faithful data and a simulated mixture of Gaussians data. (b) Then, its applicability to predictive modeling[3] is discussed using data from the Dartmouth Health Atlas, CDC and the Census Bureau. Depending on the nature of the predictor and the data source, averaged values are provided at hospital, county, HRR/HSA or state levels. Thus "individual" will refer to either a single hospital or a single county as these are at the finest granularity level in the corresponding studies. The CUDIA model is used to impute the aggregated features at the individual-level, and its results are compared to predictive modeling using only higher level data. The workflow of the CUDIA framework is described in Figure 3.

7.1 Imputation Qualities

Before going through the regression analyses using the CUDIA model, we present simple experiments to indicate the usefulness of the imputation based on CUDIA. The Old Faithful dataset consists of two features, 'duration' and 'interval'. To fit our purpose, the dataset is partitioned into 7 groups, and the average values of the 'interval' for each partition are used as s_p (thus, 'duration': x_o, 'interval': x_u). The CUDIA model is used to impute $\hat{x}_u$ at individual level in this setting. Figure 4 shows the results with different λ's, where K is set as 2 in the CUDIA model. As the value of λ_M (Ridge penalty) increases, the imputed features tend to be near to zero. The exact determination of λ_M and K might be difficult, as there is no universal criterion for the optimal K and λ_M. In our framework, λ_M and K will be determined through cross-validation in predictive modeling tasks. The baseline imputation model uses the averaged 'interval' for each partition as its individual feature. Figure 5 shows the mean squared errors (MSE) for both models, one using the baseline and the other using the CUDIA model. The imputation based on the CUDIA model clearly exhibits reduced MSE.

Another example is provided using a simulated mixture of Gaussians data. 960 data points are generated from a mixture of four two-dimensional Gaussian distributions, then they are partitioned into 8 partitions. The first column entries of the data are preserved, while the other column information is aggregated. Figure 6 shows both the result from the CUDIA imputation and the baseline imputation. Apparently, the model captures the underlying distribution, even though some information is only given in an aggregated format.

[3]Targets are chosen arbitrarily to illustrate the applicability of the CUDIA framework.

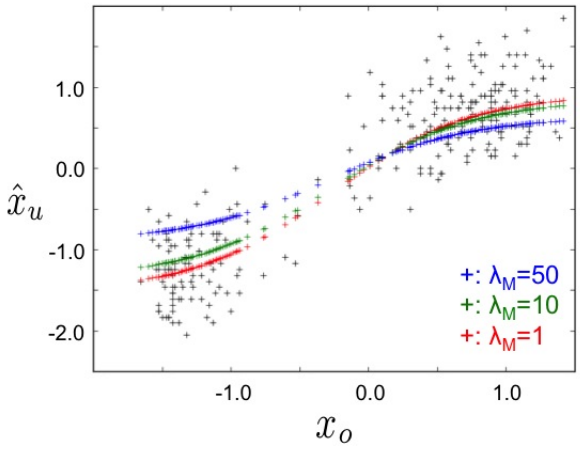

Figure 4: Old Faithful dataset. The original dataset is plotted using black '+'. The imputed features with various λ_M's are shown in different colors. All features are standardized centered at zero.

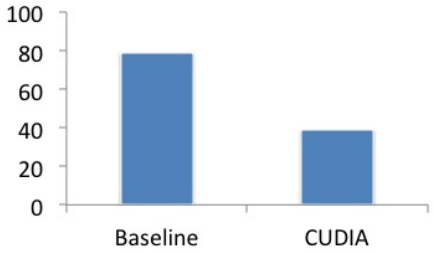

Figure 5: Imputation Accuracy on Old Faithful dataset. Vertical axis indicates MSE with respect to the original feature values.

7.2 Dartmouth Health Atlas: Case 1

The Dartmouth Health Atlas dataset [**?**] is composed of several tables with different levels of aggregation. For example, the number of beds in a hospital can be accessed at the hospital-level, whereas the medical/surgical discharge rates can only be obtained at State/HRR/HSA levels. Although the number of beds is strongly related with the medical/surgical discharge rates [**?**], this information cannot be directly used as they are aggregated at a different level. Table 2 describes the subset of the Dartmouth data used in this experiment. Only data from the 5 most populous states (CA, FL, IL, NY, TX) was used so as to have a higher value of number of hospitals per state. For this subset, the "complete data" would have consisted of five variables at the hospital-level, of which two are actually available only at the

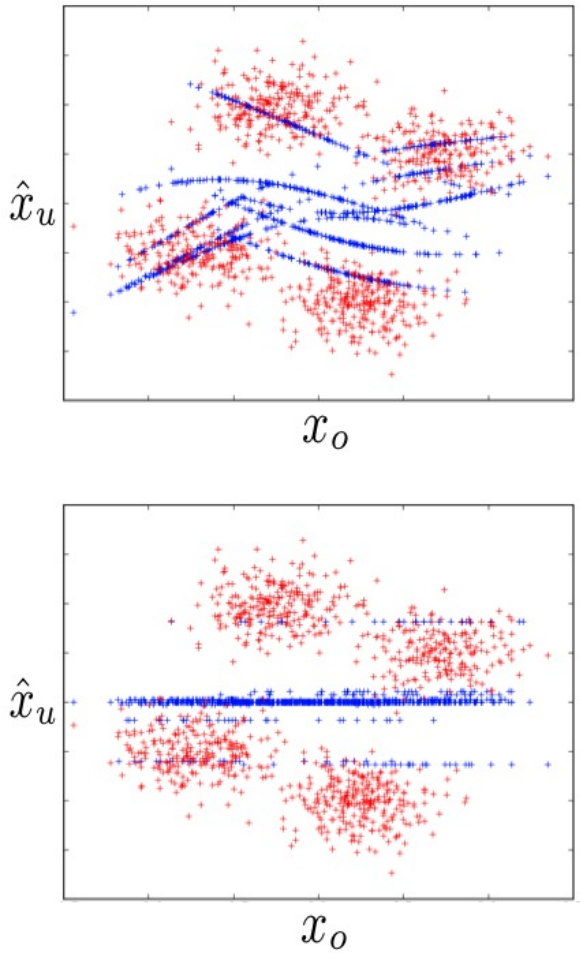

Figure 6: Mixture of Gaussians. Red '+' denotes the complete data, and blue '+' indicates the imputed data: (a) CUDIA imputation (top), (b) Baseline imputation (bottom). CUDIA captures the underlying distribution.

state level. The CUDIA model can be used to impute the unobserved features ($\vec{x}_u$).

Ridge regression is used to perform predictive modeling on three kinds of datasets: 1) 'hospital-level' dataset alone, 2) imputed complete dataset using 'state-level' summaries and 3) imputed dataset using the CUDIA model, then 5-fold CV is performed. Note that LASSO or other regression techniques can be used in this framework, and Ridge regression is just a choice in this paper. Furthermore, if a predictive modeling task is a classification task, then other classification algorithms such as Decision Trees, SVM and Logistic regression can be used depending on their performances. One fold is used for testing, and the remaining folds are again divided into training and validation sets (80/20). 'λ' (in Ridge regression) is learned for each run. Each run has different λ values. Table 3 shows the results. $K = 5$ gives the best R^2 value among all the alternatives. Table 4 shows the coefficients of Linear regression when $K = 5$. The

Table 2: Dataset Description. Target is not included when performing the imputation. The top 5 biggest population states are selected to maintain large enough N_p. (Case 1)

	Hospital-level
1	Hospital beds(**Target**)
2	Home health agency visits per decedent
3	Percent of deaths occurring in hospital

	State-level
1	Medical discharge rate
2	Surgical discharge rate

Table 3: Regression Results on Dartmouth datasets. R^2s over 5-fold cv are listed. As $K < P$, $K > 5$ is not an option.

Dataset	Case 1	Case 2	Case 3
No Imputation	0.549 (±0.023)	0.551 (±0.029)	0.654 (±0.053)
State-level Imputation	0.558 (±0.020)	0.585 (±0.045)	0.662 (±0.043)
CUDIA Imp. ($K = 2$)	0.551 (±0.030)	0.524 (±0.032)	0.682 (±0.056)
CUDIA Imp. ($K = 3$)	0.553 (±0.022)	0.547 (±0.029)	0.685 (±0.057)
CUDIA Imp. ($K = 4$)	0.578 (±0.028)	0.598 (±0.029)	0.688 (±0.057)
CUDIA Imp. ($K = 5$)	**0.597** (±0.027)	**0.601** (±0.030)	**0.689** (±0.057)

imputed medical discharge rate is positively correlated with the number of beds in a hospital.

Figure 7 shows the imputed 'Medical discharge rate' with respect to 'Percent of deaths occurring in hospital' when $K = 5$. The imputed features shows the highly non-linear relationship to the fully observed features, as the imputation is based on distinct cluster centroids. This non-linear imputation captures non-linear relationships among features in real-life complex datasets.

7.3 Dartmouth + External Source: Case 2

State-level summaries of health-related indicators can be obtained from various external sources. For example, the Center for Disease Control and Prevention (CDC) publishes annual state-level health statistics, that covers aging, cancer, diabetes, etc. In this experiment, the Dartmouth dataset is used with an external dataset from StateMaster.com, which

Table 4: Coefficients of Linear regression when $K = 5$. All features are standardized. (Case 1)

Independent Variable	Coefficient
Home health agency visits per decedent	0.276
Percent of deaths occurring in hospital	0.646
Medical discharge rate	**0.618**
Surgical discharge rate	**-0.057**

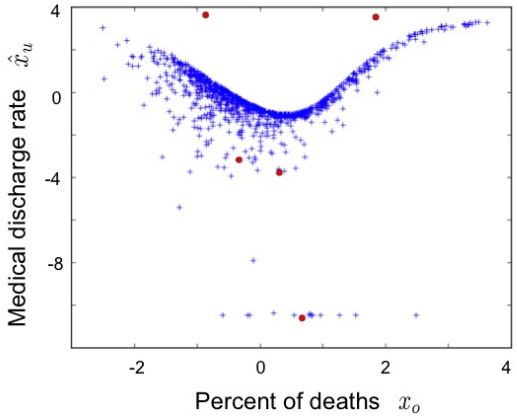

Figure 7: Imputed Dartmouth Dataset at Hospital-level ($K = 5$). Cluster centroids are plotted using red dots. Horizontal axis shows 'Percent of death occurring in hospital' and vertical axis indicates the imputed 'Medical discharge rate' at hospital-level. All features are standardized. (Case 1)

Table 5: Coefficients of Linear regression using the external source when $K = 5$. All features are standardized. (Case 2)

Independent Variable	Coefficient
Home health agency visits per decedent	0.249
Percent of death occurring in hospital	0.359
Healthcare spending	**0.280**
Admissions	**0.177**
Adult physical disabilities	**-0.051**

provides multiple state-level statistics for free. The hospital-level Dartmouth dataset from the previous experiment is used as is. The state-level dataset is replaced with the external dataset, which has state-level 1) healthcare spending, 2) hospital admissions and 3) adult physical disabilities information. All these features are not available in the Dartmouth data. As in the previous experiment, three datasets are formed. Table 3 shows the R^2 results using 5-fold CV. Imputation using the CUDIA model leads to a 9% increase in R^2 value compared to the base model without imputation. Table 5 shows the coefficients of Linear regression when $K = 5$. The imputed healthcare spending exhibits the strongest correlation with hospital spending, as one may expect.

7.4 Dartmouth + External Source: Case 3

In this experiment, Medicare part-A reimbursement at HSA-level is predicted based on Medicare part-B reimbursement and an additional external information. In the Dartmouth dataset, 'Selected Medicare Reimbursement' table contains the columns of Medicare reimbursement part-A and part-B at HSA-level. Although Medicare part-A is closely related to part-B, additional features, such as income or education levels, can be incorporated not only improving the performance of the regression but also providing richer

Table 6: Coefficients of Linear regression using the external source when $K = 4$. All features are standardized. (Case 3)

Independent Variable	Coefficient
Medicare Part-B	0.611
Income per capita	**-0.189**
Healthcare spending	**-0.080**
Education level (Bachelor or higher)	**-0.209**

Table 7: Coefficients of Linear regression on CDC diabetes dataset when $K = 5$. All features are standardized. 'Obesity rate' is set as a target.

Independent Variable	Coefficient
Diabetes	0.027
Physical inactivity	1.00
Income per capita	**-0.234**
Healthcare spending	**0.352**
Education level (Bachelor or higher)	**0.139**

interpretations. The external state-level features used in this experiment are 1) income per capita, 2) total healthcare spending and 3) education level (ratio of bachelors or higher). The experiment is performed using three datasets, which are prepared as in the previous experiments. Table 3 shows the results. Table 6 exhibits the coefficients of Linear regression when $K = 4$. The imputed 'education level' and Medicare Part-A are negatively correlated.

7.5 CDC Diabetes Dataset

The Center for Disease Control and Prevention (CDC) [**?**] provides county-level estimates of 1) obesity, 2) diabetes and 3) physical inactivity. In this experiment, we predict the county-level obesity rate using the other features in the CDC dataset and additional state-level features. The state-level features used in this experiment are the same as in the previous experiment (Dartmouth Case 3). The top 5 biggest states are used, as some smaller states have very few counties. Table 8 shows the R^2 results. The imputed dataset using the CUDIA model gives the best result. The state-level imputed dataset yields a poorer result than the dataset with no imputation. This indicates that the uncertainty in the state-level imputation of the added variables over-rode any extra benefits that these variables could have provided. Table 7 depicts the coefficients when $K = 5$. While the imputed 'income per capita' at county-level shows a negative correlation, both imputed 'healthcare spending' and 'education level' are positively correlated with the target (obesity rate at county-level).

7.6 Census Bureau Health Insurance Dataset

The U.S. Census Bureau [**?**] provides county-level estimates of insured population ratio by income levels. Income levels are divided into three overlapping groups: 1) all income levels, 2) at or below 200% of poverty threshold and 3) at or below 250% of poverty threshold. Suppose we want to see which other factors affect propensity of poor people to buy healthcare insurance at the county level. The state-level

Table 8: Regression Results on CDC Diabetes and Census Bureau Dataset. R^2s over 5-fold cv are listed.

Dataset	CDC Diabetes	Census Bureau
No Imputation	0.408	0.512
	(±0.050)	(±0.048)
State-level Imputation	0.398	0.504
	(±0.051)	(±0.049)
CUDIA Imp. ($K = 2$)	0.408	0.513
	(±0.047)	(±0.048)
CUDIA Imp. ($K = 3$)	0.405	0.513
	(±0.052)	(±0.049)
CUDIA Imp. ($K = 4$)	0.422	0.510
	(±0.047)	(±0.048)
CUDIA Imp. ($K = 5$)	**0.426**	**0.520**
	(±0.043)	(±0.043)

Table 9: Coefficients of Linear regression on Census Bureau dataset when $K = 5$. All features are standardized. 'Percent insured for the below 200% of poverty' is set as a target.

Independent Variable	Coefficient
Percent insured for all income levels	2.10
Income per capita	**-0.436**
Healthcare spending	**1.227**
Education level (Bachelor or higher)	**-2.524**

dataset in the previous experiment is used to determine if other factors play a role. Table 8 shows the regression results using the CUDIA model and the coefficients when $K = 5$ are described in Table 9. 'Income per capita' and 'education level' are negatively correlated with the target (percent insured for the below 200% of poverty group). This result indicates that the imputed county-level summaries for both income per capita and education level implicitly inform us of the sizes of poverty group at county-level. Moreover, the imputed healthcare spending at county-level exhibits a positive relationship. Thus these imputed features provide a richer interpretation of the predictive model while simultaneously improving the prediction accuracy.

8. CONCLUDING REMARKS

In this paper, aggregated statistics over certain partitions are utilized to identify clusters and impute features that are observed only as more aggregated values. The imputed features are further used in predictive modeling (Ridge regression in this paper), leading to improved R^2 values. The experiments provided in this paper are illustrative of the generality of the propsed framework and its applicability to several healthcare related datasets in which individual records are often not available, and different information sources reflect different types and levels of aggregation. Empirical studies on larger and richer datasets are forthcoming.

CUDIA is quite scalable, and in particular, the deterministic hard clustering version of the CUDIA model can be readily applied to massive datasets. Furthermore, the square loss function on $\vec{x}_o$ can be generalized to Bregman divergence, or equivalently, one can cater to any noise function from the exponential family of probability distributions [**?**]. One restriction of the current model is that the number of clusters (K) cannot be more than the number of partitions specified by the data provider (P). This is why we had to stop at K=5 for several of the results even though the R^2 values were improving with with increasing K. Adding more partitions, e.g., incorporating data from more than 5 states, should reflect in further improvements in the results.

$$\vec{\pi}_p = (p(\mathcal{C}_1|\text{partition p}),\ ...,\ p(\mathcal{C}_K|\text{partition p}))^T$$
$$= (\pi_{p1},\ ...,\ \pi_{pK})^T$$

Acknowledgments

This research was supported by NSF IIS-1016614 and by TATP grant 01829.