CUDIA: Probabilistic Cross-level Imputation using Individual Auxiliary Information

Yubin Park, The University of Texas at Austin
Joydeep Ghosh, The University of Texas at Austin

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 such as Hospital Referral Regions (HRR) or Hospital Service Areas (HSA). Such levels constitute partitions over the underlying individual level data, which may not match the groupings that would have been obtained if one clustered the data based on individual-level attributes. Moreover, treating 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? In this paper, we seek a better utilization of variably aggregated datasets, which are possibly assembled from different sources. We propose a novel "cross-level" imputation technique that models the generative process of such datasets using a Bayesian directed graphical model. The imputation is based on the underlying data distribution and is shown to be unbiased. This imputation can be further utilized in a subsequent predictive modeling, yielding improved accuracies. The experimental results using a simulated dataset and the Behavioral Risk Factor Surveillance System (BRFSS) dataset are provided to illustrate the generality and capabilities of the proposed framework.

Categories and Subject Descriptors: G.3 [**Probability and Statistics**]: Probabilistic algorithms

General Terms: Algorithms

Additional Key Words and Phrases: Clustering, Privacy Preserving Data Mining, BRFSS

ACM Reference Format:
Park, Y. and Ghosh, J., 2012. CUDIA: Probabilistic Cross-level Imputation using Individual Auxiliary Information. ACM Trans. Intell. Syst. Technol. 9, 4, Article 39 (August 2012), 23 pages.
DOI = 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

1. INTRODUCTION

In healthcare-related studies, individual patient or hospital data may contain private or confidential information such as medical conditions. Public disclosure of raw data potentially introduces ethical or legal issues. Thus, in many cases, disclosure of sensitive raw data requires legally authorized protocols such as Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule [HIPAA Compliance Assistance 2003], which includes several de-identification techniques [Emam and Fineberg 2009], or data providers' consent to share. Adherence to these regulations can consume considerable time and effort, and may result in only a limited number of attributes or features being provided to researchers. On the other hand, such measures may be provided publicly at higher or more aggregated levels, such as state-level, county-

This work is supported by NSF IIS-1016614 and by TATP grant 01829. Author's addresses: Y. Park and J. Ghosh, Electrical and Computer Engineering Department, The University of Texas at Austin.

DOI 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

level summaries or averages over health zones such as Hospital Referral Regions (HRR) or Hospital Service Areas (HSA) (for example see http://www.data.gov/health or http://www.cdc.gov/datastatistics/). Averaged statistics over such levels protect privacy by blending individual information in group summaries, thereby making individual data subjects invisible or unidentifiable over sensitive data. This kind of aggregate information may provide richer feature spaces compared to publicly available individual-level data without infringing privacy and legal issues.

Aggregate information has been frequently used across various domains such as political science, ecological and healthcare-related studies due to its relatively easier access. A common practice to deal with such information is to regard group summaries as representatives for the individuals in the same group. Individual-level feature interactions can be inferred by applying several statistical tools, known as "cross-level inferences" [Achen and Shively 1995], [King 1997], in which partitions based on such levels consist of homogeneous characteristic individuals (the constancy assumption). However, this assumption is valid usually under restrictive conditions such as very small sized partitions. Even worse, the constancy assumption is difficult to verify in real applications. In fact, many partitions derived from such levels are composed of a heterogeneous population, which causes the ecological fallacy. Thus, the applicability and the effectiveness of cross-level inference are still controversial [Freedman 1999].

Aggregate variables in healthcare data are not the only problem that hinders individual-level statistical analyses or data mining research. Many healthcare datasets have limited scope of variables or features as the survey has a pre-defined purpose. In such cases, designing additional sets of surveys might be inappropriate due to cost or temporal dynamics. Although combining multiple datasets from different sources can be an alternative solution, their aggregation levels generally differ in their geographical regions or administrative units. For example, 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. To the best of our knowledge, systematic integration of multiple data sources with different aggregation levels has not been studied thoroughly, and we firstly propose a theoretical framework on combining and utilizing such datasets. We expect that a proper utilization of such data might (i) help to reduce any extra cost or (ii) extend the scope of current healthcare research.

In this paper, we seek a better utilization of such aggregated information for augmenting individual-level data. Suppose two datasets from possibly multiple sources are available for research where their aggregation levels are also different. Figure 1 shows an example of data presented at different levels of aggregation, state-level and county-level. We refer to the dataset with a finer granularity as the individual-level dataset, and the other dataset as the aggregate-level dataset. Assuming that the dataset of interest is generated by a mixture model that represents underlying heterogeneous groups, we introduce a novel generative process that 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 more effectively. These "cross-level" imputed features better reflect the underlying distribution of the data, thus a subsequent predictive model using such extended information shows improved performance. Many datasets in the healthcare domain are divided into multiple tables containing different levels of aggregation (sometimes obtained from different sources), and the suggested methodology in this paper can be useful in increasing the utility in such scenarios.

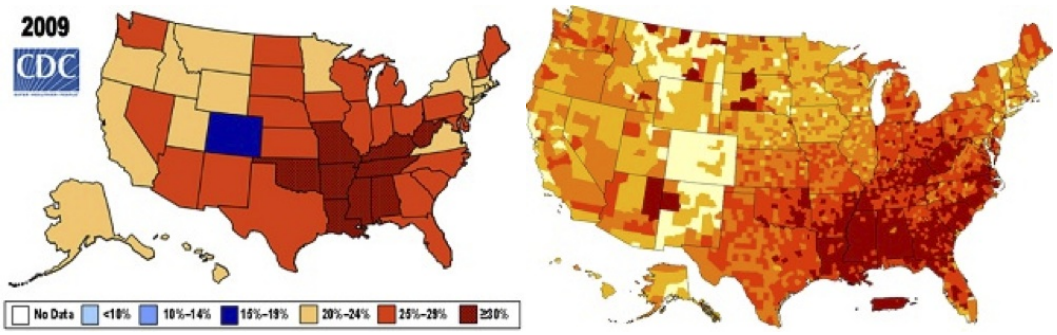

Fig. 1. (a) State-level obesity estimates (left) and (b) county-level diabetes estimates (right). Treating state-level summaries as representative county-level estimates might lead to ecological fallacy. Both figures are adopted from Centers for Disease Control and Prevention (CDC, http://www.cdc.gov/).

The rest of the paper is organized as follows: We begin by reviewing traditional statistical cross-level imputation techniques and then outline various inference mechanisms that will be used extensively in our approach. In Section 3, we approach the problem by modeling the data generation process. We start from a generic Bayesian clustering model, then step-by-step, we impose additional constraints and transform the simple model to suit the problem setting. After presenting the final model, its model parameter estimation technique is explained in Section 4. Due to the complexity of the model, a new approximate Monte Carlo Expectation Maximization (MCEM) algorithm is developed, which is more computationally efficient than a generic MCEM technique. Moreover, a deterministic algorithm that can be used as a parameter initialization method is derived as a valuable artifact of our probabilistic approach. Using the learned model parameters, in Section 4, we propose a "cross-level imputation" formula, which basically enables us to estimate the masked individual values for the aggregate features. The imputation procedure is shown to be a "unbiased" estimator, and it statistical properties are analyzed in detail. To highlight the effectiveness of our approach, we first examine the imputation quality using a simulated dataset in Section 6.1. Then we demonstrate various possible settings in real applications using the Behavioral Risk Factor Surveillance System (BRFSS) dataset in Section 6.2. Finally, we discuss the possible constraints of our framework and future work in Section 7.

2. RELATED WORK

In this section, we summarize 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.

Imputation techniques in statistics. In statistics, there is a vast literature on imputation techniques that are mainly used to substitute missing values in data [Rubin 2004]. 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 developed to supersede this original method. As a special case, when geographical information is missing in the data, geo-imputation techniques are widely used, where the imputation is taken from approximate locations derived from associated data [Henry and Boscoe 2008]. Regression estimation [Tabachnick and Fidel 2001] is another widely used imputation technique in statistics. In regression estimation, the variable with missing data is treated as the dependent variable, while the

other variables are treated as the independent variables. A normal regression is performed based on this setting, then the missing values are replaced by the regression results. Regression estimation assumes enough number of complete individual samples, which is not the case in our setting. If missing values are rather sparse, a Bootstrap technique can be used to improve a subsequent predictive modeling performance [Brownstone and Valletta 2001]. However, these traditional techniques are based on individual-level data, and some of them have limited applicability.

Ecological studies. 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. Most of the ecological analyses are based on ecological regression, which uses the Goodman's "constancy assumption" [Goodman 1953], [Goodman 1959], [King 1997]. The constancy assumption states that behavior within an ecological group does not depend on the other specific characteristics of the group i.e. a group consists of homogenous individuals. 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 [Robinson 1950], which is also known as the "ecological fallacy". In many cases, the constancy assumption may not hold because regional and contextual effects on ecological groups cannot be overlooked, and one ecological group is rarely homogeneous in its behavior.

Ecological regression analysis based on the constancy assumption is vulnerable to "confounding factors" and "aggregation bias". Traditionally, the aggregation bias has been tackled in two ways: (i) by assuming a quadratic model rather than a linear model, or (ii) by calculating interval estimates for unobserved individual features rather than point estimates. In the first method, a quadratic model is obtained by relaxing the constancy assumption [Achen and Shively 1995]. In this framework, an individual in a specific ecological group is no longer independent of the group, and this relationship is specified by a linear model, resulting in a quadratic model at aggregation level. However, the added assumption is not verifiable in most of the cases, and the interpretation of the results becomes harder. In the second method, unobserved individual features are bounded to satisfy aggregated information constraints. This technique is also known as the "method of bounds" [Duncan and Davis 1953]. But the bounds are too broad to be informative in practice, and are typically only used as a sanity check tool.

Despite their theoretical instabilities, ecological analyses continue to be used due to relatively easier access to the aggregate data [Freedman 1999]. Fortuitously, in recent years, it has been reported that auxiliary individual-level information can help to reduce the ecological fallacy [Wakefield and Salway 2001]. In the Hierarchical Related Regression (HRR) framework, auxiliary individual-level information represents a small fraction of the individual samples that constitute the aggregate information [Jackson et al. 2008], [Jackson et al. 2009]. This setting is useful when acquisition costs of getting individual data are expensive, so detailed information is only obtained from a small portion of the entire population. The HRR model relates the regression coefficients from both aggregate and individual data. This analysis has been shown to reduce the ecological bias, but the type of the auxiliary information used in HRR is different from the setting in this paper. In our setting, auxiliary individual-level information has no overlapping feature or column with available aggregate-level data. This happens because aggregated features are privacy sensitive and hence cannot be revealed at individual level for even a small subset of the population. We instead focus on a generative process of such data, and 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.

Inference algorithms in Bayesian Graphical Models. In Bayesian graphical models such as the model presented in this paper, inference can often be challenging. The Expectation Maximization (EM) algorithm is a popular approach when latent variables are present in the models. However, many sophisticated models such as Latent Dirichlet Allocation (LDA) [Blei et al. 2003] have intractable posterior distributions for the latent variables. To approximate the posterior distributions, other techniques such as variational EM algorithm, Gibbs sampling and collapsed Gibbs sampling have been proposed. Although their computational complexities and assumptions are slightly different, their performances are often comparable [Asuncion et al. 2009]. In this paper, we demonstrate an approximated Gibbs sampling approach, which is specialized for our setting. Then we propose a related deterministic algorithm that is not only much faster but also scalable to massive datasets.

3. CLUSTERING MODEL

We denote the set of attributes or features that are available at the individual level by $\vec{x}_o$, where "individual" refers to entities at the finest resolution available. 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 known for each individual. 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 distinct 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[1] to cover such situations, and in this paper we consider a common partitioning to keep the notation and exposition simple.

Suppose we want to find K clusters denoted by $\{\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_K\}$ in the complete data. Note that the clusters are based on the full features $(\vec{x}_o, \vec{x}_u)$, while the partitions typically reflect geographical or administrative units, so that the partitions don't match with the intrinsic clusters. To cater to the unobserved data, an assumption of conditional independence is made: $p(\vec{x}_o, \vec{x}_u|\mathcal{C}_k) = p(\vec{x}_o|\mathcal{C}_k)p(\vec{x}_u|\mathcal{C}_k)$ for any $\mathcal{C}_k$. 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. 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, an 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 $\vec{\alpha}$. As $\vec{x}_u$ and $\vec{x}_o$ are independent given $\mathcal{C}_k$, they can be separated using different nodes. Figure 1 (b) shows a modified clustering model that accommodates the aggregated nature of the unobserved variables. 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**

[1]Section 5.1 outlines how this extension can be achieved after we have introduced the single aggregation level case.

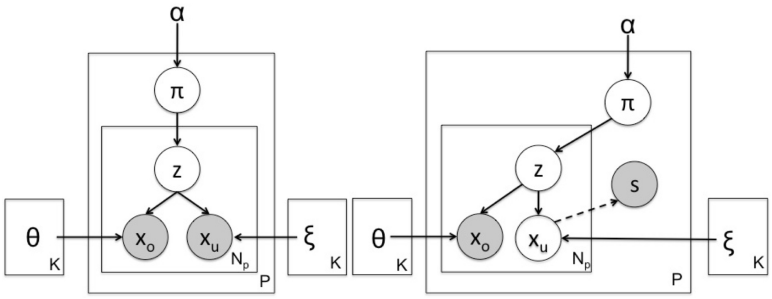

Fig. 2. (a) Clustering models when complete data is available (left) and (b) when only aggregates $\vec{s}$ are observed instead of $\vec{x}_u$ (right).

of mean statistics and the **Central Limit Theorem** (CLT), $\vec{s}_p$ can be approximated as being generated from a normal distribution as follows:

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

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

where $\vec{\eta}_k = E[\vec{x}_u|\mathcal{C}_k],\ \mathbf{T}_k^2 = Var[\vec{x}_u|\mathcal{C}_k]$. 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 $\mathbf{\Sigma}_p^2 = \sum_{k=1}^{K} \frac{\pi_{pk}(\vec{\eta}_k \cdot \vec{\eta}_k^T + \mathbf{T}_k^2) - \vec{\mu}_p \cdot \vec{\mu}_p^T}{N_p}$.
  For $\vec{x}_o$ in $\mathcal{D}_x^p$,
    Sample $\vec{z} \sim Multinomial(\vec{\pi}_p)$.
    Sample $\vec{x}_o \sim \prod_{k=1}^{K} p(\vec{x}_o|\vec{\theta}_k)^{z_k}$.

$\vec{\pi}$ 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:

$$p(\mathbf{x}, \mathbf{s}|\vec{\eta}, \vec{\theta}, \vec{\alpha}) = \sum_{\mathbf{z}} \int_{\boldsymbol{\pi}} \prod_{p=1}^{P} p(\vec{s}_p|\vec{\pi}_p, \vec{\eta}) p(\vec{\pi}_p|\vec{\alpha}) \prod_{i=1}^{N_p} \prod_{k=1}^{K} p(\vec{x}_i|\vec{\theta}_k)^{z_{ik}} p(\vec{z}_i|\vec{\pi}_p) d\boldsymbol{\pi} \tag{3}$$

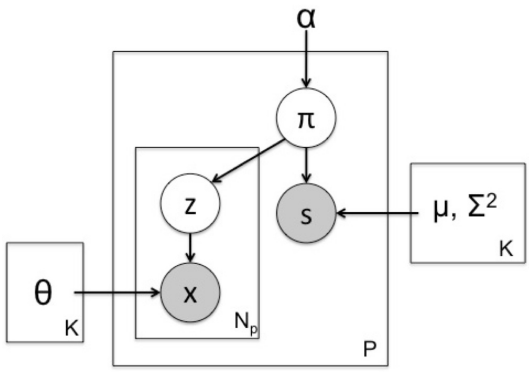

Fig. 3. Graphical Model of CUDIA.

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

$$p(\boldsymbol{\pi}, \mathbf{z}|\vec{\eta}, \vec{\theta}, \vec{\alpha}, \mathbf{x}, \mathbf{s}) = \frac{p(\mathbf{x}, \mathbf{s}, \boldsymbol{\pi}, \mathbf{z}|\vec{\eta}, \vec{\theta}, \vec{\alpha})}{p(\mathbf{x}, \mathbf{s}|\vec{\eta}, \vec{\theta}, \vec{\alpha})}. \tag{4}$$

The key inferential problem is how to calculate this posterior distribution. A generic EM algorithm [Dempster et al. 1976] cannot be applied, since the normalization constant of its posterior distribution in Equation (4) is intractable. Collapsed Gibbs sampling [Liu 1994] 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 [Booth and Hovert 1999] using this approximation technique.

4.1. E-step: Gibbs Sampling

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

$$p(\mathbf{x}, \mathbf{s}, \boldsymbol{\pi}, \mathbf{z}|\vec{\eta}, \vec{\theta}, \vec{\alpha}) = \prod_{p=1}^{P} p(\vec{s}_p|\vec{\pi}_p, \vec{\eta}) p(\vec{\pi}_p|\vec{\alpha}) \prod_{i=1}^{N_p} \prod_{k=1}^{K} p(\vec{x}_i|\vec{\theta}_k)^{z_{ik}} p(\vec{z}_i|\vec{\pi}_p). \tag{5}$$

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

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

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

However, sampling $\vec{\pi}$ is problematic as Equation (6) is not a trivial distribution. Instead of sampling directly from Equation (6), Metropolis-Hastings (MH) algorithm can be used with a proposal density $Dirichlet(\vec{\alpha})$. This algorithm is described in Algorithm 1.

Sampling from a Dirchlet distribution might be computationally heavy in some programming languages such as Numpy in Python[2]. As an alternative, the prior distribution of $\vec{\pi}$ can be replaced by a Logistic Normal distribution or a Uniform distribution

[2]In Numpy, a Dirichlet sample is generated from multiple gamma distributions.

ALGORITHM 1: MH Algorithm using Dirichlet proposal density.

Input: Initial value $\vec{\pi}_p^{(0)}$
Output: Gibbs sample $\vec{\pi}_p^{(I_{Max})}$
$index$ = 0;
repeat
  $\vec{\pi}_p^{(new)} \sim Dir(\vec{\alpha})$;
  $\zeta \sim Uniform(0,1)$;
  Set $n(z_{\cdot k}^{(j)})$ as the count of $z_{\cdot k}^{(j)} = 1$;
  $g(\vec{\pi}_p^{(new)}, \vec{\pi}_p^{(index)}) \leftarrow (p(\vec{s}_p|\vec{\pi}_p^{(new)}, \vec{\eta})p(\vec{\pi}_p^{(new)}|\vec{\alpha})^2)/(p(\vec{s}_p|\vec{\pi}_p^{(index)}, \vec{\eta})p(\vec{\pi}_p^{(index)}|\vec{\alpha})^2)$;
  $Threshold \leftarrow g(\vec{\pi}_p^{(new)}, \vec{\pi}_p^{(j)}) \prod_k^K (\pi_{pk}^{(new)}/\pi_{pk}^{(index)})^{n(z_{\cdot k}^{(j)})}$;
  if $\zeta < Threshold$ **then** $\vec{\pi}_p^{(index+1)} \leftarrow \vec{\pi}_p^{(new)}$, **else** $\vec{\pi}_p^{(index+1)} \leftarrow \vec{\pi}_p^{(index)}$;
until $index < I_{Max}$;

ALGORITHM 2: Gibbs sampling E-Step

Input: $\mathbf{x}, \mathbf{s}, \vec{\eta}, \vec{\theta}, \vec{\alpha}$
Output: $\boldsymbol{\pi}, \mathbf{z}$
$index$ = 0;
repeat
  Sample $\pi_p^{(index)}$ using Algorithm 1;
  Set $E[z_k|\pi_p^{(index)}, \mathbf{x}] \propto p(\mathbf{x}|\vec{\theta}_k)\pi_{pk}^{(index)}$;
  Set $n(z_{\cdot k}) \leftarrow E[z_k|\pi_p^{(index)}, \mathbf{x}]$;
until $index < N_{Gibbs}$;
Set $E[z_k|\mathbf{x}] \propto \sum_{j=1}^{N_{Gibbs}} E[z_k^{(j)}|\pi_p^{(j)}, \mathbf{x}]$;
Set $\vec{\pi}_p \propto \sum E[\vec{z}|\mathbf{x}]$;

by modifying the CUDIA model, so that we can adopt a different proposal density function according to the modified model. In our empirical evaluation, different prior distributions showed marginal differences in their performances. 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 avoided given a large enough size of N_p for low dimensional $\vec{x}$'s.

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 a low dimensional $\vec{x}$ is given, 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 is described in Algorithm 2.

The last line of the algorithm is derived by using the Partition Theorem of conditional expectation [Grimmett and Stirzaker 2001]. As a result, the actual sampling process occurs only in MH sampling (Algorithm 1). In this paper, we used a burning period of 10 samples, and $N_{Gibbs} \approx 50$ to 100 [Agarwal and Chen 2009]. Our empirical results show that with this small number of samples, the algorithm converges with reasonable speed.

4.2. M-step: Parameter Estimation

The 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:

$$\mathbf{S}_i = [s_{1i}, s_{2i}, ..., s_{Pi}]^T,\ \mathbf{W} = diag(N_1, N_2, ..., N_P),\ \mathbf{H} = [\vec{\eta}_1, \vec{\eta}_2, ..., \vec{\eta}_K]^T \tag{8}$$

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

$$\tag{10}$$

Note that

$$\mathbf{H}_{\cdot i} = \begin{pmatrix} \eta_{1i} \\ \eta_{2i} \\ ... \\ \eta_{Ki} \end{pmatrix},\ \hat{\mathbf{\Pi}} = \begin{pmatrix} \hat{\pi}_{11} & \hat{\pi}_{12} & ... & \hat{\pi}_{1K} \\ \hat{\pi}_{21} & \hat{\pi}_{22} & ... & \hat{\pi}_{2K} \\ ... & ... & ... & ... \\ \hat{\pi}_{P1} & \hat{\pi}_{P2} & ... & \hat{\pi}_{PK} \end{pmatrix}. \tag{11}$$

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:

$$\mathbf{S}_i \approx \hat{\mathbf{\Pi}} \cdot \mathbf{H}_{\cdot i} \tag{12}$$

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

$$\mathbf{H}_{\cdot i}^* = (\hat{\mathbf{\Pi}}^T\mathbf{W}\hat{\mathbf{\Pi}})^{-1}\hat{\mathbf{\Pi}}^T\mathbf{W}\mathbf{S}_i. \tag{13}$$

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 a more stable solution, regularization techniques can be incorporated. For example, if a Ridge penalty is used, then $\mathbf{H}$ becomes:

$$\mathbf{H}_{\cdot i}^* = (\hat{\mathbf{\Pi}}^T\mathbf{W}\hat{\mathbf{\Pi}} + \lambda\mathbf{I})^{-1}\hat{\mathbf{\Pi}}^T\mathbf{W}\mathbf{S}_i. \tag{14}$$

The entire inference algorithm is described in Algorithm 3.

ALGORITHM 3: Gibbs CUDIA EM algorithm

Input: $\mathbf{x}, \mathbf{s}$
Output: $\vec{\eta}, \vec{\theta}, \vec{\alpha}$
$index$ = 0;
repeat
  (E-Step) Algorithm 2;
  (M-Step) Learn $\vec{\alpha}$ and $\vec{\theta}$;
  $\mathbf{H}_{\cdot i}^* = (\hat{\mathbf{\Pi}}^T\mathbf{W}\hat{\mathbf{\Pi}} + \lambda\mathbf{I})^{-1}\hat{\mathbf{\Pi}}^T\mathbf{W}\mathbf{S}_i$;
until *Convergence*;

4.3. Deterministic Hard Clustering

The CUDIA model leads to an intuitive deterministic hard clustering algorithm. Starting from the log-likelihood of CUDIA, the hard clustering objective function is obtained

as (see Appendix A for details):

$$\min_{\mathbf{z},\vec{\mu},\vec{\eta}} \sum_{p} \{ \sum_{k,n_p} z_{n_p k} \parallel \vec{x}_{n_p} - \vec{\mu}_k \parallel^2 \} + \beta \parallel \vec{s}_p - \sum_k \frac{\sum_{n_p} z_{n_p k}}{N_p} \vec{\eta}_k \parallel^2 \tag{15}$$

$$= \min_{\mathbf{z},\vec{\mu},\vec{\eta}} \sum_{p,k,n_p} z_{n_p k} \parallel \vec{x}_{n_p} - \vec{\mu}_k \parallel^2 + \frac{\beta}{KN_p} \parallel \vec{s}_p - \sum_k \hat{\pi}_{pk} \vec{\eta}_k \parallel^2 \tag{16}$$

where $\hat{\pi}_{pk} = \frac{\sum_{n_p} z_{n_p k}}{N_p}$ and β is a parameter that determines weights on the group average statistics. Local minima of this objective function can be found by alternating minimization steps between $\mathbf{z}$ and $(\vec{\mu}, \vec{\eta})$ as in Algorithm 4. One iteration of this algorithm

ALGORITHM 4: Deterministic CUDIA Algorithm

Input: $\mathbf{x}, \mathbf{s}$
Output: $\vec{\eta}, \vec{\theta}, \boldsymbol{\pi}, \mathbf{z}$
repeat
(Assignment Step)
$k^* = \arg\min_k \parallel \vec{x}_{n_p} - \vec{\mu}_k \parallel^2 - 2(\vec{s}_p - \mathbf{H}^T \hat{\vec{\pi}}_p)^T \vec{\eta}_k (\frac{\beta}{KN_p})$;
if $k = k^*$ **then** $z_{n_p k} \leftarrow 1$, **else** $z_{n_p k} \leftarrow 0$;
(Update Step)
$\vec{\mu}_k \leftarrow \sum_n (z_{nk} \vec{x}_n)/N_k$, $\vec{\pi}_p \leftarrow \sum_{n_p} \vec{z}_{n_p}/N_p$;
$\mathbf{H}_{\cdot i} \leftarrow (\hat{\mathbf{\Pi}}^T \mathbf{W} \hat{\mathbf{\Pi}} + \lambda \mathbf{I})^{-1} \hat{\mathbf{\Pi}}^T \mathbf{W} \mathbf{S}_i$;
until *Convergence*;

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 the "k-means" algorithm, promising its scalability to massive datasets. Moreover, this algorithm can be used as an initialization step for the probabilistic algorithm (Algorithm 3), 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 property between Exponential family and the family of loss functions represented by Bregman divergences [Banerjee et al. 2005]. Given two vectors $\vec{x}$ and $\vec{\mu}$, the Bregman divergence is defined as:

$$d_\phi(\vec{x}, \vec{\mu}) = \phi(\vec{x}) - \phi(\vec{\mu}) - \langle \vec{x} - \vec{\mu}, \nabla\phi(\vec{\mu}) \rangle \tag{17}$$

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 cannot be appropriately modeled using the Gaussian distribution, as in the Bregman Hard/Soft Clustering algorithms [Banerjee et al. 2005]. Table I shows the relationship between specific Bregman divergences and their corresponding Exponential family distributions. The results in [Banerjee et al. 2005] state that minimizing the negative log-likelihood is the same as minimizing the corresponding expected Bregman divergence. For example, if clustering data points are generated by a mixture of Gaussian distributions, the

Table I. Bregman divergence and Exponential family.

Distribution	$p(x;\theta)$	μ	$\phi(\vec{\mu})$	$d_\phi(\vec{x},\vec{\mu})$
1-D Gaussian	$\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(x-a)^2}{2\sigma^2})$	a	$\frac{1}{2\sigma^2}\mu^2$	$\frac{1}{2\sigma^2}(x-\mu)^2$
d-D Gaussian	$\frac{1}{\sqrt{(2\pi\sigma^2)^d}}\exp(-\frac{\|\vec{x}-\vec{a}\|^2}{2\sigma^2})$	$\vec{a}$	$\frac{1}{2\sigma^2}\parallel\vec{\mu}\parallel^2$	$\frac{1}{2\sigma^2}\parallel\vec{x}-\vec{\mu}\parallel^2$
1-D Exponential	$\lambda\exp(-\lambda x)$	$1/\lambda$	$\mu\log\mu-\mu$	$x\log(\frac{x}{\mu})-(x-\mu)$
d-D Multinomial	$\frac{N!}{\prod_{j=1}^d x_j!}\prod_{j=1}^d q_j^{x_j}$	$[Nq_j]_{j=1}^{d-1}$	$\sum_{j=1}^d \mu_j\log\frac{\mu_j}{M}$	$\sum_{j=1}^d x_j\log\frac{x_j}{\mu_j}$

maximum likelihood parameters can be obtained by minimizing the squared loss function, which is the corresponding Bregman divergence for Gaussian distributions. Using this bijection and adopting the idea from the Bregman Hard Clustering algorithm, the deterministic algorithm of CUDIA can be extended by modifying the assignment step of Algorithm 4 as follows:

—**Assignment Step**

$z_{n_pk^*} \leftarrow 1$, if

$$k^* = \arg\min_k d_\phi(\vec{x}_{n_p},\vec{\mu}_k) - 2(\vec{s}_p - \mathbf{H}^T\hat{\vec{\pi}}_p)^T\vec{\eta}_k(\frac{\beta}{KN_p}) \tag{18}$$

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

ϕ can be chosen based on the distribution of $\vec{x}$ and the update step remains the same. Note that $\vec{s}_p$ follows a Gaussian distribution according to the Central Limit Theorem regardless of the underlying distribution of $\vec{x}_u$'s. Thus, the second term in Equation (18) remains the same, but the first term is changed to capture various Exponential distributions. The update step remains the same as in Algorithm 4, since a unique minimizer of a Bregman divergence is given by its mean (see Proposition 1 in [Banerjee et al. 2005]).

This extended algorithm captures various distributions while maintaining the original complexity. Furthermore, the linkage between Bregman divergences 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.

5. 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:

$$p(\vec{x}_u|\vec{x}_o,\vec{\pi}_p) = \sum_k p(\vec{x}_u,z_k|\vec{x}_o,\vec{\pi}_p) = \sum_k \frac{p(\vec{x}_u,z_k,\vec{x}_o,\vec{\pi}_p)}{p(\vec{x}_o)} \tag{19}$$

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

Equation (20) provides the exact imputation formula for any $p(\vec{x}_u|z_k)$, depending on the form of the cluster-conditional pdf of the unobserved features. For example, if $\vec{x}_u|z_k$ is generated from an Exponential family distribution with a mean $\vec{\eta}_k$ and a covariance $\delta^2\mathbf{I}$, the imputation formula obtained is:

$$\hat{\vec{x}}_u \leftarrow \sum_{k=1}^K \vec{\eta}_k E[z_k|\vec{x}_o,\vec{\pi}_p]. \tag{21}$$

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:

$$E[z_k|\vec{x}_o, \vec{\pi}_p] = \frac{\pi_{pk}\mathbf{exp}(-d_\phi(\vec{x}_o, \vec{\mu}_k))}{\sum_l \pi_{pl}\mathbf{exp}(-d_\phi(\vec{x}_o, \vec{\mu}_l))}. \tag{22}$$

For example, if $\vec{x}_o$'s are generated from a mixture of Gaussians, which means $d_\phi(\vec{x}, \vec{\mu}) = \frac{1}{2\sigma^2} \parallel \vec{x} - \vec{\mu} \parallel^2$, then Equation (22) becomes:

$$E[z_k|\vec{x}_o, \vec{\pi}_p] \propto \pi_{pk}\mathbf{exp}(-\frac{\parallel \vec{x}_o - \vec{\mu}_k \parallel^2}{2\sigma^2}). \tag{23}$$

As another example, if $\vec{x}_o$'s are generated from a mixture of d-D multinomial distributions, using the bijective relationship, we get $d_\phi(\vec{x}, \vec{\mu}) = \sum_{j=1}^{d} x_j \log \frac{x_j}{\mu_j}$ (KL-divergence). Then, Equation (22) becomes:

$$E[z_k|\vec{x}_o, \vec{\pi}_p] \propto \pi_{pk} \prod_{j=1}^{d} (\frac{x_{oj}}{\mu_{kj}})^{x_{oj}}. \tag{24}$$

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.

The imputation formula, Equation (21), calculates an unbiased estimate for $\vec{x}_u$. Moreover, the variance of the imputation is inversely proportional to the size of the data. The detailed properties of this imputation are derived and explained in Appendix B.

5.1. Extension to Different Aggregation Levels

In this section, we show how CUDIA can be applied to data that have differently aggregated variables, by considering the case when two variables aggregated at two different levels. The extension from two variables/levels to more will be clearer when we fully describe the process.

We presume that these two variables are present at individual-level, namely $\vec{x}_u^a$ and $\vec{x}_u^b$, but not observed at individual-level. Thus, the underlying "complete" dataset is represented as $\mathcal{D}_x = \{(\vec{x}_o, \vec{x}_u^a, \vec{x}_u^b)_1, (\vec{x}_o, \vec{x}_u^a, \vec{x}_u^b)_2, ..., (\vec{x}_o, \vec{x}_u^a, \vec{x}_u^b)_N\}$. For each unobserved variable, we have different partitionings, $\mathcal{P}^a$ and $\mathcal{P}^b$ where $\mathcal{P}^a \neq \mathcal{P}^b$. Then, aggregated variables, $\vec{s}_{p^a}$ and $\vec{t}_{p^b}$, are derived from their corresponding partitionings:

$$\vec{s}_{p^a} = \frac{1}{N_{p^a}} \sum_{i=1}^{N} \vec{x}_{ui}^a \mathbf{1}_{(\vec{x}_{ui}^a \in \mathcal{D}_x^{p^a})} \text{ and } \vec{t}_{p^b} = \frac{1}{N_{p^b}} \sum_{i=1}^{N} \vec{x}_{ui}^b \mathbf{1}_{(\vec{x}_{ui}^b \in \mathcal{D}_x^{p^b})}.$$

As in the CUDIA process, a probabilistic generative process can be used to model the observed variables $\vec{x}_o$, $\vec{s}_{p^a}$, and $\vec{t}_{p^b}$; however, the resultant generative process would involve necessarily a deeper hierarchy structure, and its extension to non-nested partitionings would be also problematic.

In lieu of building more complex probabilistic models, we introduce two simple approaches, which are easily extensible to more number of different aggregation levels. The first approach is a brute-force parallel application of the CUDIA imputation. In this approach, each aggregated variable is paired with the individual-level variable(s)

$\vec{x}_o$, and then the CUDIA imputation is applied to each pair of variables:

$$\hat{\vec{x}}_u^a \leftarrow \text{CUDIA-Imputation}(\vec{x}_o, \vec{s}_{p^a})$$
$$\hat{\vec{x}}_u^b \leftarrow \text{CUDIA-Imputation}(\vec{x}_o, \vec{t}_{p^b})$$

However, this approach ignores information between $\vec{s}_{p^a}$ and $\vec{t}_{p^b}$, which may be useful in some cases. To utilize such information, we suggest the second approach, which sequentially imputes individual-level variables. In the second approach, one aggregated variable is firstly paired with the individual-level variable, then its individual-level value is estimated using the CUDIA imputation. This imputed variable can be further utilized as another (estimated) individual-level variable when imputing the other aggregated variable:

$$\hat{\vec{x}}_u^a \leftarrow \text{CUDIA-Imputation}(\vec{x}_o, \vec{s}_{p^a})$$
$$\hat{\vec{x}}_u^b \leftarrow \text{CUDIA-Imputation}((\vec{x}_o, \hat{\vec{x}}_u^a), \vec{t}_{p^b})$$

These two approaches can be easily implemented from the original CUDIA algorithm with minor modification, thus we primarily focus on the common partitioning case throughout. Note that the performance of these two approaches, and other possible extensions need more thorough analysis, and we leave these for our future work.

6. EXPERIMENTAL RESULTS

In this section, we provide two kinds of experimental results. (i) First, imputation quality of the CUDIA model is assessed using a simulated mixture of Gaussians data. (ii) Then, its applicability to predictive modeling[3] is discussed using the data from the Behavioral Risk Factor Surveillance System (BRFSS).

6.1. Imputation Quality

We demonstrate CUDIA's properties for imputation using a simulated dataset. The dataset is generated by a mixture of three 2-D Gaussians ($K = 3$) as shown in Figure 4(a). We generated 2000 samples, then partitioned into ten groups according to the randomly generated mixture coefficients ($\mathbf{\Pi}$). Thus, $P = 10$ and $N_p = 200\ \forall p$. The mixture coefficients are:

$$\mathbf{\Pi} = \begin{pmatrix} \pi_{1,1} & \pi_{1,2} & \pi_{1,3} \\ \pi_{2,1} & \pi_{2,2} & \pi_{2,3} \\ \cdots & \cdots & \cdots \\ \pi_{10,1} & \pi_{10,2} & \pi_{10,3} \end{pmatrix} = \begin{pmatrix} 0.166 & 0.023 & 0.811 \\ 0.270 & 0.387 & 0.343 \\ \cdots & \cdots & \cdots \\ 0.580 & 0.174 & 0.246 \end{pmatrix}. \quad (25)$$

From this dataset, we assume the first column (x-axis column) is the auxiliary individual-level information, and the second column (y-axis column) is aggregated within each partition. If the unobserved individual-level second column is imputed by its corresponding aggregated value i.e. everyone in the same partition shares the same feature value, then the resultant dataset is as in Figure 4(b). We can observe that this naïve imputation scheme does not reflect the underlying heterogeneous distributions. Next, we run our CUDIA model over these individual- and aggregate-level datasets to discover the underlying mixture distributions. Figure 4(c) shows the CUDIA imputation based on Equation (21). The CUDIA imputation captures the hidden underlying mixture distributions, and the imputation follows the mean statistics of each intrinsic cluster. Figure 5 shows the Mean Squared Error (MSE) between the true and the imputed data points. The CUDIA imputation achieves lower MSE as well as lower variance compared to the naïve imputation.

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

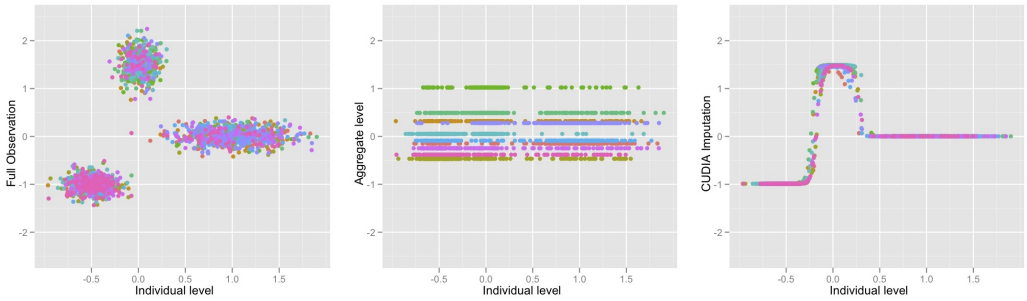

Fig. 4. (a) Simulated dataset with the individual level data (x-axis) and the true individual values (y-axis) for the aggregate data (left). (b) Direct imputation using the aggregate level data (center). (c) CUDIA imputation (right). Each partition is represented by a different color.

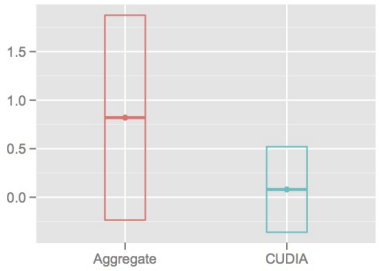

Fig. 5. Imputation accuracy (Mean Squared Error) on the simulated dataset.

6.2. BRFSS dataset

In this section, we provide the experimental results using a real world dataset in various settings.

Dataset description. We demonstrate the proposed method using the BRFSS 2009 dataset. BRFSS (Behavioral Risk Factor Surveillance System) [4] is the world's largest telephone health survey since 1984, tracking health conditions and risk behaviors in the United States. The data are collected monthly in all 50+ states in the United States. The dataset contains information on a variety of diseases like diabetes, hypertension, cancer, asthma, HIV etc, and in this paper, we mainly focus on diabetes rather than other diseases [5]. The 2009 dataset contains more than 400,000 records and 405 variables and the diabetic (positive class) ratio is 12%. Empty and less informative columns are dropped and we finally chose six variables to perform our experiments. The selected variables are: Age, Body Mass Index (BMI), Education level, Income level, Hypertension and Hyper-cholesterol.

In many cases, revealing personal disease records or medical conditions can be problematic, or even cause traumatic situations e.g. HIV. Rather than having the raw individual disease records, suppose the data is provided at aggregate level such as state-level or county-level summaries. The aggregation level we chose in this paper is the US census division as shown in Figure 6. For each division, the important feature distributions are described in Figure 7. Although the distributions are slightly different across each division, we can observe that they do not reflect the true clusters of the individual-level features.

[4]http://www.cdc.gov/brfss/
[5]Targets are chosen arbitrarily to illustrate the applicability of the CUDIA framework.

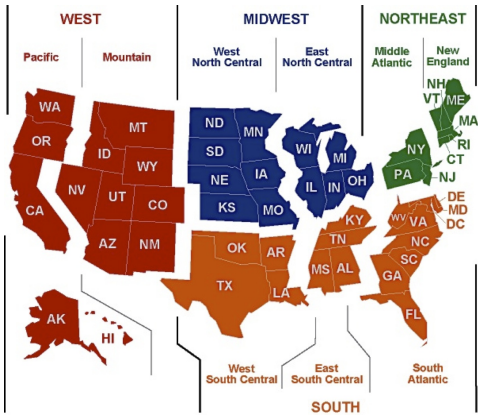

Fig. 6. Census Regions and Divisions of the United States. This picture is adopted from http://www.eia.gov /emeu/reps/maps/us_census.html.

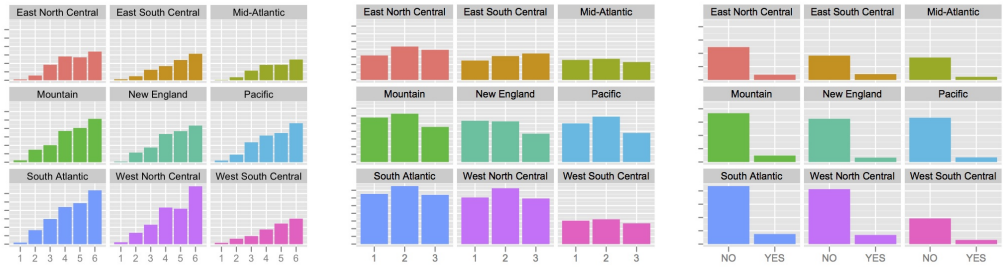

Fig. 7. BRFSS dataset description for each division. (a) Age (left), (b) BMI (center) and (c) Diabetes (right).

6.3. Aggregated Target

In this section, we focus on diabetes records that are aggregated at the US division level as in Figure 7(c). We use (i) Age, (ii) BMI, (iii) Education level and (iv) Income level as the individual-level features and the aggregated diabetes ratio as the aggregate-level features. This individual-level dataset along with the division-level aggregated diabetes records are given as the inputs to the CUDIA model. Although the individual-level features are numeric values, their values are grouped ranging from three to six levels. To prevent the singular variance problem in the EM algorithm, their values are perturbed with a negligible Uniform noise before the learning process. After all the parameters in the CUDIA model are learned, the hidden individual-level diabetic condition (diabetes or healthy) is imputed based on the underlying distribution. Note that the imputation formula produces a probabilistic estimate of how he/she is likely to have diabetes. Hence the imputation quality can be measured by Receiver Operating Characteristic (ROC) curve and Area Under ROC (AUROC) values. Figure 8 shows the ROC curves and the AUROC values from the aggregated diabetes dataset, ranging from $K = 3$ to $K = 9$. We compare the performance of CUDIA with the base model, which makes everyone in the same division share the corresponding average diabetic rate. We can observe that all CUDIA models outperforms the base model in ROC space.

The CUDIA model provides another valuable information about the data, which is the underlying distribution. Table II shows the learned parameters from the model. Noticeably, Cluster 7 exhibits a high risk for diabetes. Their profiles can be described as “higher age”, “obese” and “middle-class”, where this relationship between obesity

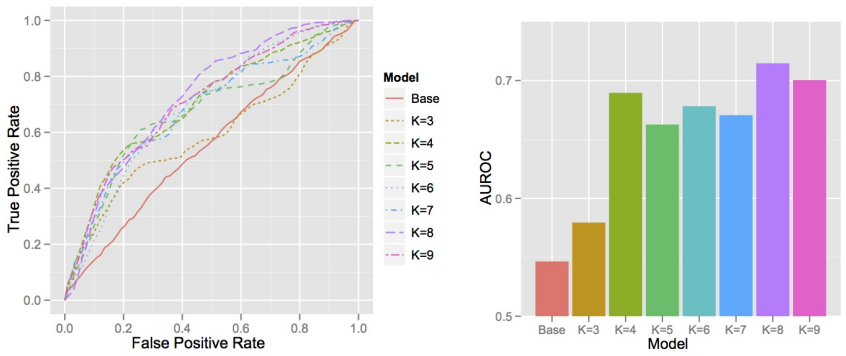

Fig. 8. Results from the aggregated diabetes dataset. (a) ROC curves (left) and (b) AUROC values (right).

Table II. θ and η values from the aggregated diabetes dataset.

Cluster Index	**Age** (θ_1)	**BMI** (θ_2)	**Education** (θ_1)	**Income** (θ_4)	**Diabetes** (η)
1	**3.904**	2.015	2.516	**4.999**	**0.065**
2	4.639	1.000	2.602	2.767	0.105
3	**4.136**	**1.790**	**4.000**	**4.999**	**0.068**
4	2.909	2.000	2.726	3.389	0.121
5	4.689	2.000	2.635	3.126	0.135
6	4.270	1.999	2.261	1.002	0.124
7	**4.534**	**2.999**	**2.480**	**2.617**	**0.233**
8	6.000	2.000	2.617	3.015	0.119
9	4.936	2.000	0.997	2.001	0.126

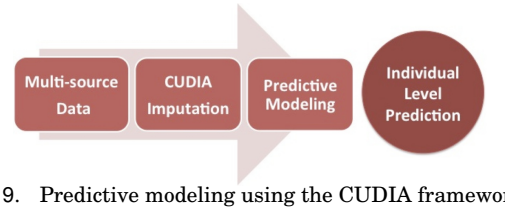

Fig. 9. Predictive modeling using the CUDIA framework.

and diabetes coincides with the medical research literature [Steppan et al. 2011]. On the other hand, Cluster 3 shows a lower risk, and their profiles can be summarized as "slim", "high education" and "high level income". Note that these cluster parameters are learned without accessing the individual diabetes information.

6.4. Aggregated Features

In this section, we consider a different setting based on the same dataset, in which the target variable is available at individual level, but other important features are masked due to privacy or legal issues. In this case, we can impute the masked features using the CUDIA model, then propagate its results to various predictive modeling algorithms. Figure 9 describes the main idea of this approach.

In this setting, Age and BMI are the individual features, Hypertension and High-cholesterol are the masked features and Diabetic condition is the target. The masked features are aggregated using the US census division mapping, and the target is only used in the predictive modeling i.e. the target is not used before the predictive modeling. As the formulated problem is a binary prediction problem, we can use any binary classifier such as SVM, Logistic regression, decision tree, Naïve Bayes, etc. If a regression problem is formulated for this setting, one can use other regression techniques such as Lasso and Ridge regression [Park and Ghosh 2011], [Park and Ghosh 2012].

Table III. θ and η values from the aggregated features dataset.

Cluster Index	**Age** (θ_1)	**BMI** (θ_2)	**Hypertension** (η_1)	**High-cholesterol** (η_2)
1	2.999	1.999	0.374	0.405
2	3.000	2.000	0.380	0.415
3	4.569	**1.000**	**0.199**	0.354
4	3.999	1.999	0.271	0.438
5	**5.999**	**1.999**	**0.437**	**0.541**
6	5.999	2.000	0.395	0.396
7	5.000	1.999	0.408	0.397
8	**4.486**	**2.999**	**0.647**	**0.504**
9	1.860	1.999	0.384	0.413

In this paper, we demonstrate this predictive modeling framework using a Logistic regression family, decision trees, Random Forests, and SVM.

Table III shows the estimated parameters when $K = 9$ from the dataset. The people belonging to Cluster 8 have higher hypertension risk as well as high-cholesterol risk. Their observed individual features are centered at the "higher age" and "obese" centroid [Carmelli et al. 1994]. On the other hand, the people from Cluster 3 have lower hypertension risk while their ages are comparably high. But interestingly, their BMI's are very low, and this supports the result. For the rest of the predictive modeling tasks, we used $K = 9$ and $\lambda = 0.1$, where λ is in Equation (14).

As in Section 6.3, the aggregated variables are estimated at individual-level using the CUDIA imputation framework. These imputed variables now form individual-level predictors, together with the two individual-level variables, Age and BMI. This newly created dataset, namely the CUDIA dataset, can be plugged into various predictive models for diabetes. On the other hand, without the CUDIA imputation, the best way to utilize the aggregated variables is to view them as the individual-level representatives, and we name this dataset as the baseline dataset. These two datasets along with the "complete" dataset (the ground truth individual-level variables) will be compared in various predictive models. We expect the CUDIA dataset would result in better prediction than the baseline dataset, as the CUDIA imputed variables will be closer to the true individual-level values than the coarse baseline variables. Figure 10 shows the imputation quality of the CUDIA imputed features as well as the baseline variables. Note that the original hypertension and high-cholesterol variables are binary variables at individual-level, while the CUDIA imputation results in numeric estimates, which are basically weighted averages of cluster centroids in Table III. These numeric estimates are better aligned with the underlying true individual-level values, and we measured AUROC's for different values of K's as well as the baseline. As can be seen, the aggregated variables (the baseline) across nine Census Divisions show almost no predictive power on their original individual-level values, resulting in nearly 0.5 AUROC values. On the other hand, the CUDIA imputed variables tend to follow the original individual-level values, even with very small K's.

6.4.1. Logistic regression with aggregated features. In some cases, the relationship between the aggregated features and the target might be of primary research interest. If we have available individual side information (in this case, age and BMI) along with the aggregated features, we can use either the CUDIA imputed values or the aggregate values (baseline approach). The Logistic regression equation is given as:

$$p(Diabetes) \sim logit(\beta_{Hyper}(Hypertension) + \beta_{Chol}(Cholesterol) + \beta_{Const}). \quad (26)$$

Figure 11 shows the Logistic regression results from three different kinds of datasets: (i) Baseline dataset (direct aggregate variable imputation), (ii) Complete dataset (full individual observation) and (iii) CUDIA dataset (CUDIA imputation). In Figure 11(a), we can observe that the coefficients from the CUDIA dataset mimics the

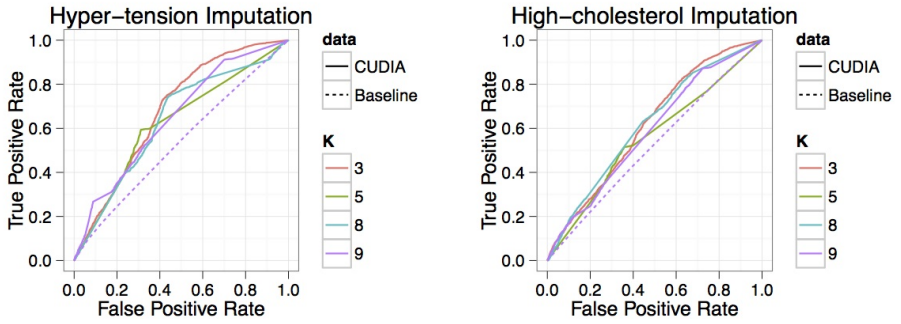

Fig. 10. ROC charts for Hyper-tension (left) and High-cholesterol (right) imputed features. The CUDIA imputed features are closer to the ground truth than the baseline features (aggregated values).

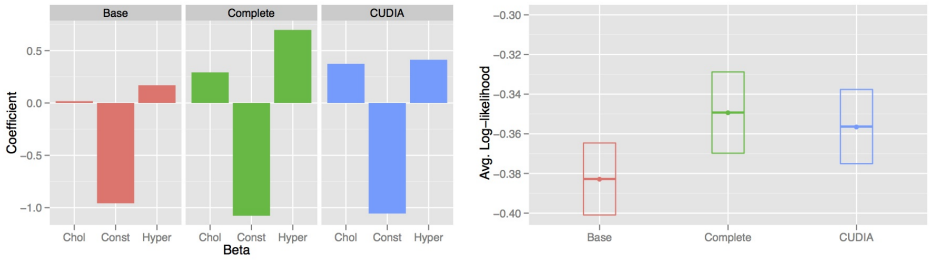

Fig. 11. Results from the Logistic regression only with the masked variables. (a) Coefficients (β) (left), (b) Average Log-likelihood on the test sets (right).

coefficients of the complete dataset quite well. 5-fold cross validation is performed and the average log-likelihood values of the hold-out samples are recorded. Figure 11(b) shows that the CUDIA dataset outperforms the baseline dataset, while it performs slightly worse than the complete dataset.

6.4.2. Logistic regression with L1 constraints. The rest of the experiments use a combination of the individual- and the aggregate-level features. The dependent variables are two individual variables (Age and BMI) and two aggregate variables (Hypertension and High-cholesterol). Unfortunately, many features in the BRFSS dataset are interdependent such as "Age" and "Income level", "BMI" and "Hypertension", etc. This property becomes even worse when the interdependent numeric values are grouped into a few number of bins, as in the BRFSS dataset. This type of problems can be alleviated if we adopt shrinkage methods, also known as regularizers such as $L1$ or $L2$ [Hastie et al. 2009]. In this paper, we demonstrate two widely used regularizers, $L1$ and $L2$.

The $L1$ regularizer is known to generate a sparser solution compared to a normal regression [Cawley et al. 2006], which can be regarded as an automatic feature selection technique. Figure 12 shows the results from the $L1$ Logistic regression. From Figure 12(a), we can observe that the hypertension affects the most in both the complete and the CUDIA datasets, but not in the baseline dataset. The coefficients for the aggregate variables from the baseline dataset are actually zeroed out due to the effect of the $L1$ regularizer. Furthermore, the average log-likelihood values from 5-cv show that the CUDIA imputation is more effective in this predictive task than the baseline imputation.

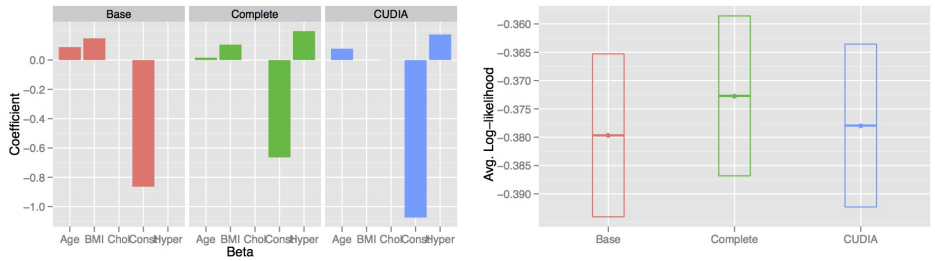

Fig. 12. Results from the Logistic regression with L1 constraints. (a) Coefficients (β) (left), (b) Average Log-likelihood on the test sets (right).

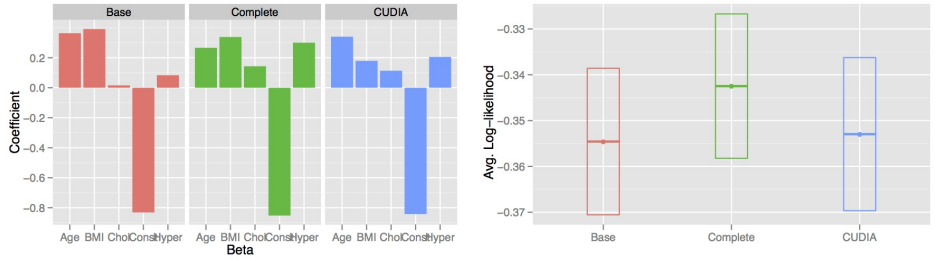

Fig. 13. Results from the Logistic regression with L2 constraints. (a) Coefficients (β) (left), (b) Average Log-likelihood on the test sets (right).

Fig. 14. Results from the Decision trees. (a) BRFSS dataset (left), (b) BRFSS and KFF datasets (right).

6.4.3. Logistic regression with L2 constraints. The results obtained on applying an $L2$ constraints (Ridge regression) are shown in Figure 13. Unlike the $L1$ case, all the coefficients have non-zero values in Figure 13(a). Note that the coefficients from the complete and the CUDIA datasets have very similar weights. Again, from Figure 13(b), we observe that the CUDIA imputation is more effective than the baseline dataset.

6.4.4. Decision Tree. Decision trees are recursive rule based classifiers. We demonstrate the impact of CUDIA using two kinds of decision trees based respectively on: (i) Gini criterion [Breiman 1984] and (ii) Entropy criterion [Quinlan 1993]. We used the decision tree package from KNIME[6], and the Minimum Description Length (MDL) principle is used for pruning.

Figure 14(a) shows the results from the decision trees. The performance is measured using Area under Receiver Operating Characteristic (AUROC) curve in both cases. Surprisingly, the CUDIA imputation recorded almost the same performance as the complete dataset. Originally, the CUDIA model is designed to model the underlying

[6]http://www.knime.org

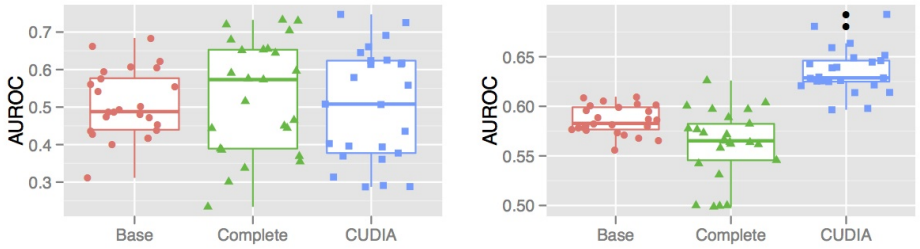

Fig. 15. Results from SVM (left) and Random Forests (right). Average AUROC's from five runs of 5-fold cross validation are presented.

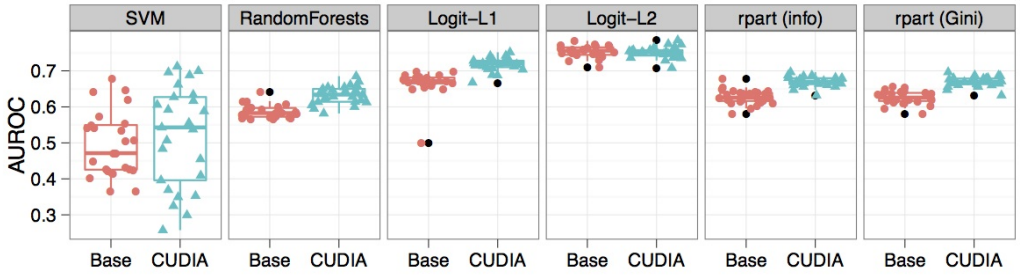

Fig. 16. Measured AUROC Results from various classifiers. Average AUROC's from five runs of 5-fold cross validation are presented. We used `e1071` for SVM, `randomForest` for Random Forests, `glmnet` for $L1$ and $L2$ logistic regressions, and `rpart` for decision trees in R packages. The presented results are based on the default settings of the packages except minor parameter changes. Note that our primary objective is to check the improvement through the CUDIA imputation against the baseline dataset, not to compare the performance of the classification algorithms. We remark that the best performance of each algorithm might be different from these results.

distribution, then the individual values are imputed utilizing the learned conditional distributions. As the recursive decision tree algorithms more focus on the conditional distributions between the target and the features than the individual values themselves, the CUDIA model shows its strength especially in decision tree algorithms.

6.4.5. Support Vector Machine and Random Forests. We provide two more demonstrative examples of the CUDIA imputation framework using Support Vector Machine (SVM) and Random Forests.[7] Figure 15 shows the results from both SVM and Random Forests. We used linear kernel for the SVM classifier, and the default setting for the rest of the parameters. As can be seen, the CUDIA imputed dataset provides better classification results than the baseline datasets in both experiments, and it performs even better than the complete dataset in the Random Forests example.

6.4.6. Performance Analysis. Figure 16 shows the performance comparison between the baseline dataset and the CUDIA imputed dataset across various classifiers. The experimental results from SVM, decision trees, and Random Forests show significant performance improvement using the CUDIA imputed dataset, contrary to the slight improvement in the logistic regression experiments. We contemplate two possible explanations for this kind of results. First, the two CUIDA-imputed features, hypertension and high-cholesterol, turn out to be rather strongly correlated with the individual-

[7]We used `e1071` and `randomForest` R packages for SVM and Random Forests, respectively.

level BMI feature, violating the feature independence assumption of linear models. Although the imputed features are correlated with the original features, these imputed features are based on underlying clusters, providing richer information about the dataset. Therefore, classifiers resistant to feature dependencies, such as SVM and decision trees, may be able to produce better predictive performance using this additional information. Second, the hypertension and high-cholesterol features are originally binary features, but the CUDIA-imputation results in real-valued estimates at individual-level, which can be interpreted as probabilistic estimates of having hypertension and high-cholesterol. In decision trees, these numeric features provide larger freedom of splitting rules differing a cutting threshold, while binary features can result in only two-way splits. In a case of datasets with aggregated binary features, the CUDIA imputation may provide higher degree of freedom in decision rules, as long as the numeric estimates are close to the features. Thus, the CUDIA imputation not only helps to reconstruct the individual values of the aggregated features, but also supports the predictive modeling using the imputed features.

6.5. Aggregate Features from a Difference Source: Multi-source example

In this section, we demonstrate a multi-source integration example using CUDIA. We use the BRFSS dataset as the individual-level information and the Kaiser Family Foundation dataset as the aggregate-level information. The Kaiser Family Foundation (KFF) is a non-profit, private foundation, which focuses on the major healthcare issues facing the nation. Statehelthfacts.org is a project of KFF, which provides various health-related statistics for all 50 states in the US. From the dataset, we selected two state-level summaries: (i) Average Fruit/Vegetable Consumption and (ii) Average Heart Disease Rate. The state-level summaries are weighted by the BRFSS sample selection bias, then averaged to make the US divisional statistics. Thus, we have Age and BMI as the individual-level data from the BRFSS dataset, and the adjusted Fruit/Vegetable Consumption and Heart Disease Rate as the aggregate-level data from the KFF dataset.

Figure 14(b) shows the results from the decision trees when the individual Diabetic condition is set as the target. We compare the performance of the CUDIA imputation with two other datasets: (i) without using the aggregate information i.e. only the BRFSS dataset and (ii) with division-level direct imputation (Base model). We used the same decision trees as in the previous experiment. The CUDIA dataset exhibits the best performance among the approaches considered.

The distributions of the imputed variables are shown in Figure 17. Note that we do not have the ground truth individual information for the KFF dataset. Although we cannot measure the imputation accuracy, we can check the quality of the results through the distributions. From Figure 17(a), we can observe that higher risk groups for heart disease contain more people with higher age and higher BMI. Moreover, fruits and vegetables are consumed more by people with lower BMI.

7. CONCLUDING REMARKS

In this paper, aggregated statistics over certain partitions are utilized to identify clusters and impute features that are observed only as aggregated values. The imputed features are further used in predictive modeling, leading to improved performance. The experiments provided in this paper are illustrative of the generality of the proposed 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.

In the CUDIA framework, the aggregate data do not need to be the actual average of the underlying individual-level data. The CLT approximation in CUDIA provides

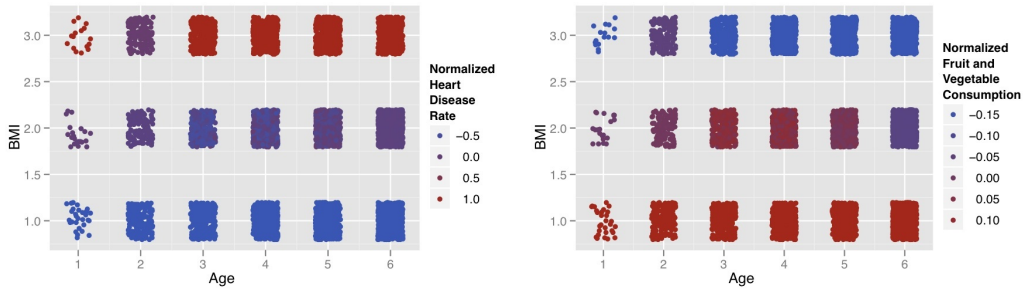

Fig. 17. Distributions of Imputed Features. (a) Heart Disease Rate (left), (b) Fruit and Vegetable Consumption (right). Age and BMI are jittered to visualize.

flexibility in this and many other practical settings. For example, in the UK census, some aggregate data are calculated using a 10% sample to maintain confidentiality. The observed statistics are not the same as the true sample average, thus the direct application of the model of Figure 2(b) is no longer valid. However, the difference between the sub-sampled average and the true sample average can be modeled using a Normal distribution, which fits the key assumption of the CUDIA approximation. As another example, to maintain confidentiality or privacy, a popular technique is to add noise to the true values. Additive Laplace or Gaussian noise are known to guarantee (ϵ, δ)-differential privacy [Dwork 2006] under certain assumptions [Dwork et al. 2006], [Dwork et al. 2006]. Adding a Gaussian noise exactly fits the assumption in the CUDIA model, so that the CUDIA model becomes no longer an approximation in this case.

CUDIA is quite scalable, and in particular, the deterministic hard clustering version can be readily applied to massive datasets. Furthermore, the square loss function on $\vec{x}_o$ can be generalized to all Bregman divergences, or equivalently, one can cater to any noise function from the exponential family of probability distributions [Banerjee et al. 2005]. One restriction of the current model is that the number of clusters (K) cannot be more than the number of partitions (P) specified by the data provider. This is why we had to stop at $K = 9$ for several of the results even though the performances were improving with increasing K. For the aggregate variables from different partitions, the CUDIA framework can be applied in either parallel or sequential way by dividing the problem with having one aggregate variable per each, and we leave the implementation and evaluation of these approaches as our future work.

ACKNOWLEDGMENTS

This work is supported by NSF IIS-1016614 and by TATP grant 01829. We would like to thank Anurekha Ramakrishnan for pre-processing and cleaning the BRFSS dataset.

REFERENCES

ACHEN, C. H. AND SHIVELY, W. P. 1995. *Cross-level Inference*. The University of Chicago Press.

AGARWAL, D. AND CHEN, B.-C. 2009. Regression-based Latent Factor Models. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

ASUNCION, A., WELLING, M., SMYTH, P., AND TEH, Y. 2009. On smoothing and inference for topic models. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*. 27–34.

BANERJEE, A., MERUGU, S., DHILLON, I. S., AND GHOSH, J. 2005. Clustering with Bregman Divergences. *Journal of Machine Learning Research 6*, 1705–1749.

BLEI, D. M., NG, A. Y., AND JORDAN, M. I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 993–1022.

BOOTH, J. G. AND HOVERT, J. P. 1999. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society: Series B 61*, 265–285.

BREIMAN, L. 1984. *Classification and regression trees*. Wadsworth International Group.

BROWNSTONE, D. AND VALLETTA, R. 2001. The bootstrap and multiple imputations: Harnessing increased computing power for improved statistical tests. *Journal of Economic Perspectives 15,* 4, 129–141.

CARMELLI, D., CARDON, L. R., AND FABSITZ, R. 1994. Clustering of hypertension, diabetes, and obesity in adult male twins: same genes or same environments? *American Journal of Human Genetics 55,* 3, 566–573.

CAWLEY, G. C., TALBOT, N. L., AND GIROLAMI, M. 2006. Sparse multinomial logistic regression via bayesian l1 regularization. In *Proceedings of the 19th Annual Conference on Neural Information Processing Systems*. 209–216.

DEMPSTER, A. P., LAIRD, N. M., AND RUBIN, D. B. 1976. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society: Series B 39*.

DUNCAN, O. D. AND DAVIS, B. 1953. An alternative to ecological correlation. *American Sociological Review 18*, 665–666.

DWORK, C. 2006. Differential privacy. In *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming*. Vol. 4052. 1–12.

DWORK, C., KENTHAPADI, K., MCSHERRY, F., MIRONOV, I., AND NAOR, M. 2006. Our data, ourselves: Privacy via distributed noise generation. In *Proceedings of the 25th International Cryptology Conference (EUROCRYPT)*. 486–503.

DWORK, C., MCSHERRY, F., NISSIM, K., AND SMITH, A. 2006. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Theory of Cryptography Conference*.

EMAM, K. E. AND FINEBERG, A. 2009. An overview of techniques for de-identifying personal health information. *Social Science Research Network*.

FREEDMAN, D. A. 1999. Ecological inference and the ecological fallacy. Technical Report 549, Department of Statistics, University of California Berkeley, CA 94720. October.

GOODMAN, L. 1953. Ecological regression and the behavior of individuals. *American Sociological Review 18*, 663–664.

GOODMAN, L. 1959. Some alternatives to ecological correlation. *American Journal of Socialogy 64*, 610–625.

GRIMMETT, G. AND STIRZAKER, D. 2001. *Probability and Random Processes* Third Ed. Oxford, Chapter 3.7, 67.

HASTIE, T., TIBSHIRANI, R., AND FRIEDMAN, J. 2009. *The Elements of Statistical Learning* Second Ed. Springer.

HENRY, K. A. AND BOSCOE, F. P. 2008. Estimating the accuracy of geographical imputation. *International Journal of Health Geographics*.

HIPAA COMPLIANCE ASSISTANCE. 2003. Summary of the HIPAA Privacy Rule. *http://www.hhs.gov/ocr/privacy/hipaa/understanding/summary/privacysummary.pdf*.

JACKSON, C., BEST, N., AND RICHARDSON, S. 2008. Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors. *Journal of Royal Statistical Society: Series A 171*, 159–178.

JACKSON, C., BEST, N., AND RICHARDSON, S. 2009. Bayesian graphical models for regression on multiple data sets with different variables. *Journal of Biostatistics 10,* 2, 335–351.

KING, G. 1997. *A Solution to the ecological inference problem: reconstructing individual behavior from aggregate data*. Princeton University Press.

LIU, J. S. 1994. The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association 89,* 427, 958–966.

PARK, Y. AND GHOSH, J. 2011. A generative framework for predictive modeling using variably aggregated, multi-source healthcare data. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Workshop on Medicine and Healthcare*. 27–32.

PARK, Y. AND GHOSH, J. 2012. A probabilistic imputation framework for predictive analysis using variably aggregated, multi-source healthcare data. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*.

QUINLAN, J. R. 1993. *C4.5: prgrams for machine learning*. Morgan kaufmann.

ROBINSON, W. S. 1950. Ecological correlations and the behavior of individuals. *American Sociological Review 15*, 351–357.

RUBIN, D. B. 2004. *Multiple Imputation for nonresponse in surveys*. Wiley-IEEE.

STEPPAN, C. M., BAILEY, S. T., BAHT, S., BROWN, E. J., BANERJEE, R. R., WRITHE, C. M., PATEL, H. R., AHIMA, R. S., AND LAZAR, M. A. 2011. The hormene resistin links obesity to diabetes. *Nature 209*, 307–312.

TABACHNICK, B. G. AND FIDEL, L. S. 2001. *Using multivariate statistics* 4th Ed. Allyn and Bacon.

WAKEFIELD, J. AND SALWAY, R. 2001. A statistical framework for ecological and aggregated studies. *Journal of Royal Statistical Society: Series A 164*, 119–137.

Received December 2011; revised May 2012; accepted August 2012

Online Appendix to: CUDIA: Probabilistic Cross-level Imputation using Individual Auxiliary Information

Yubin Park, The University of Texas at Austin
Joydeep Ghosh, The University of Texas at Austin

A. HARD CLUSTERING DERIVATION

The log-likelihood of the CUDIA model is given by:

$$\log p(\mathbf{x},\mathbf{s},\mathbf{z}|\boldsymbol{\pi},\vec{\eta},\vec{\theta},\vec{\alpha}) = \log \prod_{p=1}^{P} p(\vec{s}_p|\vec{\pi}_p,\vec{\eta}) \prod_{i=1}^{N_p}\prod_{k=1}^{K} p(\vec{x}_i|\vec{\theta}_k)^{z_{ik}} p(\vec{z}_i|\vec{\pi}_p) \tag{27}$$

$$= \sum_{p=1}^{P} \log p(\vec{s}_p|\vec{\pi}_p,\vec{\eta}) + \sum_{p=1}^{P}\sum_{i=1}^{N_p}\sum_{k=1}^{K} z_{ik} \log(\pi_{pk} p(\vec{x}_i|\vec{\theta}_k)) \tag{28}$$

$$= \sum_{p} \log p(\vec{s}_p|\vec{\pi}_p,\vec{\eta}) + \sum_{p,i,k} z_{ik}(\log \pi_{pk} + \log p(\vec{x}_i|\vec{\theta}_k)), \tag{29}$$

where $\boldsymbol{\pi}$ is treated as a model parameter.

$p(\vec{s}_p|\vec{\pi}_p,\vec{\eta})$ is a Gaussian distribution according to the Central Limit Theorem, where its mean and variance are $\vec{\mu}_p$ and $\boldsymbol{\Sigma}_p^2$, respectively. Suppose $p(\vec{x}_i|\vec{\theta}_k)$ is a Gaussian distribution with a mean $\vec{\mu}_k$ and a diagonal covariance matricx $\epsilon\mathbf{I}$, and $\boldsymbol{\Sigma}_p^2$ has a form of $\frac{\epsilon}{\beta}\mathbf{I}$. Then, the log-likelihood becomes:

$$= -\sum_{p} \frac{\beta \parallel \vec{s}_p - \vec{\mu}_p \parallel^2}{\epsilon} - \sum_{p,i,k} z_{ik} \frac{\parallel \vec{x}_i - \vec{\mu}_k \parallel^2}{\epsilon} - \text{const.} \tag{30}$$

$$\propto -\sum_{p} \beta \parallel \vec{s}_p - \vec{\mu}_p \parallel^2 - \sum_{p,i,k} z_{ik} \parallel \vec{x}_i - \vec{\mu}_k \parallel^2 \tag{31}$$

Note that the maximum likelihood estimator of $\vec{\mu}_p$ is given as $\sum_k \frac{\sum_{n_p} z_{n_p k}}{N_p} \vec{\eta}_k$. By replacing $\vec{\mu}_p$ with its maximum likelihood estimator and changing the sign of the log-likelihood, we get an approximate deterministic clustering objective function as follows:

$$\min_{\mathbf{z},\vec{\mu},\vec{\eta}} \sum_{p} \{ \sum_{k,n_p} z_{n_p k} \parallel \vec{x}_{n_p} - \vec{\mu}_k \parallel^2 \} + \beta \parallel \vec{s}_p - \sum_{k} \frac{\sum_{n_p} z_{n_p k}}{N_p} \vec{\eta}_k \parallel^2 . \tag{32}$$

A local minimum of Equation (32) can be obtained by alternating the minimization steps between (i) $\mathbf{z}$ and (ii) $\vec{\mu}$, $\vec{\eta}$. This alternating minimization mechanism directly leads to Algorithm 4.

B. PROPERTIES OF THE CUDIA IMPUTATION

In this section, we show the basic properties of the CUDIA imputation, including the bias and variance of this imputation.

DOI 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

B.1. Unbiasedness of $\hat{x}_u$

Using (i) the law of iterated expectations and (ii) linearity of expectation,

$$E[\hat{x}_u] = E[\sum_{k=1}^{K} \eta_k E[z_k|\vec{x}_o, \vec{\pi}_p]] = \sum_{k=1}^{K} \eta_k E[E[z_k|\vec{x}_o, \vec{\pi}_p]] = \sum_{k=1}^{K} \eta_k E[z_k] = E[x_u]. \tag{33}$$

The expectation of the estimated $\hat{x}_u$ is the same as the expectation of the unobserved x_u. Thus, the imputation formula provides unbiased estimators for the x_u's. This property holds regardless of the distribution of x_u.

B.2. Variance of η

Recall the observed sample statistics (sample average) of a given partition p is:

$$s_p \sim \mathcal{N}(\mu_p, \sigma_p^2) \tag{34}$$

$$\mu_p = \sum_{k=1}^{K} \pi_{pk}\eta_k \tag{35}$$

$$\sigma_p^2 = \frac{\sum_{k=1}^{K} \pi_{pk}(\eta_k^2 + \tau_k^2) \ - \mu_p^2}{N_p} \propto \frac{1}{N_p}. \tag{36}$$

π_{pk} represents the mixing proportion of the kth component in the partition p. The linearity of expectation naturally leads to Equation (35). From the properties of mixture distributions, the variance of $\vec{x}_u$ in the partition p is given by:

$$Var[x_u|x_u \in \text{partition } p] = \sum_{k=1}^{K} \pi_{pk}(\eta_k^2 + \tau_k^2) \ - \mu_p^2. \tag{37}$$

Applying the Central Limit Theorem, we get Equation (36).

Suppose all the parameters of the CUDIA model are learned correctly, which means the log-likelihood reaches the global optimum. However, the sample means we used to learn the model are inherently noisy based on the Central Limit Theorem. This results in the noisy estimation of η's regardless of the learning methods used. As Equation (13) is the optimal solution in this setting, if all the parameters are learned correctly, then Equation (13) should also hold. Equation (13) gives another interesting interpretation, if we view $\mathbf{S}$ as "dependent variables" and $\mathbf{\Pi}$ as "independent variables" in a Linear regression formulation.

THEOREM B.1. *If all the parameters are learned correctly and $N_p = M, \ \forall p$, then*
(a) $\hat{\eta}$ is normally distributed.
(b) The means and variances are given by

$$E[\hat{\eta}_k] = \eta_k \tag{38}$$

$$Var[\hat{\eta}_k] \propto \frac{1}{M} \tag{39}$$

$$Cov[\hat{\eta}_i, \hat{\eta}_j] \propto \frac{1}{M}, \text{ where } i \neq j \text{ and } 0 < i, j < K. \tag{40}$$

PROOF. From Equation (13),

$$\begin{aligned} E[\mathbf{H}^*] &= E[(\mathbf{\Pi}^T\mathbf{\Pi})^{-1}\mathbf{\Pi}^T\mathbf{S}] = E[(\mathbf{\Pi}^T\mathbf{\Pi})^{-1}\mathbf{\Pi}^T(\mathbf{\Pi}\mathbf{H} + \epsilon)] \\ &= E[\mathbf{H}] + E[(\mathbf{\Pi}^T\mathbf{\Pi})^{-1}\mathbf{\Pi}^T\epsilon] = E[\mathbf{H}] + E[E[(\mathbf{\Pi}^T\mathbf{\Pi})^{-1}\mathbf{\Pi}^T\epsilon|\mathbf{\Pi}]] \\ &= E[\mathbf{H}] + E[(\mathbf{\Pi}^T\mathbf{\Pi})^{-1}\mathbf{\Pi}^T E[\epsilon|\mathbf{\Pi}]] = E[\mathbf{H}]. \end{aligned}$$

This proves Equation (38) in the result (b). Moreover, since $\mathbf{S}$ is normally distributed, a linear combination of $\mathbf{S}$ is also normal. Thus, $\mathbf{H}$, which is a linear combination of $\mathbf{S}$, is normal.

The estimator $\mathbf{H}^*$ can be written as:

$$\mathbf{H}^* = \mathbf{H} + (\mathbf{\Pi}^T\mathbf{\Pi})^{-1}\mathbf{\Pi}^T\epsilon. \tag{41}$$

Thus, the variance of $\mathbf{H}^*$ is the same as the variance of $(\mathbf{\Pi}^T\mathbf{\Pi})^{-1}\mathbf{\Pi}^T\epsilon$. Let $\mathbf{Q}_\pi = (\mathbf{\Pi}^T\mathbf{\Pi})^{-1}\mathbf{\Pi}^T$. Then,

$$Var[\mathbf{H}^*] = \mathbf{Q}_\pi\mathbf{\Sigma}_\epsilon^2\mathbf{Q}_\pi^T \text{ , where } \mathbf{\Sigma}_\epsilon^2 = \begin{pmatrix} \sigma_1^2 & 0 & ... & 0 \\ 0 & \sigma_2^2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \sigma_P^2 \end{pmatrix} \text{ and } \mathbf{Q}_\pi \equiv \begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1P} \\ q_{21} & q_{22} & \cdots & q_{2P} \\ \cdots & \cdots & \cdots & \cdots \\ q_{K1} & q_{K2} & \cdots & q_{KP} \end{pmatrix}. \tag{42}$$

Then,

$$Var[\eta_k] = \sum_{p=1}^{P} q_{kp}^2\sigma_p^2 = \sum_{p=1}^{P} q_{kp}^2 \frac{\sum_{k=1}^{K}\pi_{pk}(\eta_k^2+\tau_k^2) \ -\mu_p^2}{N_p} \tag{43}$$

$$= \frac{\sum_{p=1}^{P} q_{kp}^2\{\sum_{k=1}^{K}\pi_{pk}(\eta_k^2+\tau_k^2) \ -\mu_p^2\}}{M} \propto \frac{1}{M}. \tag{44}$$

Moreover,

$$Cov[\eta_i,\eta_j] = \sum_{p=1}^{P} q_{ip}q_{jp}\sigma_p^2 = \frac{\sum_{p=1}^{P} q_{ip}q_{jp}\{\sum_{k=1}^{K}\pi_{pk}(\eta_k^2+\tau_k^2) \ -\mu_p^2\}}{M} \propto \frac{1}{M}. \tag{45}$$

This proves Theorem (B.1). □

B.3. Variance of $\hat{x}_u$

The estimated $\hat{x}_u$ is a linear combination of $\hat{\eta}$'s. Theorem (B.1) naturally leads to the next theorem.

THEOREM B.2. *If* $\hat{x}_u = \sum_{k=1}^{K}\hat{\eta}_k E[z_k|\vec{x}_o,\vec{\pi}_p]$*, then*

$$Var[\hat{x}_u] \propto \frac{1}{M}. \tag{46}$$

PROOF. Let $a_k = E[z_k|\vec{x}_o,\vec{\pi}_p]$ to simplify the notation. Then,

$$Var[\hat{x}_u] = Var[\sum_{k=1}^{K} a_k\hat{\eta}_k] = \sum_{k=1}^{K} a_k^2 Var[\hat{\eta}_k] + \sum_{i\neq j} a_i a_j Cov[\hat{\eta}_i,\hat{\eta}_j] \propto \frac{1}{M}. \tag{47}$$

The last line of the equation comes from Theorem (B.1). This proves Theorem (B.2). □

LEMMA B.3. *The mean squared error,* $MSE(\hat{x}_u)$*, is inversely proportional to the size of the aggregation* M*.*

PROOF.

$$MSE(\hat{x}_u) = Var[\hat{x}_u] + (Bias(\hat{x}_u, x_u))^2 = Var[\hat{x}_u] \propto \frac{1}{M}. \tag{48}$$

□