CUDIA: Probabilistic Cross-level Imputation using Individual Side Information

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

Due to privacy or legal issues, aggregate data publication is a common practice in healthcare and medical research. However, to find out valuable individual level relationships from the aggregate data, many data mining algorithms suffer from the aggregation bias and the information loss, or require rather strict assumptions, which are usually unverifiable. Furthermore, even if individual level data are available, as many healthcare studies are performed with a pre-specified goal, a limited scope of variables constraints the range of the research focus. 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 aggregate datasets, which are possibly from different sources. By modeling the generative process of such datasets using a Bayesian directed graphical model, we propose a novel "cross-level" imputation technique. The imputation is based on the underlying data distribution and shown to be unbiased. This imputation can be further utilized in a subsequent predictive modeling, showing improved performances than just imputing the aggregate information as it is. 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. Probabilistic Cross-level Imputation using Individual Side Information. ACM Trans. Intell. Syst. Technol. 9, 4, Article 39 (November 2011), 23 pages.
DOI = 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

1. INTRODUCTION

Aggregate data publication, such as state or county summaries, is a common practice in healthcare and medical studies, as individual record publication might infringe privacy or legal issues. Specifically, revealing someone's disease or medical condition records might cause a severe traumatic situation when the information reflects a conflicting social perception such as STD or HIV. In this case, the aggregate data publication alleviates such risks, while providing a overall description about data. Due to this reason, many of the health-related features and indicators are publicly available at a highly aggregated level (for example see http://www.data.gov/health or http://www.cdc.gov/datastatistics/). However, the aggregation process damages a significant amount of individual information. If a data mining goal is not to observe or discover global or regional interactions, many data mining algorithms suffer from the

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

aggregation bias (also known as the ecological bias) or the lost information. Several statistical methods and assumptions have been suggested to make “cross-level inference” overcoming this issues [Achen and Shively 1995], [King 1997], but their validity and effectiveness are still controversial [Freedman 1999].

Aggregate variables in healthcare studies are not only the problem that hinders data mining research. Many of healthcare related surveys are designed with a specific predefined purpose, so that the variables in one dataset are limited in their scope. Designing another set of survey is inappropriate in many cases due to the cost and temporal dynamics. Combining multiple datasets from different sources can be a solution in this case, but in general, their aggregation levels differ in their sizes or populations. 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. As many health or healthcare indicators are available at different aggregated levels rather than providing an entry for each individual, a proper utilization of such data is crucial to facilitate healthcare data mining research without any extra cost.

Suppose two datasets from possibly different sources are available for research, where their aggregation levels are also different. By referring the dataset with a finer granularity as an individual level dataset, the other dataset becomes an aggregate level dataset. 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 “cross-level” imputed features reflect the underlying distribution of the data, thus a subsequent 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. Even further, for multiple datasets with multiple aggregation levels, our approach can be applied recursively to maximize the granularities of the datasets.

The rest of the paper is organized as follows: We begin by reviewing traditional statistical imputation techniques, ecological studies, and various inference mechanisms that will be used extensively in our approach in Section 2. In Section 3, we approach the problem by modeling the data generation process, which is essentially how the aggregate data are created. We start from a generic Bayesian clustering model, then step-by-step, we impose additional constraints and transform the simple model into our approach to exactly describe 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 MCEM algorithm is developed, which is computationally efficient than a generic MCEM technique. Moreover, a deterministic algorithm, which 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 is shown to be a“unbiased” estimator, and it statistical properties are analyzed in detail. Experimental verification of the proposed model is followed in Section 6 using a simulated dataset and the Behavioral Risk Factor Surveillance System (BRFSS) dataset. Finally, we discuss the limitation and the future work in Section 7.

2. RELATED WORK

In this section, we present 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, 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 technique is widely used, where the imputation is taken from approximate locations derived from associate data [Henry and Boscoe 2008]. Regression estimation [Tabachnick and Fidel 2001] is another widely used imputation technique in statistics. In the 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 regression results for the missing values are imputed. Nevertheless, the regression estimation assumes enough number of individual samples, which is not the case in our setting. On the other hand, 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 are limited in their applicabilities.

Ecological study. 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 [Robinson 1950], which is also known as the “ecological fallacy”. Most of the controversial ecological analyses were 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 doesn’t depend on the group specific characteristics. However, in general, the constancy assumption doesn’t 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 from “confounding” and “aggregation bias”. Traditionally, the aggregation bias has been tackled in two ways: a) assuming a quadratic model rather than a linear model, b) 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 from 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 as in the original ecological regression, and the interpretation of the results becomes harder. In the second method, unobserved individual features are bounded satisfying 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 rather used as a sanity check tool.

Despite its theoretical instability, ecological analyses will continue to be used being benefited from 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 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 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. Note that, individual and aggregate level are defined relatively to each other.

Inference algorithms in Bayesian Graphical Models. In Bayesian graphical models such as the model presented in this paper, inferential problems pose key challenges in most cases. EM algorithm is the most popular approach when latent variables are present in models. However, many sophisticated models such as LDA [Blei et al. 2003] 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 have been proposed. Although their computational complexities and assumptions are slightly different, their performances are marginally the same [Asuncion et al. 2009]. In this paper, we demonstrate an approximated Gibbs sampling approach, which is specialized for our setting. Then we further introduce its deterministic algorithm, 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 by $\vec{x}_o$, where "individual" refers to entities at the highest 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 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 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 to cover such situations, and in this paper we consider a common partitioning to keep the notation and exposition simple.

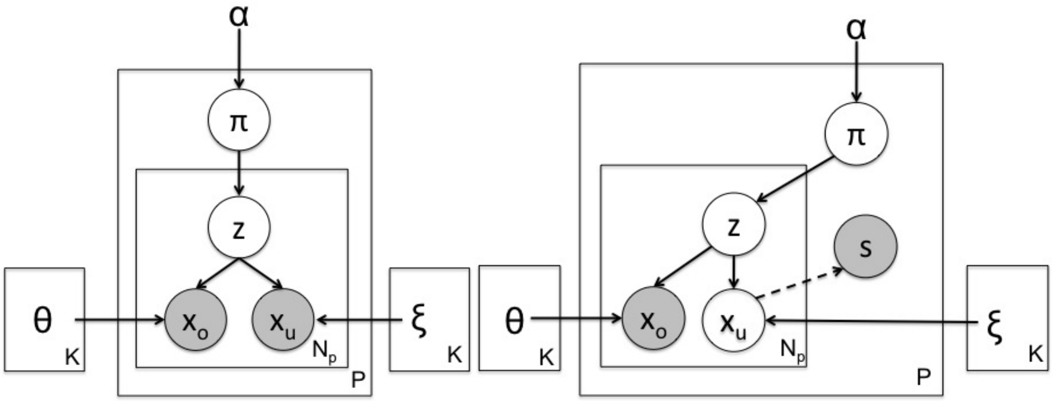

Fig. 1. (a) Clustering models when complete data is available (left) and (b) when only aggregates $\vec{s}$ are observed instead of $\vec{x}_u$ (right).

Suppose we want to find K clusters in the complete data, denoted by $\{\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_K\}$. Note that the clusters are based on the full features $(\vec{x}_o, \vec{x}_u)$, while the partitions are arbitrary, 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 $\vec{\pi}_p = (p(\mathcal{C}_1|\mathcal{D}^p_x), p(\mathcal{C}_2|\mathcal{D}^p_x), ..., p(\mathcal{C}_K|\mathcal{D}^p_x))^T = (\pi_{p1}, \pi_{p2}, ..., \pi_{pK})^T$, 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 any distinct 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 $\vec{\alpha}$. 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 $\mathbf{T}^2_k$ 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:

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

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

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

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

Essentially, $\vec{\eta}_k$ and $\mathbf{T}^2_k$ 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 effi-

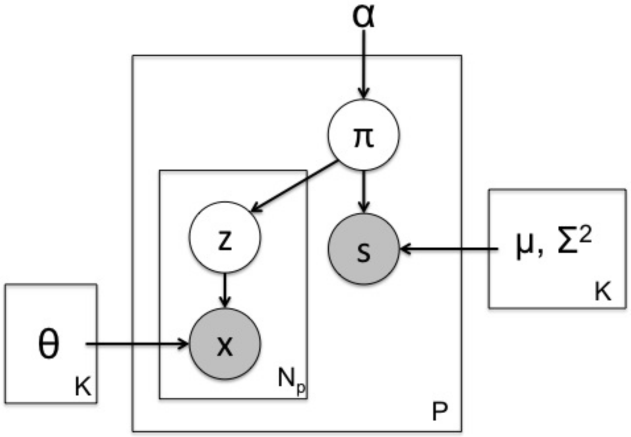

Fig. 2. Graphical Model of CUDIA.

cient **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 $\vec{x}_o$ 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:

$$
\begin{aligned}
& p(\mathbf{x}, \mathbf{s}|\vec{\eta}, \vec{\theta}, \vec{\alpha}) && (5)\\
& = \sum_{\mathbf{z}} \int_{\boldsymbol{\pi}} \prod_{p=1}^{P} p(\vec{s}_p|\vec{\pi}_p, \vec{\eta}) p(\vec{\pi}_p|\vec{\alpha}) \prod_{i=1}^{N_p} \prod_{k=1}^{K} p(\vec{x}_i|\vec{\theta}_k)^{z_{ik}} p(\vec{z}_i|\vec{\pi}_p) d\boldsymbol{\pi} && (6)\\
& = \int_{\boldsymbol{\pi}} \prod_{p=1}^{P} p(\vec{s}_p|\vec{\pi}_p, \vec{\eta}) p(\vec{\pi}_p|\vec{\alpha}) \prod_{i=1}^{N_p} \prod_{k=1}^{K} \sum_{\mathbf{z}} p(\vec{x}_i|\vec{\theta}_k)^{z_{ik}} p(\vec{z}_i|\vec{\pi}_p) d\boldsymbol{\pi}. && (7)
\end{aligned}
$$

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

$$
p(\boldsymbol{\pi}, \mathbf{z}|\vec{\eta}, \vec{\theta}, \vec{\alpha}, \mathbf{x}, \mathbf{s}) = \frac{p(\mathbf{x}, \mathbf{s}, \boldsymbol{\pi}, \mathbf{z}|\vec{\eta}, \vec{\theta}, \vec{\alpha})}{p(\mathbf{x}, \mathbf{s}|\vec{\eta}, \vec{\theta}, \vec{\alpha})}. \qquad (8)
$$

The key inferential problem is how to calculate this posterior distribution. However, a generic EM algorithm [Dempster et al. 1977] cannot be applied, since the normalization constant of its posterior distribution in Equation (8) 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 $p(\mathbf{x}, \mathrm{s}, \boldsymbol{\pi}, \mathbf{z}|\vec{\eta}, \vec{\theta}, \vec{\alpha})$. 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:

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

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

$$\vec{\pi}_p^{(j+1)} \sim p(\vec{\pi}|\vec{z}_1^{(j)}, \vec{z}_2^{(j)}, ..., \vec{z}_{N_p}^{(j)}, \vec{s}_p, \vec{\eta}, \vec{\alpha}) \tag{9}$$

$$\vec{z}_i^{(j+1)} \sim p(\vec{z}|\vec{\pi}_p^{(j+1)}, \vec{x}_i, \vec{\theta}). \tag{10}$$

However, sampling $\vec{\pi}$ is problematic as Eq. (9) is not a trivial distribution. Instead of sampling directly from Eq. (9), Metropolis-Hastings (MH) algorithm can be used with a proposal density $Dirichlet(\vec{\alpha})$. This algorithm is described in Algorithm 1.

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**
        $\vec{\pi}_p^{(index+1)} \leftarrow \vec{\pi}_p^{(index)}$;
    **end**
**until** $index < I_{Max}$;
```

The sampling from a Dirchlet distribution might be computationally heavy in some programming languages. As an alternative, the prior distribution of $\vec{\pi}$ can be replaced by a Logistic Normal distribution or a Uniform distribution by modifying the CUDIA model, so that we can adopt a different proposal density function according to the modified model. In our empirical evaluation over the different prior distributions, it 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 removed, when a large enough data size of N_p and a small dimension of $\vec{x}$ are provided.

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 is described in Algorithm 2:

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 $E[z_k|\pi_p^{(index)}, \mathbf{x}] \propto p(\mathbf{x}|\vec{\theta}_k)\pi_{pk}^{(index)}$;
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 $\vec{\pi}_p \propto \sum E[\vec{z}|\mathbf{x}]$;
```

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 the MH sampling. In this paper, we used a burning period of 10 samples, and $N_{Gibbs} \approx 50$ to 100 [Agarwal and Chen 2009]. 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 [Wei and Tanner 1990] :

$$\mathbf{S}_i = [s_{1i}, s_{2i}, ..., s_{Pi}]^T \tag{11}$$

$$\hat{\mathbf{\Pi}} = [\hat{\vec{\pi}}_1, \hat{\vec{\pi}}_2, ..., \hat{\vec{\pi}}_P]^T, \text{ where } \hat{\vec{\pi}}_p = \frac{\sum_{i=1}^{N_{Gibbs}} \vec{\pi}_p^{(i)}}{N_{Gibbs}} \tag{12}$$

$$\mathbf{W} = diag(N_1, N_2, ..., N_P) \tag{13}$$

$$\mathbf{H} = [\vec{\eta}_1, \vec{\eta}_2, ..., \vec{\eta}_K]^T \tag{14}$$

To help understanding, their expressive forms are given by:

$$\mathbf{S}_i = \begin{pmatrix} s_{1i} \\ s_{2i} \\ \dots \\ s_{Pi} \end{pmatrix}, \; \mathbf{H}_{\cdot i} = \begin{pmatrix} \eta_{1i} \\ \eta_{2i} \\ \dots \\ \eta_{Ki} \end{pmatrix}, \; \hat{\mathbf{\Pi}} = \begin{pmatrix} \hat{\pi}_{11} & \hat{\pi}_{12} & \dots & \hat{\pi}_{1K} \\ \hat{\pi}_{21} & \hat{\pi}_{22} & \dots & \hat{\pi}_{2K} \\ \dots & \dots & \dots & \dots \\ \hat{\pi}_{P1} & \hat{\pi}_{P2} & \dots & \hat{\pi}_{PK} \end{pmatrix}. \tag{15}$$

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:

$$\mathbf{S}_i \approx \hat{\mathbf{\Pi}} \cdot \mathbf{H}_{\cdot i} \tag{16}$$

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

$$\mathbf{H}^*_{\cdot i} = (\hat{\mathbf{\Pi}}^T \mathbf{W} \hat{\mathbf{\Pi}})^{-1} \hat{\mathbf{\Pi}}^T \mathbf{W} \mathbf{S}_i. \tag{17}$$

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:

$$\mathbf{H}^*_{\cdot i} = (\hat{\mathbf{\Pi}}^T \mathbf{W} \hat{\mathbf{\Pi}} + \lambda \mathbf{I})^{-1} \hat{\mathbf{\Pi}}^T \mathbf{W} \mathbf{S}_i. \tag{18}$$

Furthermore, the regularizer term, λ, can be utilized when $P < K$, which makes CUDIA under-determined. But we leave this to the future work. The entire inference algorithm is described in Algorithm 3.

ALGORITHM 3: Gibbs CUDIA EM algorithm

```
Input: $\mathrm{x}, \mathrm{s}$
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 Converge;
```

4.3. Deterministic Hard Clustering

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

$$\min_{\mathbf{z},\vec{\mu},\vec{\eta}} \sum_p \{ \sum_{k,n_p} z_{n_p k} \parallel \vec{x}_{n_p} - \vec{\mu}_k \parallel^2 \} + \beta \parallel \vec{s}_p - \sum_k \frac{\sum_{n_p} z_{n_p k}}{N_p} \vec{\eta}_k \parallel^2 \tag{19}$$

$$= \min_{\mathbf{z},\vec{\mu},\vec{\eta}} \sum_{p,k,n_p} z_{n_p k} \parallel \vec{x}_{n_p} - \vec{\mu}_k \parallel^2 + \frac{\beta}{KN_p} \parallel \vec{s}_p - \sum_k \hat{\pi}_{pk} \vec{\eta}_k \parallel^2 \tag{20}$$

where $\hat{\pi}_{pk} = \frac{\sum_{n_p} z_{n_p k}}{N_p}$ 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 z and $(\vec{\mu}, \vec{\eta})$ as in Algorithm 4. One iteration of this algorithm costs $\Theta(KN)$. For a fixed number of iterations I, the overall complexity is therefore $\Theta(KNI)$, which is

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
        $z_{n_p k} \leftarrow 0$;
    end
    (Update Step)
    $\vec{\mu}_k \leftarrow \sum_n (z_{nk}\vec{x}_n)/N_k$, $\vec{\pi}_p \leftarrow \sum_{n_p} \vec{z}_{n_p}/N_p$;
    $\mathbf{H}_{\cdot i} \leftarrow (\hat{\mathbf{\Pi}}^T \mathbf{W} \hat{\mathbf{\Pi}} + \lambda_M \mathbf{I})^{-1} \hat{\mathbf{\Pi}}^T \mathbf{W} \mathbf{S}_i$;
until Converge;
```

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 [Banerjee et al. 2005]. Given two vectors $\vec{x}$ and $\vec{\mu}$, the Bregman divergence is defined as:

$$d_\phi(\vec{x}, \vec{\mu}) = \phi(\vec{x}) - \phi(\vec{\mu}) - \langle \vec{x} - \vec{\mu}, \nabla\phi(\vec{\mu}) \rangle \tag{21}$$

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 [Banerjee et al. 2005]. Table I 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:

— **Assignment Step**

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

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

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.

Table I. 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	$\sum_{j=1}^{d}\mu_j\log\frac{\mu_j}{M}$	$\sum_{j=1}^{d}x_j\log\frac{x_j}{\mu_j}$

Table II. An example of the CUDIA Imputation. The imputed values are personalized according to the individual observations.

Indiv. level	**Aggr. level**	**CUDIA imputation**
$\vec{x}_1=(6,2)^T$	$\vec{s}_1=(0.09826,0.01024)^T$	$\hat{\vec{x}}_{u,1}\vert\vec{x}_1=(-0.0082,0.0280)^T$
$\vec{x}_2=(3,3)^T$	$\vec{s}_1=(0.09826,0.01024)^T$	$\hat{\vec{x}}_{u,2}\vert\vec{x}_2=(0.1491,0.0779)^T$
$\vec{x}_3=(4,3)^T$	$\vec{s}_1=(0.09826,0.01024)^T$	$\hat{\vec{x}}_{u,3}\vert\vec{x}_3=(0.1420,0.0768)^T$
...	...	
$\vec{x}_{N-1}=(6,2)^T$	$\vec{s}_P=(-0.02818,-0.03053)^T$	$\hat{\vec{x}}_{u,N-1}\vert\vec{x}_{N-1}=(-0.0082,0.0280)^T$
$\vec{x}_N=(2,1)^T$	$\vec{s}_P=(-0.02818,-0.03053)^T$	$\hat{\vec{x}}_{u,N}\vert\vec{x}_N=(-0.1186,-0.0725)^T$

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:

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

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

$$=\sum_k p(\vec{x}_u|z_k)p(z_k|\vec{x}_o,\vec{\pi}_p). \tag{24}$$

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:

$$\hat{\vec{x}}_u\leftarrow\sum_{k=1}^{K}\vec{\eta}_k E[z_k|\vec{x}_o,\vec{\pi}_p] \tag{25}$$

where the covariance of $\vec{x}_u$ is assumed to be $\delta^2\mathbf{I}$. Table II shows an example of this imputation method. Based on the observed individual values, the model finds an appropriate cluster assignment, then performs a personalized imputation. The numbers in Table II is based on the BRFSS dataset, which will be detailed in Section 6.

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:

$$E[z_k|\vec{x}_o,\vec{\pi}_p]\propto\frac{exp(-d_\phi(\vec{x}_o,\vec{\mu}_k))}{\sum_l exp(-d_\phi(\vec{x}_o,\vec{\mu}_l))}\pi_{pk}. \tag{26}$$

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.

5.1. Unbiasedness of $\hat{x}_u$

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

$$E[\hat{x}_u] = E[\sum_{k=1}^{K} \eta_k E[z_k|\vec{x}_o, \vec{\pi}_p]] = \sum_{k=1}^{K} \eta_k E[E[z_k|\vec{x}_o, \vec{\pi}_p]] \tag{27}$$

$$= \sum_{k=1}^{K} \eta_k E[z_k] = E[x_u]. \tag{28}$$

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, as we have not used any of its property.

5.2. Variance of η

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

$$s_p \sim \mathcal{N}(\mu_p, \sigma_p^2) \tag{29}$$

$$\mu_p = \sum_{k=1}^{K} \pi_{pk}\eta_k \tag{30}$$

$$\sigma_p^2 = \frac{\sum_{k=1}^{K} \pi_{pk}(\eta_k^2 + \tau_k^2)\ - \mu_p^2}{N_p} \propto \frac{1}{N_p}. \tag{31}$$

π_{pk} represents the mixing proportion of the kth component in the partition p. The linearity of expectation naturally leads to Equation (30). From the properties of mixture distributions, the variance of $\vec{x}_u$ in the partition p is given by:

$$Var[x_u|x_u \in \text{partition } p] = \sum_{k=1}^{K} \pi_{pk}(\eta_k^2 + \tau_k^2)\ - \mu_p^2. \tag{32}$$

Applying the Central Limit Theorem, we get Equation (31).

Suppose all the parameters of the CUDIA model is learned correctly, which means the log-likelihood reached the global optimum. However, the sample means we used to learn the model is inherently noisy based on the Central Limit Theorem. This results in the noisy estimation of η's regardless of learning methods. As Equation (17) is the optimal solution in this setting, if all the parameters are learned correctly, then Equation (17) should also hold. Equation (17) gives another interesting interpretation, if we view **S** as "dependent variables" and $\mathbf{\Pi}$ as "independent variables" in a Linear regression.

THEOREM 5.1. *If all the parameters are learned correctly and $N_p = M$, $\forall p$, then*
(a) $\hat{\eta}$ is normally distributed.
(b) The mean and variance are given by

$$E[\hat{\eta}_k] = \eta_k \tag{33}$$

$$Var[\hat{\eta}_k] \propto \frac{1}{M} \tag{34}$$

$$Cov[\hat{\eta}_i, \hat{\eta}_j] \propto \frac{1}{M}, \textit{ where } i \neq j \textit{ and } 0 < i, j < K. \tag{35}$$

PROOF. From Eqaution (17),

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

This proves Equation (33) in the result (b). Moreover, since **S** is normally distributed, a linear combination of **S** is also normal. Thus, **H**, which is a linear combination of **S**, is normal.

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

$$\mathbf{H}^* = \mathbf{H} + (\mathbf{\Pi}^T\mathbf{\Pi})^{-1}\mathbf{\Pi}^T\epsilon. \tag{36}$$

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,

$$Var[\mathbf{H}^*] = \mathbf{Q}_\pi\mathbf{\Sigma}_\epsilon^2\mathbf{Q}_\pi^T \text{ , where } \mathbf{\Sigma}_\epsilon^2 = \begin{pmatrix}\sigma_1^2 & 0 & ... & 0\\ 0 & \sigma_2^2 & ... & 0\\ ... & ... & ... & ...\\ 0 & 0 & ... & \sigma_P^2\end{pmatrix} \text{ and } \mathbf{Q}_\pi \equiv \begin{pmatrix}q_{11} & q_{12} & ... & q_{1P}\\ q_{21} & q_{22} & ... & q_{2P}\\ ... & ... & ... & ...\\ q_{K1} & q_{K2} & ... & q_{KP}\end{pmatrix}. \tag{37}$$

Then,

$$Var[\eta_k] = \sum_{p=1}^{P} q_{kp}^2\sigma_p^2 = \sum_{p=1}^{P} q_{kp}^2 \frac{\sum_{k=1}^{K}\pi_{pk}(\eta_k^2+\tau_k^2)\ -\mu_p^2}{N_p} \tag{38}$$

$$= \frac{\sum_{p=1}^{P} q_{kp}^2\{\sum_{k=1}^{K}\pi_{pk}(\eta_k^2+\tau_k^2)\ -\mu_p^2\}}{M} \propto \frac{1}{M}. \tag{39}$$

Moreover,

$$Cov[\eta_i,\eta_j] = \sum_{p=1}^{P} q_{ip}q_{jp}\sigma_p^2 = \frac{\sum_{p=1}^{P} q_{ip}q_{jp}\{\sum_{k=1}^{K}\pi_{pk}(\eta_k^2+\tau_k^2)\ -\mu_p^2\}}{M} \propto \frac{1}{M}. \tag{40}$$

This proves Theorem (5.1). □

5.3. Variance of $\hat{x}_u$

The estimated $\hat{x}_u$ is a linear combination of $\hat{\eta}$'s. Theorem (5.1) naturally leads to the next theorem.

THEOREM 5.2. *If $\hat{x}_u = \sum_{k=1}^{K}\hat{\eta}_k E[z_k|\vec{x}_o,\vec{\pi}_p]$, then*

$$Var[\hat{x}_u] \propto \frac{1}{M}. \tag{41}$$

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

$$Var[\hat{x}_u] = Var[\sum_{k=1}^{K} a_k\hat{\eta}_k] = \sum_{k=1}^{K} a_k^2 Var[\hat{\eta}_k] + \sum_{i\neq j} a_i a_j Cov[\hat{\eta}_i,\hat{\eta}_j] \propto \frac{1}{M}. \tag{42}$$

The last line of the equation comes from Theorem (5.1). This proves Theorem (5.2). □

LEMMA 5.3. *The mean squared error, $MSE(\hat{x}_u)$, is inversely proportional to the size of the aggregation M.*

PROOF.

$$MSE(\hat{x}_u) = Var[\hat{x}_u] + (Bias(\hat{x}_u, x_u))^2 = Var[\hat{x}_u] \propto \frac{1}{M}. \tag{43}$$

□

Surprisingly, the bigger the size of the aggregation M, the more accurate the estimate of $\hat{x}_u$. However, if we assume a finite number of samples, say N, then there is a critical size of M, beyond that, the estimate doesn't get better.

LEMMA 5.4. *If $M > N/K$, then the CUDIA model becomes under-determined.*

PROOF. The CUDIA model assumes $P > K$, where $P = N/M$. □

6. EXPERIMENTAL RESULTS

In this section, we provide two kinds of experimental results. (a) First, imputation quality of the CUDIA model is assessed using a simulated mixture of Gaussians data. (b) Then, its applicability to predictive modeling[1] is discussed using the data from the Behavioral Risk Factor Surveillance System (BRFSS).

6.1. Imputation Quality

We demonstrate the CUDIA imputation using a simulated datasets. Two 2-D Gaussians are used as mixture components, which are centered at $(-0.5, -0.5)$ and $(0.5, 0.5)$ respectively, both having the same covariance matrix $0.07\mathbf{I}$. With these two Gaussians, we follow the CUDIA generative process. First, the Dirichlet prior parameters are given as $\vec{\alpha} = (1, 1)$ to make each partition to have almost equal amount of clusters. For each partition, a mixing coefficient vector $\vec{\pi}$ is drawn from the Dirichlet prior, where $M = N_p,\ \forall p$. We generated total 960 random samples (N) from the CUDIA generative model, and the aggregation size M is 160. The x-axis values are regarded as individual level data points and the y-axis values are aggregated. After generating the datasets, the masked y-axis values (x_u in CUDIA) are imputed using the CUDIA imputation algorithm.

Figure 3(a) shows the complete dataset, where the y-axis values are not aggregated yet. When given this kind of multi-level datasets, a typical imputation method is to make everyone in the same partition share the same feature values, which is essentially the average statistics. Figure 3(b) describes this naïve imputation, and the CUDIA imputation is shown in Figure 3(c). We can observe that the CUDIA model captures the underlying distribution of the generative model without accessing the masked individual data points. Figure 4 shows the Mean Squared Error (MSE) between the true and the imputed data points. The CUDIA imputation achieves the lower MSE simultaneously having the lower variance compared to the naïve imputation.

6.2. BRFSS dataset

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

Dataset description. We demonstrate the proposed method using the BRFSS 2009 dataset. BRFSS (Behavioral Risk Factor Surveillance System) [2] 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

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

[2] http://www.cdc.gov/brfss/

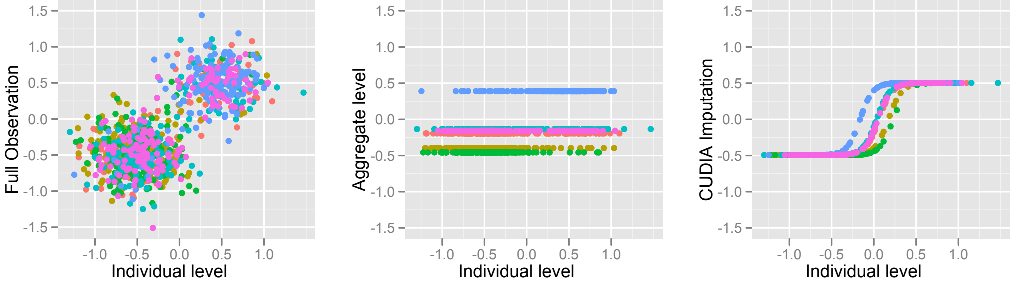

Fig. 3. (a) Simulated dataset with the individual level data (x-axis) and the true individual values (y-axis) for the aggregate data (left). (b) Direct imputation using the aggregate level data (center). (c) CUDIA imputation (right). A different color represents each partition.

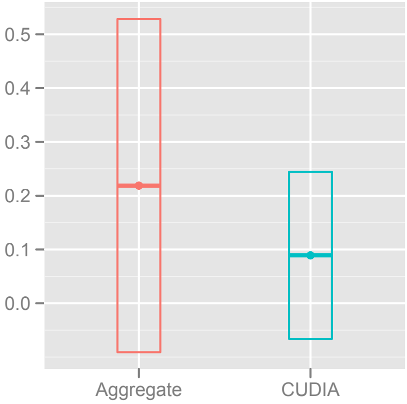

Fig. 4. Imputation accuracy (Mean Squared Error) on the simulated dataset.

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 [3]. The 2009 dataset contains more than 400,000 records and 405 variables and the diabetic (positive class) ratio is 12%. Empty columns and little informative columns are dropped and 22 predictors are finally chosen, including Hypertension, Body Mass Index (BMI), age, education, income, etc. The aggregation level we chose in this paper is the US census division as shown in Figure 5. For each division, the important feature distributions are described in Figure 6. Although the distributions are slightly different from each division, we can observe that they do not reflect the true clusters of the features.

6.3. Aggregated Target

In many cases, revealing personal disease records can be problematic, or even cause traumatic situations e.g. HIV. Rather than the individual disease records, suppose the data is provided at aggregate level such as state-level or county-level summaries. In this paper, we focus on diabetes records that are aggregated at the US division level

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

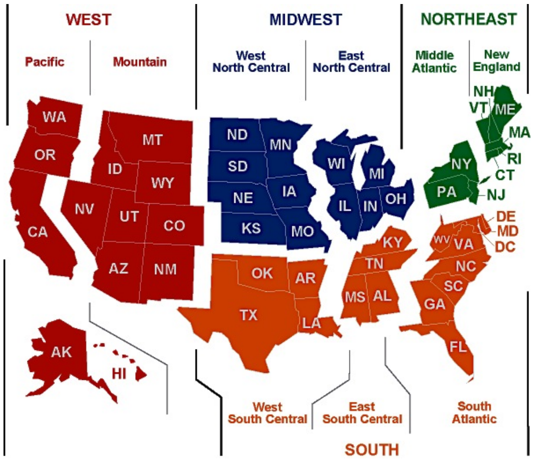

Fig. 5. Census Regions and Divisions of the United States. This picture is adopted from http://www.eia.gov /emeu/reps/maps/us_census.html.

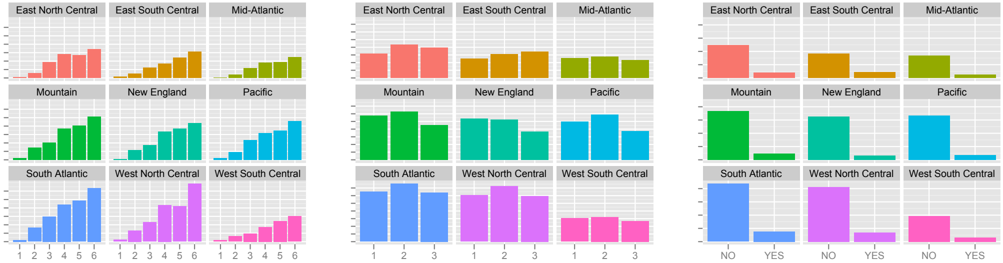

Fig. 6. BRFSS dataset description for each division. (a) Age (left), (b) BMI (center) and (c) Diabetes (right).

as in Figure 6(c). Then, (i) Age, (ii) BMI, (iii) Education level and (iv) Income level are provided as the individual 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 four features represent numeric values, their values are grouped ranging from 3 to 6 levels. To prevent the singular variance problem in the EM algorithm, their values are perturbed with a Uniform noise before the learning process. After all the parameters in the CUDIA model are converged, the estimate for the masked variable, diabetes, is individually imputed. Since the masked variable is a binary feature, the imputation quality can be measured as follows:

$$\text{Average Log-likelihood} = \frac{1}{|T|}\sum_{i=1}^{|T|} \log \hat{t}_i^{t_i}(1-\hat{t}_i)^{1-t_i}, \tag{44}$$

where t is the original target value, $\hat{t}$ is the CUDIA imputed value and $|T|$ represents the total number of the data points.

Figure 7 shows the results with varying λ and K in the model. In Figure 7(b), the lift is calculated based on the performance of the naïve imputation method, which imputes all the same values in the same division. This base performance recorded -0.3532 average log-likelihood. From the figure, we can observe that the CUDIA imputation outperforms the naïve imputation. Although the performance increases as K increases,

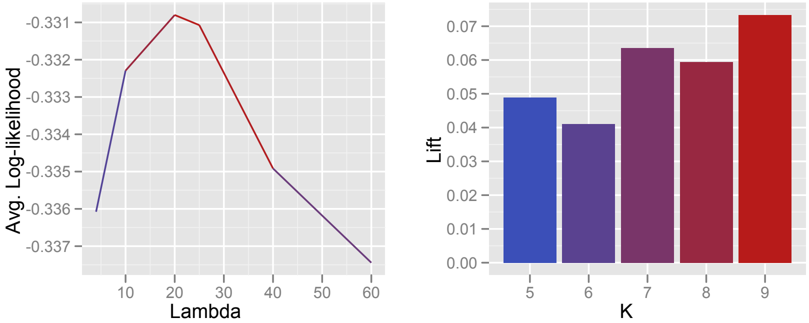

Fig. 7. Results from the aggregated diabetes dataset. (a) λ vs. Average Log-likelihood when $K = 7$ (left), (b) K vs. Lift when $\lambda = 20.0$ (right).

Table III. 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

we cannot test the cases when $K > 9$ since these cases make Equation (17) to be underdetermined.

The CUDIA model provides another valuable information about the data, the underlying distribution. Table III shows the learned parameters from the model. Noticeably, Cluster 7 exhibits a high risk diabetes. Their profiles can be described as “higher age”, “obese” and “middle-class”, where this relationship between obesity and diabetes coincides with the medical research [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 of the problem, 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 predictive modeling algorithms such as decision trees or regressions. Figure 8 describes the main idea of this approach.

We use the BRFSS dataset with the individual features (age and BMI), the masked features (Hypertension and high-cholesterol) and the target (diabetes). The masked features are aggregated using the US census division mapping, and the target is only used in 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 from this setting, one can use other regression techniques such as Lasso and Ridge regression [Park and Ghosh 2011], [Park and Ghosh 2012]. In this paper, we demonstrate this predictive modeling framework using a Logistic regression family and decision trees.

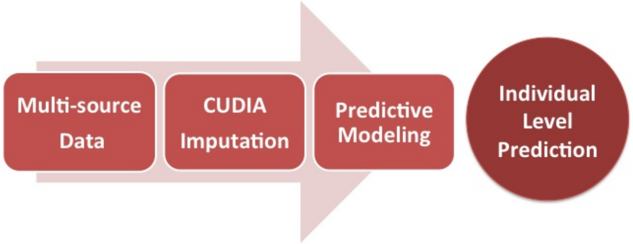

Fig. 8. Predictive modeling using the CUDIA framework.

Table IV. 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

Table IV shows the estimated parameters from the dataset. The people belong 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. However, their BMI's are very low, and this explains the result. For the rest of the predictive modeling tasks, we used $K = 9$ and $\lambda = 10.0$, where λ is in Equation (18).

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

$$Diabetes \sim logit(\beta_{Hyper}(Hypertension) + \beta_{Chol}(Cholesterol) + \beta_{Const}). \qquad (45)$$

Figure 9 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 9(a), we can observe that the coefficients from the CUDIA dataset follows the relationship of the complete dataset. 5-fold cross validation is performed and their average log-likelihood values are recorded. Figure 9(b) shows the CUDIA dataset outperforms the baseline dataset, while slightly worse than the complete dataset.

Logistic regression with L1 constraints. The rest of the experiments use the combination of the individual and the aggregate values. Thus, 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", "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

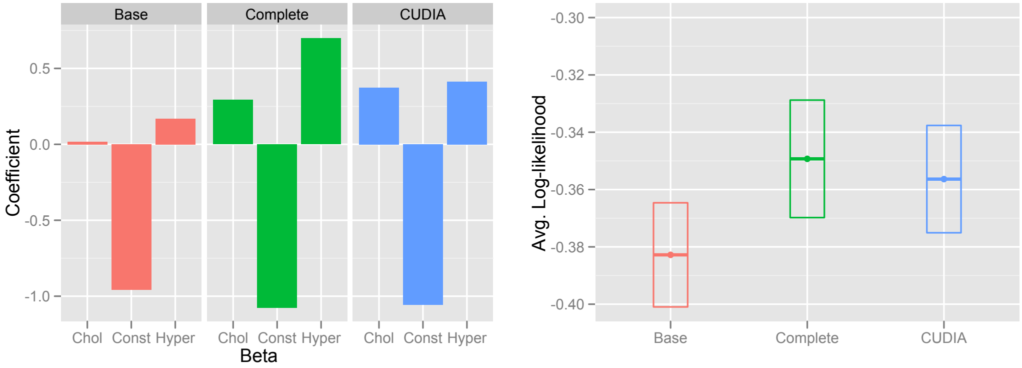

Fig. 9. Results from the Logistic regression only with the masked variables. (a) Coefficients (β) (left), (b) Average Log-likelihood on the test sets (right).

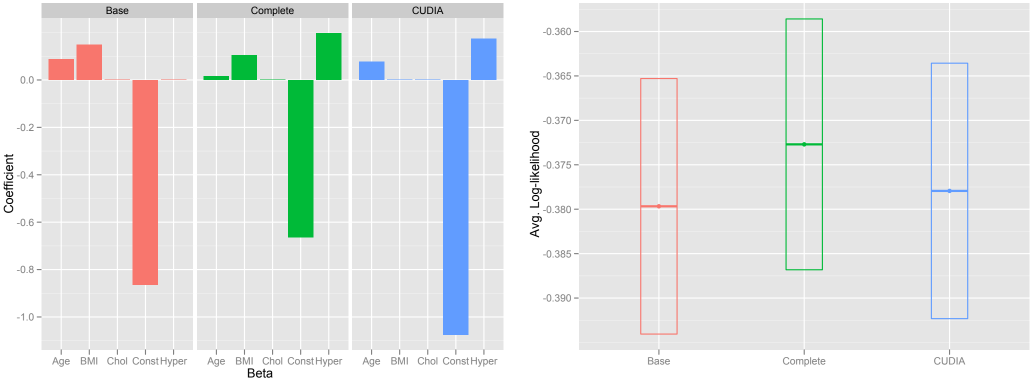

Fig. 10. Results from the Logistic regression with L1 constraints. (a) Coefficients (β) (left), (b) Average Log-likelihood on the test sets (right).

shrinkage methods, also known as regularizers such as $L1$ or $L2$ [Hastie et al. 2009]. In this paper, we demonstrate two regularizers, $L1$ and $L2$.

The $L1$ regularizer is known to generate a sparser solution compared to a normal regression [Scholkopt et al. 2007], which can be regarded as an automatic feature selection technique. Figure 10 shows the results from the $L1$ Logistic regression. From Figure 10(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 $L1$ regularizer. Furthermore, the average log-likelihood values from 5-cv show that the CUDIA imputation is actually effective in this predictive task than the baseline imputation.

Logistic regression with L2 constraints. $L2$ constraint is another popular choice among many regularizers. Using the same dataset by differing the regularizer, we have the results, which is shown in Figure 11. Unlike the $L1$ case, all the coefficients have non-zero values in Figure 11(a). Plus, the coefficients from the complete and the CUDIA datasets have very similar weights to each other. Again, from Figure 11(b),

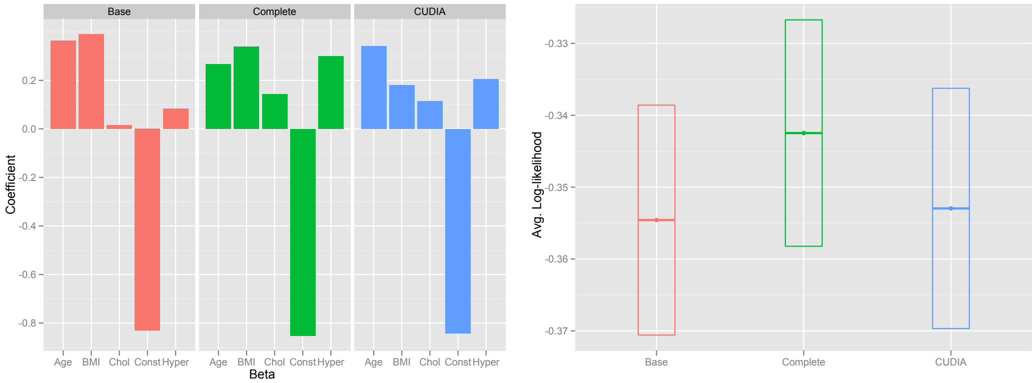

Fig. 11. Results from the Logistic regression with L2 constraints. (a) Coefficients (β) (left), (b) Average Log-likelihood on the test sets (right).

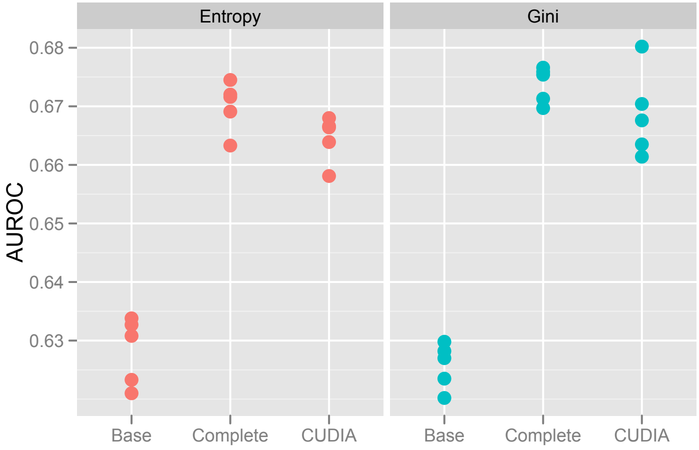

Fig. 12. Results from the Decision trees. (a) Entropy criterion (left), (b) Gini criterion (right).

the CUDIA imputation is more effective in the $L2$ Logistic regression case than the baseline dataset.

Decision Tree. Decision trees are recursive rule based classifiers, and we demonstrate using two kinds of decision trees: (i) Gini criterion [Breiman 1984] and (ii) Entropy criterion [Quinlan 1993]. We used the decision tree package from KNIME[4], and the Minimum Description Length (MDL) principle is used in the pruning. The combined dataset from the aggregate and the individual values is used as the same in the previous experiments.

Figure 12 shows the results from the decision trees. The performances are 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. Furthermore, one experiment from the Gini criterion decision tree

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

actually outperformed the rest of the complete dataset performances. Originally, the CUDIA model is designed based from the underlying distribution, then the individual values are imputed baed on the conditional distributions. As the recursive decision tree algorithms more focus on the conditional distributions between the target and the features, not the individual value, the CUDIA model shows its strength especially in decision tree algorithms.

7. DISCUSSIONS

CUDIA can be viewed as an approximate graphical model of the original generative process (Figure 1(b)). Nevertheless, in many practical settings, this approximation becomes a part of data publication properties.

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 Figure 1(b) model is no longer valid. The difference between the sub-sampled average and the true sample average can be modeled using a Normal distribution, 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. Laplace or Gaussian noise are known to protect the (ϵ, δ)-differential privacy with certain assumptions [Dwork et al. 2006], [Dwork et al. 2006]. Adding a Gaussian noise exactly fits the assumption of the CUDIA model, so that the CUDIA model becomes no longer an approximation in this case. Finally, in many real datasets, sizes of aggregation usually range from 1,000 to 100,000 or even more. To make the exact inference on the Figure 1(b) model, a simultaneous optimization across the aggregation is needed, which is intractable considering the sizes of the hidden variables.

The derivation of the CUDIA model is based on the CLT approximation and subsequent removal of the unobserved random variables, $\vec{x}_u$'s. The CLT assumption in CUDIA is valid in many cases at least indirectly as well as necessary due to the intractability of the original model. The CUDIA model not only helps the inference to be tractable, but also captures many practical settings in real datasets.

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 modelings, leading to improved performances. 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. 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 [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 specified by the data provider (P). This is why we had to stop at $K = 9$ for several of the results even though the performances were improving with with increasing K.

ACKNOWLEDGMENTS

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. 2009. Regression-based latent factor models. In *KDD '09*. 19–28.

ASUNCION, A., WELLING, M., SMYTH, P., AND TEH, Y. 2009. On smoothing and inference for topic models. In *UAI 2009*.

BANERJEE, A., MERUGU, S., DHILLON, I., AND GHOSH, J. 2005. Clustering with Bregman divergences. *Jl. Machine Learning Research (JMLR) 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.

DEMPSTER, A. P., LAIRD, N. M., AND RUBIN, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Society. Series B (Methodological) 39,* 1, 1–38.

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

DWORK, C., KENTHAPADI, K., MCSHERRY, F., MIRONOV, I., AND NAOR, M. 2006. Our data, ourselves: Privacy via distributed noise generation. In *EUROCRYPT*.

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

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*.

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. *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 *KDD 2011 Workshop on Medicine and Healthcare*.

PARK, Y. AND GHOSH, J. 2012. A probabilistic imputation framework for predictive analysis using variably aggregated, multi-source healthcare data. In *ACM SIG IHI 2012*.

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.

SCHOLKOPT, B., PLATT, J. C., AND HOFFMANN, T. 2007. Sparse multinomial logistic regression via bayesian l1 regularization. In *Neural Information Processing Systems (NIPS)*.

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.

WEI, G. C. G. AND TANNER, M. A. 1990. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association 85,* 411, 699–704.

Received November 2011; revised .; accepted .