LUDIA: An Aggregate-Constrained Low-Rank Reconstruction Algorithm to Leverage Publicly Released Health Data

Yubin Park
The University of Texas at Austin
yubin.park@utexas.edu

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

ABSTRACT

In the past few years, the government and other agencies have publicly released a prodigious amount of data that can be potentially mined to benefit the society at large. However, data such as health records are typically only provided at aggregated levels (e.g. per State, per Hospital Referral Region, etc.) to protect privacy. Unfortunately aggregation can severely diminish the utility of such data when modeling or analysis is desired at a per-individual basis. So, not surprisingly, despite the increasing abundance of aggregate data, there have been very few successful attempts in exploiting them for individual-level analyses. This paper introduces LUDIA, a novel low-rank approximation algorithm that utilizes aggregation constraints in addition to auxiliary information in order to estimate or "reconstruct" the original individual-level values from aggregate data. If the reconstructed data are statistically similar to the original individual-level data, off-the-shelf individual-level models can be readily and reliably applied for subsequent predictive or descriptive analytics. LUDIA is more robust to nonlinear estimates and random effects than other reconstruction algorithms. It solves a Sylvester equation and leverages multi-level (also known as hierarchical or mixed-effect) modeling approaches efficiently. A novel graphical model is also introduced to provide a probabilistic viewpoint of LUDIA. Experimental results using a Texas inpatient dataset show that individual-level data can be reasonably reconstructed from county-, hospital-, and zip code-level aggregate data. Several factors affecting the reconstruction quality are discussed, along with the implications of this work for current aggregation guidelines.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

Data aggregation, Low rank approximation, Multi-level model

KDD'14, August 24–27, 2014, New York, NY, USA.

http://dx.doi.org/10.1145/2623330.2623659 .

1. INTRODUCTION

Individual-level datasets that contain one or more records per person are rich sources for data mining applications. In the healthcare domain, the application of advanced data mining methods on individual level records across large populations can enable major breakthroughs in both personalized and population-level healthcare, leading to much improved, more cost-effective and timely diagnoses and interventions [33]. However, such data often contain a substantial amount of privacy-sensitive attributes. In practice, privacy concerns are typically addressed through multiple Statistical Disclosure Limitation (SDL) techniques [10], such as data aggregation [1], data swapping [8, 13], top-coding, feature generalization such as k-anonymity [36] or l-diversity [28], and additive random noise with measurement error [17]. Each method has distinct utility and risk aspects. Often an appropriate mix of disclosure limitation techniques is carefully chosen by domain experts and statisticians. For example, Centers for Medicare and Medicaid Services applied six different SDL techniques when publishing synthetic public use files[1]: variable reduction, suppression, substitution, imputation, data perturbation, and coarsening.

Among various SDL approaches, data aggregation is currently the most widely used. Data aggregation is a process of summarizing individual-level data into a small set of representative values such as mean and median statistics computed over groups that are typically geographically or administratively defined (such as county, hospital group, state, etc). This process is straightforward to apply on diverse datasets: wireless sensor networks [22], regional healthcare statistics [7], and government data [9]. Moreover, such aggregate data can be efficiently and effortlessly generated in RDBMS [29] and statistical programming languages [37]. Data collecting agencies publish various aggregate datasets at different levels of aggregation (including individual-level for non-sensitive information). In particular, the U.S. government's open data project, data.gov has recently released a substantial amount of regional and topic-based aggregate data regarding agriculture, education, and energy. Centers for Disease Control and Prevention annually publishes various regional statistics related to aging, cancer, and diabetes. Other notable sources of aggregated health data are dartmouthatlas.org and healthdata.gov.

The use of aggregate data is typically limited to group-level studies, often referred to as ecological studies for his-

[1] http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/SynPUFs/DE_Syn_PUF.html

Table 1: Illustrative health data files: artificial individual-level data (left) and aggregate-level summary (right) [24].

ID	Age	Length	State
1	19	1 day	TX
2	35	2 days	CA
3	3	10 days	FL
6	68	100 days	FL
⋮	⋮	⋮	⋮

State	Avg. Hospital Charge
CA	$ 2,706
FL	$ 1,809
NY	$ 1,954
TX	$ 2,001
⋮	⋮

toric reasons. Applying the result from aggregate data to individual-level inference often results in the classic problem of ecological fallacy [35]. Ecological fallacy occurs when aggregate-level statistics are misinterpreted as individual-level inferences. For example, the high correlation between "per capita consumption of dietary fat" and "breast cancer" in different countries [6] does not imply that dietary fat causes breast cancer.

There have been many attempts to circumvent the ecological fallacy while analyzing aggregate data. This is because individual-level data acquisition is usually expensive, and it is sometimes legally and ethically implausible. Duncan [11] developed the method of bounds that uses the constraints of contingency tables, but the bounds are often uninformative in real applications [14]. The constancy assumption, suggested by Goodman [21], allows an individual-level interpretation of ecological regression. Suppose that we want to check the relationship between Length of Stay (LoS) and Hospital Charge (HC) variables from state-level aggregate data:

$$\mathrm{HC}_{\mathrm{state}} \sim c_{\mathrm{state}} + \beta_{\mathrm{state}}\mathrm{LoS}_{\mathrm{state}}$$

The constancy assumption states that daily hospital charge rates are the same across different states i.e. $\beta_{\mathrm{state}} = \beta$ and $c_{\mathrm{state}} = c$. Of course, this assumption is rarely true in real datasets; for this example, it is more natural to assume that each state has a different daily charge rate, thereby indicating that multi-level modeling can be used [19]. Such an approach, however, is under-identified and can't be solved using aggregate data, since we have more parameters than observations. King [25, 26] proposed a Bayesian prior-based multi-level approach to overcome the limitation of Goodman's assumption, but Freedman [15] criticized that King's method cannot be validated on the basis of aggregate data.

We provide a novel approach for addressing the ecological fallacy dilemma by leveraging available sources of individual-level data for which the values of the partitioning or aggregation variable is known. For example, an aggregation variable can indicate state, county, or zip codes, that can be used to link to aggregate-level dataset that is aggregated along such geographical regions. In practice, it is not difficult to collect multiple datasets with different levels of aggregations from multiple agencies, so little added data-collection expense is involved.

Table 1 shows a simple, illustrative example of two health datasets. Non-sensitive fields are published at individual-level, while a sensitive field (hospital charge) is aggregated over the partition variable "state". Our approach is substantially different from previous ecological fallacy solutions where only aggregate data were considered.

We use a two-stage approach to avoid the ecological fallacy. We first reconstruct the masked individual-level variables from aggregate data, then apply multi-level regression models to the reconstructed data. In other words, we first synthesize "pseudo individual-level" data that are statistically similar to the original (unseen) individual-level data. Not only multi-level regression models, but also numerous off-the-shelf data mining algorithms can be easily applied to such pseudo individual-level data. Our reconstruction algorithm is based on two key observations:

- Aggregation is a linear transformation, thus it preserves several algebraic properties including the associative property.
- Using a proper data model, additional individual-level data can provide statistical clues for the reconstruction of the masked columns. From the previous hospital charge example, if we know *a priori* that hospital charge (aggregate-level) is a function of length of stay (individual-level), we roughly expect that a person with a longer stay may have paid more than a person who stayed only a day. We demonstrate that such clues can be captured using a low rank model.

We demonstrate our reconstruction algorithm on both simulated and real datasets. Many factors contribute to the reconstruction quality, for example, the number of data points per aggregation and correlation strength with other columns. These factors will be illustrated in Section 6 using Texas Inpatient Public Use Files. The main contributions of this paper are:

- We formulate a data model, LUDIA, that reconstructs individual values from aggregate values.
- We derive efficient algorithms for solving optimization problems associated with LUDIA.
- We show that our reconstructed data can capture aggregate-level random effects, thus the reconstructed data can be used for multi-level analyses as well as more sophisticated data mining applications.

The first two contributions will be illustrated in Section 3, the last contribution will be explained in Section 4. Experimental results are provided in Section 6, followed by discussions in Section 7.

2. PRELIMINARIES & RELATED WORK

This section starts by setting up the notation of this paper, and visiting two key existing approaches for tackling aggregate data. We extend these approaches to reconstruct the original individual-level data, and briefly discuss their modeling assumptions and limitations.

Aggregation is a compressive linear transformation, which we denote as $\mathbf{A}$. For example, suppose that there are five individuals from two different groups: the first two from Group A and the last three from Group B. Individual-level observations, say $\mathbf{y} = \begin{bmatrix}1 & 2 & 3 & 4 & 5\end{bmatrix}^\top$, can be aggregated into two groups by multiplying an aggregation matrix defined as follows:

$$\mathbf{s} = \begin{bmatrix}1.5\\ 4\end{bmatrix} = \mathbf{A}\mathbf{y} = \begin{bmatrix}1/2 & 1/2 & 0 & 0 & 0\\ 0 & 0 & 1/3 & 1/3 & 1/3\end{bmatrix}\mathbf{y}$$

Table 2 summarizes the notation of this paper.

Table 2: Notation. For simplicity but without loss of generality, we use $d = 1$ in this paper.

Symbol	Explanation
$\mathbf{X}$	$n \times m$, individual-level matrix
$\mathbf{x}_i$	$1 \times m$, ith row of $\mathbf{X}$
$\mathbf{y}$	$n \times d$, masked individual-level vector
$\mathbf{A}$	$p \times n$, aggregation matrix
$\mathbf{s}$	$p \times d$, aggregate-level vector i.e. $\mathbf{Ay}$
$\mathbf{U}, \mathbf{V}$	$n \times r$, $m \times r$ low-rank matrices

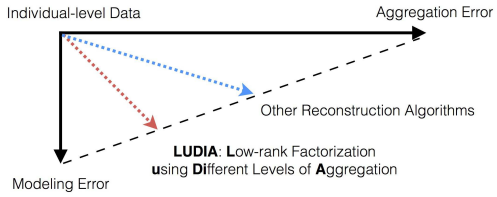


Figure 1: Reconstruction triangle and LUDIA.

The processes of aggregation and reconstruction can be illustrated as follows:

$$\text{(Compression)} \qquad \mathbf{Ay} \xrightarrow{\text{compressive linear}} \mathbf{s}$$

$$\text{(Reconstruction)} \qquad \hat{\mathbf{y}} \xleftarrow{\text{low-rank modeling}} \text{Recon}(\mathbf{s}, \mathbf{A}, \mathbf{X})$$

where $\mathbf{X}$ represents individual-level data, and Recon is a reconstruction algorithm. To give a brief overview, our reconstruction algorithm, LUDIA (**L**ow-rank factorization **U**sing **Di**fferent levels of **A**ggregation), is a constrained low-rank factorization algorithm that can capture multi-level effects. Figure 1 illustrates the overall idea of LUDIA and other reconstruction algorithms. We have two sources of errors that construct our reconstruction triangle: aggregation and modeling errors. LUDIA reduces the aggregation error using a low-rank model, but the LUDIA error is lower-bounded by the modeling error.

To illustrate existing approaches for aggregate data, let us consider the previous "hospital charge vs. length of stay" example. When using aggregate data, three approaches have been popular:

- The neighborhood model [16], proposed by Freedman, will imply that hospital charges are more influenced by geographical attributes rather than the length of stay variable, since each geographical partition is assumed to contain a homogeneous population group.
- Ecological regression [20], suggested by Goodman, will assume that the effect size of length of stay is the same across different states, based on the constancy assumption. According to the constancy assumption, geographical partitions are treated as different batches of i.i.d. experiments.
- Ecological inference, also known as King's method [26], combines the method of bounds and Goodman's ecological regression. King's method is a multi-level approach that models different effect sizes for different states. The multi-level parameters are first characterized by their acceptable regions using the method of bounds, then their joint distributions are modeled under three assumptions [27]: uni-modal joint distribution, absence of spatial correlation, and independence between multi-level coefficients and dependent variables. However, these assumptions are not verifiable on the basis of aggregate-level data [15], and this method requires manual tuning of the parameter distributions. In short, ecological inference is a method with many knobs and unverifiable assumptions, and we do not include this method in our baseline methods.

These previous approaches have been developed to tackle aggregate data, and need to be slightly modified to synthesize individual-level data. Imagine that we now obtained individual-level length of stay data[2] $\mathbf{X}$ and each individual's location information $\mathbf{A}$. To reconstruct the masked hospital charge data $\mathbf{y}$, two direct extensions from the previous approaches can be considered:

- Moore-Penrose (MP) solution is an extension of the neighborhood model. As the neighborhood model only focuses on the aggregation matrix $\mathbf{A}$, the reconstructed values are obtained by applying the Moore-Penrose pseudo-inverse of $\mathbf{A}$ to the aggregate data:
$$\hat{\mathbf{y}}_{\text{MP}} = \mathbf{A}^{+}\mathbf{s}$$
where $\mathbf{A}^{+} = \mathbf{A}(\mathbf{A}\mathbf{A}^{\top})^{-1}$.
- Ecological Regression (ER) solution is an extension of Goodman's ecological regression. Assuming that the effect sizes are the same across different states, we obtain the regression parameter $\boldsymbol{\beta}$ from the aggregate data, then apply to the individual-level covariate:
$$\hat{\mathbf{y}}_{\text{ER}} = \boldsymbol{\beta}_{\text{ER}}\mathbf{X}$$
where $\boldsymbol{\beta}_{\text{ER}} = (\mathbf{Z}^{\top}\mathbf{Z})^{-1}\mathbf{Z}^{\top}\mathbf{s}$ and $\mathbf{Z}$ is the aggregate-level representation of $\mathbf{X}$ i.e. $\mathbf{AX}$.

MP and ER exhibit different failure modes. MP ignores the effects of individual-level covariates, which may substantially leverage the utility of aggregate data. On the other hand, ER relies on the constancy assumption, which is rarely true in real settings.

For our hospital charge example, daily charge rates are significantly different across city and rural areas (see Section 6). This geographical variation on daily charge rates can be expressed as follows:

$$\begin{aligned} y_i &= \mathbf{x}_i\boldsymbol{\beta}_{\text{state}} + \zeta_{\text{state}} + e_i & e_i &\sim \text{N}(0, \sigma_e^2) \\ \boldsymbol{\beta}_{\text{state}} &= \boldsymbol{\beta}_{\text{global}} + \boldsymbol{\eta}_{\text{state}} & \boldsymbol{\eta}_{\text{state}} &\sim \text{N}(0, \sigma_\eta^2\mathbf{D}) \\ \zeta_{\text{state}} &\sim \text{N}(0, \sigma_\zeta^2) \end{aligned}$$

where ζ_{state} and $\boldsymbol{\eta}_{\text{state}}$ represent state-level biases for the intercept and slope; they are called random intercept and random slope, respectively. Assuming that we have two states A and B, and individuals listed by state, this multi-level approach [18] can be written in a matrix form:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_{n-1} \\ \mathbf{x}_n \end{bmatrix} \boldsymbol{\beta}_{\text{global}} + \begin{bmatrix} \mathbf{x}_1 & \mathbf{0} & 1 & 0 \\ \mathbf{x}_2 & \mathbf{0} & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{x}_{n-1} & 0 & 1 \\ \mathbf{0} & \mathbf{x}_n & 0 & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{\eta}_A \\ \boldsymbol{\eta}_B \\ \zeta_A \\ \zeta_B \end{bmatrix} + \mathbf{E}$$

[2]**Note**: This simple example has only one individual level (LoS) and one aggregated (HC) feature, and one level of aggregation, called "State", so as to convey the concepts most easily. Our approach readily generalizes to multiple individual and aggregate variables as well as multiple levels of aggregations, as will be seen later

We define new matrices $\boldsymbol{\gamma}$ (random effects) and $\mathbf{G}$ (covariates for random effects) to obtain a compact form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_{\text{global}} + \mathbf{G}\boldsymbol{\gamma} + \mathbf{E} \tag{1}$$

Aggregate data are obtained through the aggregation operation as follows:

$$\begin{aligned} \mathbf{A}\mathbf{y} &= \mathbf{A}\mathbf{X}\boldsymbol{\beta}_{\text{global}} + \mathbf{A}\mathbf{G}\boldsymbol{\gamma} + \mathbf{A}\mathbf{E} \\ \Rightarrow \quad \mathbf{s} &= \mathbf{Z}\boldsymbol{\beta}_{\text{global}} + \mathbf{A}\mathbf{G}\boldsymbol{\gamma} + \mathbf{A}\mathbf{E} \end{aligned}$$

As can be seen, the ER solution is valid only if

- $\boldsymbol{\gamma} = 0$ (no random effect)
- $((\mathbf{A}\mathbf{X})^\top \mathbf{A}\mathbf{X})^{-1}(\mathbf{A}\mathbf{X})^\top \mathbf{A}\mathbf{y} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$

These two conditions are rarely realistic in real applications.

MP and ER are formulated based on two orthogonal assumptions. MP assumes that only geographical partitions affect the dependent variable, while ER posits that geographical partitions are merely random groupings. These assumptions are necessary to obtain some meaningful results from aggregate data, as the ecological fallacy is, in fact, the problem of statistical under-identification [34]. However, the direct extensions from the previous approaches do not utilize the full potential of auxiliary individual-level data.

A recent breakthrough in the use of aggregate data to augment individual-level models was made by Park and Ghosh [30, 32, 31]. The suggested model, CUDIA, is a probabilistic clustering algorithm that utilizes both aggregated and individual-level data. CUDIA models the data points as being generated from a mixture of exponential family distributions. The parameters of CUDIA are estimated using a Monte-Carlo Expectation Maximization (MCEM) algorithm. Although CUDIA can reasonably reconstruct the data based on the estimated cluster centers, the primary objective of CUDIA is still clustering rather than reconstruction. Furthermore, the presented MCEM algorithm is not scalable to large-scale data. We show that CUDIA is, in fact, a special case of LUDIA with a non-negative constraint on $\mathbf{U}$ (see Section 5). LUDIA generalizes CUDIA with a more flexible representation of $\mathbf{U}$. This generalization provides an efficient optimization algorithm that is suitable for large-scale data.

3. LUDIA

LUDIA is a low-rank factorization algorithm using aggregate data. We first describe the underlying data model of LUDIA, then formulate LUDIA's objective function. Because of the non-trivial aggregation constraint, we derive a customized minimization approach that uses the Sylvester equation.

3.1 Low-rank Data Model

LUDIA employs a bottom-up approach starting from individual level data. We first design a data model for a complete matrix $\mathbf{D} = \begin{bmatrix}\mathbf{X} & \mathbf{y}\end{bmatrix}$, then formulate an objective function when $\mathbf{y}$ is masked and only $\mathbf{s} = \mathbf{A}\mathbf{y}$ is provided. The data model for LUDIA is based on the low-rank approximation theory as follows:

$$\begin{bmatrix}\mathbf{X} & \mathbf{y}\end{bmatrix} = \mathbf{U}\mathbf{V}^\top + \mathbf{E} = \mathbf{U}\begin{bmatrix}\mathbf{V}_x^\top & \mathbf{v}_y^\top\end{bmatrix} + \mathbf{E} \tag{2}$$

where $\mathbf{U} \in \mathbb{R}^{n\times r}$, $\mathbf{V} \in \mathbb{R}^{m\times r}$, $\mathbf{E} \in \mathbb{R}^{n\times m}$, and $r \le \min(n, m)$. Note that we divided $\mathbf{V}$ into two block matrices: $\mathbf{V}_x$ and $\mathbf{v}_y$, so that $\mathbf{X} \approx \mathbf{U}\mathbf{V}_x$ and $\mathbf{y} \approx \mathbf{U}\mathbf{v}_y$.

The main objective of this paper is to reconstruct the masked values, $\mathbf{y}$. In theory, under certain assumptions such as an underlying low-rank structure and a uniform missing mechanism, missing values in a matrix can be reconstructed. Candes and Recht [5] showed that, for matrix entries that are missing at random, they can be exactly recovered if the number of observations exceed a certain threshold value. However, the settings for the matrix completion problem are not suitable for our problem, since we consider a situation wherein one or more columns of a matrix is entirely missing, but its aggregated statistics are given.

We approximate the original matrix using two low-rank matrices. This problem is different from the matrix completion problem [23]. Low-rank approximation is typically posed as a minimization problem as follows:

$$\min \|\hat{\mathbf{D}} - \mathbf{D}\|_F^2 \quad \text{s.t.} \quad \text{rank}(\hat{\mathbf{D}}) \le r$$

where $\mathbf{D}$ and $\hat{\mathbf{D}}$ are both $n\times m$ matrices, and $r \le \min(n, m)$. The Eckart-Young-Mirsky theorem [12] says that rank r approximation of the data matrix $\mathbf{D}$ is given as follows:

$$\hat{\mathbf{D}} = \mathbf{U}\boldsymbol{\Gamma}\mathbf{V}^\top = (\mathbf{U}\boldsymbol{\Gamma}^{1/2})(\boldsymbol{\Gamma}^{1/2}\mathbf{V}^\top) = \tilde{\mathbf{U}}\tilde{\mathbf{V}}^\top$$

where $\mathbf{U}$, $\boldsymbol{\Gamma}$, $\mathbf{V}$ are $n\times r$, $r\times r$, $m\times r$ truncated Singular Vector Decomposition matrices, respectively. The data model in Equation 2 is, however, inapplicable to our reconstruction application. The model should instead reflect the constraint that $\mathbf{y}$ is masked and only $\mathbf{s} = \mathbf{A}\mathbf{y}$ is provided.

3.2 Aggregation Constraint

A novel optimization problem for three latent matrices $\mathbf{y}$, $\mathbf{U}$, and $\mathbf{V}$ is proposed as follows:

$$\begin{aligned} \min_{\mathbf{y},\mathbf{U},\mathbf{V}} \ & \|\begin{bmatrix}\mathbf{X} & \mathbf{y}\end{bmatrix} - \mathbf{U}\mathbf{V}^\top\|_F^2 \\ \text{subject to} \ & \mathbf{A}\mathbf{y} = \mathbf{s} \end{aligned} \tag{3}$$

A simultaneous minimization over $\mathbf{y}$, $\mathbf{U}$, and $\mathbf{V}$ is a difficult non-convex optimization problem. However, minimization over one set of variables alone is a convex problem.

We tackle this problem by removing the equality constraint. The equality constraint on $\mathbf{y}$ can be eliminated if we fix the other two variables. Given that $\mathbf{U}$ and $\mathbf{V}$ are fixed, the optimality condition [4] is given as:

$$\mathbf{A}\mathbf{y}^* = \mathbf{s} \quad \text{and} \quad \nabla f(\mathbf{y}^*) + \mathbf{A}^\top\boldsymbol{\Psi}^* = 0$$

where $f(\mathbf{Y}) = \|\mathbf{X} - \mathbf{U}\mathbf{V}_x^\top\|_F^2 + \|\mathbf{y} - \mathbf{U}\mathbf{v}_y^\top\|_2^2$ and $\boldsymbol{\Psi}^* \in \mathbb{R}^p$ is a dual variable. $\mathbf{Y}^*$ is optimal if and only if there exists $\boldsymbol{\Psi}^*$ satisfying the optimality conditions. It turns out that, for this system, $\mathbf{y}^*$ can be solved in a closed form.

To eliminate the constraint, we solve Karush-Kuhn-Tucker (KKT) equations as follows:

$$\nabla f(\mathbf{y}^*) + \mathbf{A}^\top\boldsymbol{\Psi}^* = 2\mathbf{y}^* - 2\mathbf{U}\mathbf{v}_y^\top + \mathbf{A}^\top\boldsymbol{\Psi}^* = 0$$

We multiply $\mathbf{A}$ on both sides of the second KKT equation, and solve for $\boldsymbol{\Psi}^*$:

$$\begin{aligned} 2\mathbf{A}\mathbf{y}^* - 2\mathbf{A}\mathbf{U}\mathbf{v}_y^\top + \mathbf{A}\mathbf{A}^\top\boldsymbol{\Psi}^* &= 0 \\ \boldsymbol{\Psi}^* &= -2(\mathbf{A}\mathbf{A}^\top)^{-1}(\mathbf{s} - \mathbf{A}\mathbf{U}\mathbf{v}_y) \end{aligned}$$

Thus, the optimal $\mathbf{y}^*$ is:

$$\mathbf{y}^* = \mathbf{U}\mathbf{v}_y^\top + \mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top)^{-1}(\mathbf{s} - \mathbf{A}\mathbf{U}\mathbf{v}_y^\top) \tag{4}$$

We plug the optimal $\mathbf{y}^*$ into the original objective function to obtain:

$$\min_{\mathbf{U},\mathbf{V}} \|\mathbf{X} - \mathbf{U}\mathbf{V}_x\|_F^2 + (\mathbf{s} - \mathbf{A}\mathbf{U}\mathbf{v}_y)^\top (\mathbf{A}\mathbf{A}^\top)^{-1}(\mathbf{s} - \mathbf{A}\mathbf{U}\mathbf{v}_y) \tag{5}$$

We have thus transformed the original objective function with three variables and an equality constraint into a simpler unconstrained objective function with two variables.

3.3 Objective Function

Although we simplified the constrained optimization problem to the non-constrained optimization problem in Equation 5, solving the objective function poses another challenge. Intuitively, one can approach the problem using an alternating minimization approach over $\mathbf{U}$ and $\mathbf{V}$. Solving for $\mathbf{U}$, however, does not have a closed form solution, because the low rank matrix $\mathbf{U}$ is surrounded by $\mathbf{A}$ and $\mathbf{v}_y$. Using a divide-and-conquer approach, we can solve for one row $\mathbf{u}_i$ of $\mathbf{U}$, and iterate over the entire rows. This divide-and-conquer approach is, however, susceptible to the sequence of rows, and cannot be generalized to an arbitrary aggregation matrix.

We propose a simple and efficient optimization solution by introducing an auxiliary variable $\mathbf{\Pi} = \mathbf{A}\mathbf{U}$ where we treat $\mathbf{\Pi}$ as an independent variable. We also relax the hard relationship between $\mathbf{\Pi}$ and $\mathbf{U}$ as a penalty term.

Combining these two tricks, our new objective function is written as follows:

$$\min_{\mathbf{U},\mathbf{V},\mathbf{\Pi}} \|\mathbf{X} - \mathbf{U}\mathbf{V}_x^\top\|_F^2 + \|\sqrt{\mathbf{W}}(\mathbf{s} - \mathbf{\Pi}\mathbf{v}_y^\top)\|_2^2 + \|\mathbf{A}\mathbf{U} - \mathbf{\Pi}\|_F^2 \tag{6}$$

where $\mathbf{W} = (\mathbf{A}\mathbf{A}^\top)^{-1}$. This objective function is LUDIA's objective function, and denote as $\mathcal{L}(\mathbf{U}, \mathbf{V}, \mathbf{\Pi})$. We now apply our alternating minimization technique to Equation 6.

3.3.1 Solving for U

First, we derive the partial derivative of the LUDIA objective function with respect to $\mathbf{U}$:

$$\frac{\partial \mathcal{L}(\mathbf{U}, \mathbf{V}, \mathbf{\Pi})}{\partial \mathbf{U}} = -\mathbf{X}\mathbf{V}_x + \mathbf{U}\mathbf{V}_x^\top \mathbf{V}_x + \mathbf{A}^\top \mathbf{A}\mathbf{U} - \mathbf{A}^\top \mathbf{\Pi} = 0$$

Rearranging the terms, we obtain:

$$\mathbf{U}\mathbf{V}_x^\top \mathbf{V}_x + \mathbf{A}^\top \mathbf{A}\mathbf{U} = \mathbf{X}\mathbf{V}_x + \mathbf{A}^\top \mathbf{\Pi}$$

This is a type of a Sylvester equation [2]. This form of equation widely appears in the field of control theory [3], and the continuous Lyapanov equation is a special case of the Sylvester equation. If $\mathbf{V}_x^\top \mathbf{V}_x$ and $\mathbf{A}^\top \mathbf{A}$ have no common eigenvalues, a unique solution exists and it is given as:

$$\text{vec}(\mathbf{U}) = (\mathbf{V}_x^\top \mathbf{V}_x \otimes \mathbf{I}_n + \mathbf{I}_r \otimes \mathbf{A}^\top \mathbf{A})^{-1} \text{vec}(\mathbf{X}\mathbf{V}_x + \mathbf{A}^\top \mathbf{\Pi})$$

where vec is a vectorization operator, and $\otimes$ represents the Kronecker product. For example, vec($\mathbf{U}$) is defined as:

$$\text{vec}(\mathbf{U}) = \begin{bmatrix} u_{1,1} & \dots & u_{n,1} & u_{1,2} & \dots & u_{1,r}, \dots u_{n,r} \end{bmatrix}^\top$$

3.3.2 Solving for Π

Next, we derive a partial derivative of the LUDIA objective function with respect to $\mathbf{\Pi}$:

$$\frac{\partial \mathcal{L}(\mathbf{U}, \mathbf{V}, \mathbf{\Pi})}{\partial \mathbf{\Pi}} = -\mathbf{W}(\mathbf{s} - \mathbf{\Pi}\mathbf{v}_y^\top)\mathbf{v}_y - \mathbf{A}\mathbf{U} + \mathbf{\Pi} = 0$$

Rearranging the terms, we obtain another Sylvester equation:

$$\mathbf{W}\mathbf{\Pi}\mathbf{v}_y^\top \mathbf{v}_y + \mathbf{\Pi} = \mathbf{W}\mathbf{s}\mathbf{v}_y + \mathbf{A}\mathbf{U}$$

The solution is given as:

$$\text{vec}(\mathbf{\Pi}) = (\mathbf{I}_r \otimes \mathbf{I}_p + \mathbf{v}_y^\top \mathbf{v}_y \otimes \mathbf{W})^{-1} \text{vec}(\mathbf{W}\mathbf{s}\mathbf{v}_y + \mathbf{A}\mathbf{U})$$

3.3.3 Solving for V

Finally, we derive closed form update equations for two block matrices $\mathbf{V}_x$ and $\mathbf{v}_y$. The partial derivative with respect to $\mathbf{V}_x$ is given as:

$$\frac{\partial \mathcal{L}(\mathbf{U}, \mathbf{V}, \mathbf{\Pi})}{\partial \mathbf{V}_x} = -\mathbf{U}^\top (\mathbf{X} - \mathbf{U}\mathbf{V}_x) = 0$$

Rearranging the terms, we obtain:

$$\mathbf{V}_x = (\mathbf{U}^\top \mathbf{U})^{-1} \mathbf{U}^\top \mathbf{X}$$

Similarly, the partial derivative with respect to $\mathbf{v}_y$ is:

$$\frac{\partial \mathcal{L}(\mathbf{U}, \mathbf{V}, \mathbf{\Pi})}{\partial \mathbf{v}_y} = -\mathbf{\Pi}^\top \mathbf{W}(\mathbf{s} - \mathbf{\Pi}\mathbf{v}_y^\top) = 0$$

Thus, the update form is:

$$\mathbf{v}_y^\top = (\mathbf{\Pi}^\top \mathbf{W}\mathbf{\Pi})^{-1} \mathbf{\Pi}^\top \mathbf{W}\mathbf{s}$$

3.4 Algorithm

Algorithm 1 summarizes our alternating minimization approach combining three different minimization equations for $\mathbf{U}$, $\mathbf{\Pi}$, and $\mathbf{V}$. The algorithm takes three input matrices: individual-level matrix $\mathbf{X}$, aggregation matrix $\mathbf{A}$, and aggregate-level matrix $\mathbf{s}$. The output of the algorithm is the reconstructed individual-level data $\hat{\mathbf{y}}$. The algorithm does not require any other parameters.

Algorithm 1: LUDIA Estimation Algorithm

Data: $\mathbf{X}, \mathbf{A}, \mathbf{s}$
Result: $\hat{\mathbf{y}}$
$r = \text{rank}(\mathbf{X})$);
$\bar{\mathbf{y}} = \mathbf{A}^+ \mathbf{s}$;
$\mathbf{U}, \mathbf{V} = \text{SVD}(\begin{bmatrix}\mathbf{X} & \bar{\mathbf{y}}\end{bmatrix}, \text{rank} = r)$;
$\mathbf{\Pi} = \mathbf{A}\mathbf{U}$;
while *not converged* **do**
  $\text{vec}(\mathbf{U}) = (\mathbf{V}_x^\top \mathbf{V}_x \otimes \mathbf{I}_n + \mathbf{I}_r \otimes \mathbf{A}^\top \mathbf{A})^{-1} \text{vec}(\mathbf{X}\mathbf{V}_x + \mathbf{A}^\top \mathbf{\Pi})$;
  $\text{vec}(\mathbf{\Pi}) = (\mathbf{I}_r \otimes \mathbf{I}_p + \mathbf{v}_y^\top \mathbf{v}_y \otimes \mathbf{W})^{-1} \text{vec}(\mathbf{W}\mathbf{s}\mathbf{v}_y + \mathbf{A}\mathbf{U})$;
  $\mathbf{V}_x = (\mathbf{U}^\top \mathbf{U})^{-1} \mathbf{U}^\top \mathbf{X}$;
  $\mathbf{v}_y^\top = (\mathbf{\Pi}^\top \mathbf{A}\mathbf{A}^\top \mathbf{\Pi})^{-1} \mathbf{\Pi}^\top \mathbf{A}\mathbf{A}^\top \mathbf{s}$;
end
$\hat{\mathbf{y}} = \mathbf{U}\mathbf{v}_y^\top$;
// correction equation;
$\hat{\mathbf{y}} = \hat{\mathbf{y}} + \mathbf{A}^+(\mathbf{s} - \mathbf{A}\hat{\mathbf{y}})$

The initialization of $\mathbf{U}$ and $\mathbf{V}$ is based on the MP solution. We first pseudo-reconstruct the masked individual-level data using MP, then run SVD on the pseudo-complete matrix. The rank parameter of the SVD algorithm is given as the rank of $\mathbf{X}$. This setting captures both our low-rank data model and a linear model defined as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$. If this linear model is the true underlying data model for the data, then the rank of the complete matrix is the same as the rank of $\mathbf{X}$.

The last line of the algorithm calibrates the final output. Recall that the optimal $\mathbf{y}^*$ was given in Equation 4. This

correction equation ensures that the aggregation of the reconstructed values are the same as the given aggregate data i.e. $\mathbf{s} = \mathbf{A}\hat{\mathbf{y}}$. However, if the aggregate values do not necessarily need to match the reconstructed values (possibly from noise or sub-sampling), we can ignore the last line of the algorithm.

4. EXTENSIONS

We illustrate two extensions of the LUDIA algorithm. The first extension shows that LUDIA can directly incorporate multi-level data models. This extended reconstruction method can capture group-level effects, which were not possible in classical frameworks. The second extension explores whether we can improve the reconstruction quality if we have multiple levels of aggregate data.

4.1 Multi-level Modeling

The ecological fallacy problem is essentially "statistical under-identification" [34]. For aggregate data analyses, the maximum degrees of freedom are limited by the number of partitions. Individual-level analyses, such as multi-level models [19], often require more parameters than the number of partitions. This under-identification problem is traditionally approached by more assumptions; Goodman's and King's assumptions are two extreme cases. These assumptions are usually unrealistic, and they are almost impossible to verify on the basis of aggregate data.

Smartly utilizing auxiliary individual-level data can provide higher degrees of freedom than the number of partitions. The key observation comes from the connection between the degrees of freedom and the rank of a full matrix. Suppose that a target $\mathbf{y}$ is a function of r degrees of freedom. Then the rank of the full matrix $\begin{bmatrix}\mathbf{X} & \mathbf{y}\end{bmatrix}$ is r, since $\mathbf{y}$ can be expressed by a linear combination of $\mathbf{X}$. Analogously, if a target is a multi-level function of r variables and p levels, then the degrees of freedom for this model is given by $(r \times p)$. To capture the variability of the target, the corresponding full matrix needs to have the rank of $(r \times p)$. In this section, we show that this rank augmentation can be seamlessly integrated with the LUDIA framework.

As illustrated in Equation 1, a multi-level model can be compactly written as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\gamma} + \mathbf{E} \approx \begin{bmatrix}\mathbf{X} & \mathbf{G}\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta} \\ \boldsymbol{\gamma}\end{bmatrix}$$

where $\boldsymbol{\gamma} \in \mathbb{R}^{l\times 1}$ is a random effect vector, and $\mathbf{G} \in \mathbb{R}^{n\times l}$ represents encoded covariates according to $\boldsymbol{\gamma}$. For this model, the degrees of freedom are given as $(r+l)$ where $r = \text{rank}(\mathbf{X})$. The full matrix has $(r+l+1)$ columns, and this matrix can be written as a product of two rank $(r+l)$ matrices.

To fully reconstruct the masked individual-level data, the rank of our low-rank model should be at least $(r+l)$. This can be achieved by augmenting the data by l:

$$\begin{bmatrix}\mathbf{X} & \mathbf{G} & \mathbf{y}\end{bmatrix} \approx \begin{bmatrix}\mathbf{U} & \tilde{\mathbf{U}}\end{bmatrix}\begin{bmatrix}\mathbf{V}_x^\top & \tilde{\mathbf{V}}_{x1} & \mathbf{v}_y^\top \\ \tilde{\mathbf{V}}_{x2} & \tilde{\mathbf{V}}_{x3} & \mathbf{v}_a^\top\end{bmatrix}$$

where $\tilde{\mathbf{U}} \in \mathbb{R}^{n\times l}$, $\tilde{\mathbf{V}}_{x\cdot} \in \mathbb{R}^{l\times l}$, and $\mathbf{v}_a \in \mathbb{R}^{1\times l}$.

Although one can run LUDIA with these augmented terms, we show that a simple post-processing approach can mimic the result from this augmentation. The block matrix $\tilde{\mathbf{V}}_x$ can be treated as a nuance parameter, since it does not directly affect the reconstruction of $\mathbf{y}$. The trick is to specify our low-rank matrices to be of a specific form as follows:

$$\begin{bmatrix}\mathbf{X} & \mathbf{G} & \mathbf{y}\end{bmatrix} \approx \begin{bmatrix}\mathbf{U} & \mathbf{G}\end{bmatrix}\begin{bmatrix}\mathbf{V}_x^\top & \mathbf{0} & \mathbf{v}_y^\top \\ \mathbf{0} & \mathbf{I}_l & \mathbf{v}_a^\top\end{bmatrix}$$

Then we do not need to estimate $\tilde{\mathbf{U}}$ and $\tilde{\mathbf{V}}_x$, but only $\mathbf{v}_a$. The augmented term $\mathbf{v}_a$ needs to minimize the second term of Equation 6:

$$\min_{\mathbf{v}_a} \|\sqrt{\mathbf{W}}(\mathbf{s} - \mathbf{A}\begin{bmatrix}\mathbf{U} & \mathbf{G}\end{bmatrix}\begin{bmatrix}\mathbf{v}_y^\top \\ \mathbf{v}_a^\top\end{bmatrix})\|_2^2$$

The solution for this minimization problem is given as follows:

$$\hat{\mathbf{v}}_a^\top = ((\mathbf{AG})^\top\mathbf{WAG})^{-1}(\mathbf{AG})^\top\mathbf{W}(\mathbf{s} - \mathbf{AU}\mathbf{v}_y^\top)$$

Using this $\mathbf{v}_a$, we calibrate the reconstruction of $\mathbf{y}$:

$$\hat{\mathbf{y}} = \mathbf{U}\mathbf{v}_y^\top + \mathbf{G}\hat{\mathbf{v}}_a^\top$$

This adjustment equation mimics the original augmentation.

This data augmentation technique for multi-level modeling is not suitable for the MP and ER frameworks. MP only focuses on the aggregation matrix, and does not involve individual-level covariates. Adding the augmented block matrix $\mathbf{G}$ requires a different approach. The number of covariates in ER is upper-bounded by the number of partitions. The simplest multi-level model, a random intercept model, requires the number of covariates to be the same as the number of partitions. LUDIA utilizes the full potential of individual-level covariates, and thus it can be easily extended to more complex models.

4.2 Aggregation Stacking

Thus far, we have considered only one source of aggregate data. There can be many levels of groupings based on geography, administration, or other factors. This section answers how one can further improve the reconstruction quality with additional aggregate data.

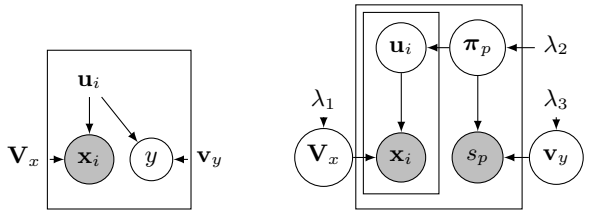
The key trick is to stack two aggregate-level datasets and create a new aggregate dataset. Algorithm 2 illustrates this approach. In the algorithm, we have two sources of aggregate data: $(\mathbf{A}_1, \mathbf{s}_1)$ and $(\mathbf{A}_2, \mathbf{s}_2)$. For example, there can be county-level and state-level aggregate data, respectively. This kind of augmentation can further improve the reconstruction accuracy. This is because we have more constraints on $\mathbf{y}$, and the degrees of freedom for $\mathbf{y}$ decrease accordingly.

Algorithm 2: LUDIA with Aggregation Stacking

Data: $\mathbf{X}, \mathbf{A}_1, \mathbf{s}_1, \mathbf{A}_2, \mathbf{s}_2$
Result: $\hat{\mathbf{y}}$
$\mathbf{A} = \begin{bmatrix}\mathbf{A}_1 \\ \mathbf{A}_2\end{bmatrix}$ and $\mathbf{s} = \begin{bmatrix}\mathbf{s}_1 \\ \mathbf{s}_2\end{bmatrix}$;
$\hat{\mathbf{y}} = \text{LUDIA}(\mathbf{X}, \mathbf{A}, \mathbf{s})$;

5. PROBABILISTIC INTERPRETATION

This section presents a probabilistic interpretation of the proposed LUDIA objective function. Figure 2a shows our low-rank model for the complete data. Note that the node for y is not shaded, since the variable is masked. To incorporate the aggregation constraint, we draw another plate that represents groupings. Figure 2b illustrates the graphical model for LUDIA. Each $\mathbf{u}_i$ in a group is assumed to be



(a) Low-rank model for complete data. (b) Low-rank model with the aggregation constraint.

Figure 2: Probabilistic Models for LUDIA.

drawn from a multivariate Gaussian centered at $\boldsymbol{\pi}_p$. Thus, the log-likelihood $\log p(\mathbf{U}, \mathbf{V}, \boldsymbol{\Pi} \mid \mathbf{X}, \mathbf{s})$ of LUDIA is written as:

$$
\begin{aligned}
&- (\mathbf{X} - \mathbf{U}\mathbf{V}_x^\top)^\top \boldsymbol{\Sigma}_x^{-1} (\mathbf{X} - \mathbf{U}\mathbf{V}_x^\top) \\
&- (\mathbf{s} - \boldsymbol{\Pi}\mathbf{v}_y^\top)^\top \boldsymbol{\Sigma}_y^{-1} (\mathbf{s} - \boldsymbol{\Pi}\mathbf{v}_y^\top) \\
&- (\boldsymbol{\Pi} - \mathbf{A}\mathbf{U})^\top \boldsymbol{\Sigma}_\pi^{-1} (\boldsymbol{\Pi} - \mathbf{A}\mathbf{U}) + \text{const.}
\end{aligned}
$$

In our setting, each row of $\mathbf{X}$ is i.i.d., thus $\boldsymbol{\Sigma}_x$ can be modeled as an identity matrix $\mathbf{I}_n$. Before characterizing $\boldsymbol{\Sigma}_y$, we first show that $\mathbf{A}\mathbf{A}^\top$ is invertible and positive-semidefinite. This property can be shown from the fact that $\text{rank}(\mathbf{A}) = p$ and $\mathbf{A}\mathbf{A}^\top \in \mathbb{R}^{p \times p}$. Moreover, the (p, p)th diagonal component of $(\mathbf{A}\mathbf{A}^\top)^{-1}$ is the same as n_p, the number of data points in group p. Thus, $\mathbf{A}\mathbf{A}^\top$ can replace $\boldsymbol{\Sigma}_y$. Finally, if we assume that $\boldsymbol{\Sigma}_\pi = \mathbf{I}_p$, then this log-likehood is actually a negative of the LUDIA objective function.

To show the connection to CUDIA, let us assume that we restrict the shape of $\mathbf{U}$ to be as follows:

$$
\mathbf{U}_C \quad \text{s.t.} \quad u_{ij} \in \{0, 1\} \quad \text{and} \quad \sum_j u_{ij} = 1
$$

In other words, each column of $\mathbf{U}$ becomes an indicator column for clusters. The rank parameter r of LUDIA is now interpreted as r different clusters, and $\mathbf{V}$ represents cluster centers. If we plug in this constraint to the LUDIA's log-likelihood function, we obtain the log-likelihood of CUDIA. Although this formulation may provide a different perspective on combining multiple sources of data, the minimization of the CUDIA objective function is more complicated to solve because of the non-negative constraint. Thus, CUDIA requires a computationally heavy MCEM algorithm, or greedy deterministic algorithm [31]. As the non-negative case is a special case of $\mathbf{U}$, we also have:

$$
\|\mathbf{D} - \mathbf{U}\mathbf{V}^\top\|_F^2 \leq \|\mathbf{D} - \mathbf{U}_C\mathbf{V}^\top\|_F^2
$$

This is why the CUDIA imputation is not so suitable for complex modeling such as multi-level modeling and non-linear estimates, while the LUDIA reconstruction provides valid inferences in such situations (see Section 6).

6. EXPERIMENTAL RESULTS

We provide experimental results using simulated data and Texas Inpatient Discharge data. A simulated dataset is used to illustrate the differences between ER, MP, and LUDIA. Next, we illustrate reconstruction tasks using actual health data. In this set of experiments, we mask sensitive columns, then show how well LUDIA can reconstruct the masked original values for different analytical tasks including non-linear estimates and multi-level modeling.

6.1 Simulated Data

We generate four different simulated datasets as follows:

- Low-Rank (LR) model emulates the model assumption of LUDIA. The parameters are given as $r = 2$ and $m = 4$. The equation for simulated data is as follows:

  $$
  \begin{bmatrix} \mathbf{X} & \mathbf{y} \end{bmatrix} = \mathbf{U}\mathbf{V}^\top + \mathbf{E}
  $$

  where $\mathbf{U}$ and $\mathbf{V}$ are drawn from the standard normal distribution, and the noise matrix $\mathbf{E}$ is drawn from a normal distribution with 0.4 standard deviation.

- Fixed Effect (FE) model emulates the model assumption of ER. We generate individual-level matrices with $m = 2$ from the standard normal distribution. The model equation is:

  $$
  \mathbf{y} = \mathbf{c} + \mathbf{X}\boldsymbol{\beta} + \mathbf{E}
  $$

  where $\mathbf{E}$ is drawn from a normal distribution with 0.2 standard deviation.

- Random Intercept (RE1) and Random Slope (RE2) model check whether the LUDIA's multi-level argument is valid. The model equation is:

  $$
  \mathbf{y} = \mathbf{c} + \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\gamma} + \mathbf{E}
  $$

  where $\boldsymbol{\gamma}$ is drawn from a normal distribution with 0.2 standard deviation.

We fix the number of partitions to be five, and vary the number of total data points. Aggregation matrices are generated using random assignment of partitions.

Figure 3 shows the reconstruction errors for different simulated data and reconstruction methods. Each cell represents a different simulated dataset, and the horizontal axes represent the number of data points per partition. The lower the curve is, the better the reconstruction quality is. MP is not affected by the number of data points per partition, but its performance is the worst from the experiments. The performance of ER is comparable with that of LUDIA for the FE dataset, but it does not capture the low-rank structure and random effects. For the random effect datasets, ER is largely affected by the number of data points per partition. LUDIA shows robust and stable performances over different datasets.

Figure 4 shows the reconstructed values compared to the original values from the RE1 dataset. The leftmost first two cells show the reconstructed values from MP and ER, respectively. In this figure, we show three different initialization methods for LUDIA: MP, ER, and random initialization methods. The alternating minimization approach of LUDIA does not guarantee the convergence to the global optimum, and the algorithm is susceptible to initial points. All three initialization methods provide comparable performances, and it would be worthwhile to investigate the better choice of initialization methods. The rest of the experiments use the MP initialization to maintain the consistency of our algorithm.

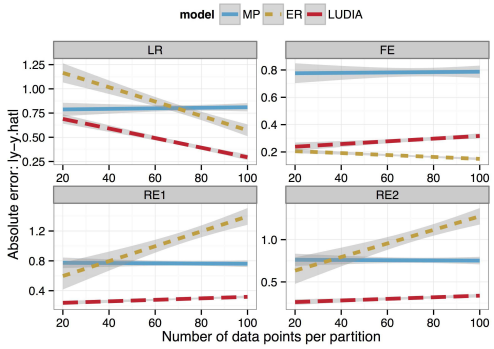


Figure 3: Reconstruction error vs. number of data points per partition. Except the FE case, the LUDIA reconstruction shows the least absolute errors.

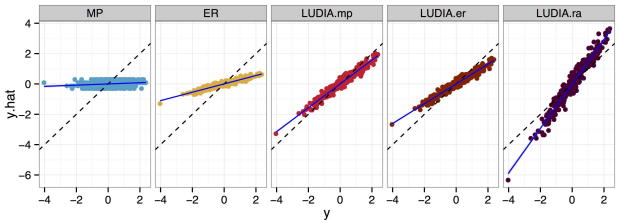


Figure 4: Reconstructed vs. original. We show three different initializations for LUDIA: MP, ER, and random initializations. All these three LUDIA reconstructions are closer to the original values, and the MP initialization performs the best.

6.2 Texas Inpatient Data

We use Texas Inpatient Public Use Data File [38] from the Texas Department of State Health Services (DSHS). Hospital billing records collected from 1999 to 2007 are publicly available through their website. Each yearly dataset contains about 2.8 millions events with more than 250 features including hospital name, county, patient ZIP codes, etc. Specifically, we use the inpatient records from Central Texas in the fourth quarter of 2006. Except for a few exempt hospitals, all the hospitals in Texas reported inpatient discharge events to DSHS. The public use data file we use is a subset of the DSHS's hospital discharge database. Our primary interest is the hospital charge for normal delivery. We aggregate the individual-level hospital charges at county-, hospital-, and ZIP code-levels. We assume that some of the individual-level covariates are available such race, specialty unit, length of stay variables.

Hospital charge is primarily a function of length of stay, but it is substantially different across regions and is also affected by many other factors:

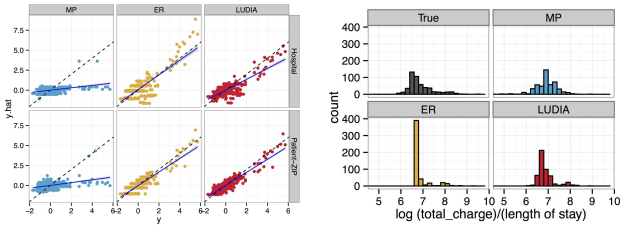
$$\text{HC} = \beta_{\text{hospital}}\text{LoS}^{\alpha} + \text{unit} + \text{severity} + \ldots + \text{error}$$

where HC and LoS represent Hospital Charge and Length of Stay, respectively. Note that the coefficient for LoS is indexed by hospital, since daily charge rate is a function of hospital. The distribution of HC is, in fact, similar to a log-normal distribution. It is a better practice to log-transform the data, before applying a linear model:

$$\log \text{HC} = \log \beta_{\text{hospital}} + \alpha \log \text{LoS} + \ldots + \text{Error}'$$

Table 3: Reconstruction Accuracy of the Texas dataset

Level	Model	MAE	MSE
County	MR	0.648 (± 0.75)	0.976 (± 3.28)
	ER	**0.466** (± 0.45)	**0.422** (± 0.87)
	LUDIA	0.514 (± 0.48)	0.497 (± 1.14)
Hospital	MR	0.609 (± 0.69)	0.851 (± 2.92)
	ER	0.513 (± 0.49)	0.501 (± 1.10)
	LUDIA	**0.435** (± 0.40)	**0.348** (± 0.68)
Patient ZIP	MR	0.589 (± 0.69)	0.824 (± 2.92)
	ER	0.319 (± 0.28)	0.184 (± 0.38)
	LUDIA	**0.289** (± 0.26)	**0.152** (± 0.34)



(a) (b)

Figure 5: (a) Reconstructed vs. original for the 3 models. (b) Estimated histograms of daily hospital charges. LUDIA histogram is the closest to the original.

This log-transformed linear model turns out to be a simple random intercept model.

Table 3 shows the reconstruction errors from three different levels of aggregation. Except for the county-level case, the LUDIA-reconstructed values are the closest to the original values with smallest variances. ER performs slightly better than LUDIA for the county-level aggregate data. This is because the multi-level effects at county-level are not distinctive enough i.e. the constancy assumption can be applied. Figure 5a illustrates the reconstructed values compared to the original values. If reconstruction is perfect, points should lie on the dotted diagonal lines. As can be seen, the MP reconstructions do not capture the tails. This is because, when the HC values are averaged, those tail values are typically cancelled out, and MP cannot infer beyond the provided average statistics. The ER reconstructions perform reasonably well, but does not capture the multi-level bias. LUDIA provides better estimates for the original values in terms of Mean Absolute Error (MAE).

The advantages of LUDIA are even more highlighted when calculating non-linear estimates. As an illustrative example, suppose that we want to estimate average daily charges. To calculate this value, we first need to reconstruct individual-level hospital charges, and then divide the reconstructed charges by the individual-level length of stay variable. In other words, average daily charges are calculated as follows:

$$\text{Average Daily Charge} = \frac{1}{n}\sum_i \frac{\hat{\text{HC}}_i}{\text{LoS}_i}$$

Figure 5b show the histograms of the estimated average daily charges. As can be seen, the histogram from LUDIA captures the asymmetrical shape of the original histogram.

As shown in Section 4, multi-level modeling can be directly integrated with LUDIA. We extract rural counties of

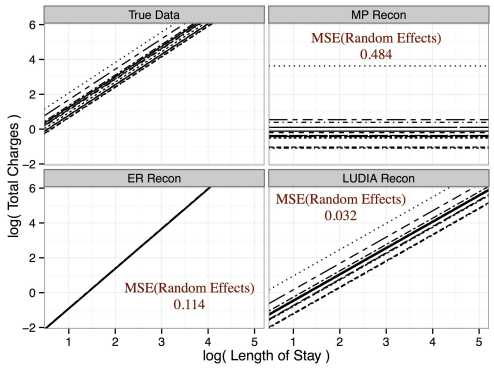


Figure 6: Multi-level modeling and the mean squared errors, shown as "MSE(Random Effects)", between the original random effects and estimated random effects. LUDIA's random effects are almost the same as the original.

Central Texas, and compare the hospital charges by applying a random intercept model. Figure 6 shows the fitted lines from the multi-level models. As can be seen, the original data clearly show the random intercept terms. It was impossible to estimate the slope term from the MP reconstructed values. For the ER reconstructed values, although the global model was similar to the original data, we cannot visually check the random intercepts. This is because ER ignores the information from the aggregation matrix. On the other hand, LUDIA provides almost the exact same random effect coefficients.

Reconstructed values from aggregate data can be used in various data mining applications. In this paper, we show a simple predictive analysis when a target column is provided in an aggregate form. By reconstructing the individual-level target values from the aggregate data, we can train a model, and then apply the model to test data as follows:

1. Combine the aggregate and individual-level data, then reconstruct the masked column
2. Train a predictive model using the pseudo complete data
3. For new data points, predict the target values using the trained model

We first divided the Texas inpatient dataset into a training (80%) and a hold-out (test, 20%) set. Assuming the total charges (target) are provided in only an aggregate form, we reconstruct the target using three different algorithms. We trained a Lasso regression model, and then measured the predictive accuracies of the target. Figure 7 shows the results from the test set. As can be seen, the LUDIA-reconstructed training dataset provides the best Lasso model in terms of MAEs. In this example, we included the performance of a model that is trained on CUDIA-reconstructed data. The CUDIA-reconstructed dataset provides better predictive accuracies than the MP- and ER-reconstructed training datasets. However, CUDIA is still a clustering algorithm, and the reconstruction from CUDIA is based on estimated cluster centers. Although CUDIA provides homogenous cluster centers, it does not generate fine-grained reconstruction like LUDIA. The predictive Lasso model trained on the CUDIA-reconstructed dataset exhibits higher MAE and

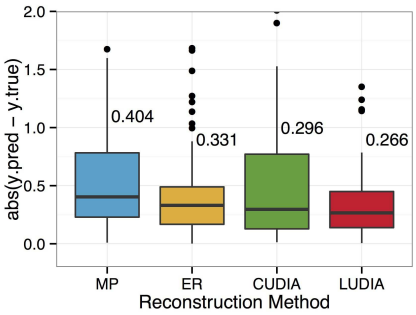


Figure 7: Predictive performance of the Lasso (`glmnet`) models trained on the reconstructed data. Absolute errors are measured using a hold-out dataset.

variances than the model trained on the LUDIA-reconstructed dataset.

7. DISCUSSION

The implication of our research can be viewed from two perspectives.

Utility perspective. Our method allows aggregated data to be effectively utilized in individual-level inferential tasks. This is particularly important since standard imputation techniques do not make use of the summary statistics provided by aggregated data that are widely available for social good. Many machine learning algorithms that require completely observed data can now be directly applied to the LUDIA-reconstructed data.

Privacy perspective. Although the reconstructed values are not guaranteed to be identical to the true values, it is clear that the estimated values are correlated with the actual values. If additional theoretical guarantees are developed, data aggregation may be no longer perfectly safe from privacy attacks. With enough auxiliary information, it is possible that private information gets revealed using techniques similar to LUDIA. This implies that, in the future, reconstruction performance will need to be considered prior to data aggregation, to guarantee that privacy requirements are met.

The proposed LUDIA framework can be extended to more complex data models. It is also worthwhile to investigate more efficient solutions for the Sylvester equation. One can also explore theoretical reconstruction guarantees that depend on the characteristics of the datasets and of the aggregation matrices.

Acknowledgements

This work is supported by NSF IIS-1016614 and by TATP grant 01829.

8. REFERENCES

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

[2] R. H. Bartels and G. W. Stewart. Solution of the matrix equation $ax + xb = c$. *Communications of the ACM*, 15(9):820–826, 1972.

[3] R. Bhatia and P. Rosenthal. How and why to solve the operater equation $ax - xb = y$. *Bulletin of the London Mathematical Society*, 29(1):1–21, 1997.

[4] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[5] E. J. Candes and B. Recht. Exact Matrix Completion via Convex Optimization. *Foundations of Computational Mathematics*, 2008.

[6] K. Carroll. Experimental evidence of dietary factors and hormone-dependent cancers. *Cancer Research*, 35:3374–3383, 1975.

[7] Centers for Disease Control and Prevension (CDC). Data and statistics. `http://www.cdc.gov/datastatistics/`, 2014.

[8] T. Dalenius and S. P. Reiss. Data-swapping: A technique for disclosure control (exteded abstract). In *Proceedings of the Section on Survey Research Methods*, 1978.

[9] Data.CMS.gov. Inpatient prospective payment system. `https://data.cms.gov/Medicare/Inpatient-Prospective-Payment-System-IPPS-Provider/97k6-zzx3`, 2014.

[10] G. T. Duncan, M. Elliot, and J.-J. Salazar-Gonzalez. *Statistical Confidentiality: Principles and Practice*. Springer, 2011.

[11] O. D. Duncan and B. Davis. An alternative to ecological correlation. *American Sociological Review*, 18:665–666, 1953.

[12] C. Eckart and G. Young. The approximation of one matrix by another lower rank. *Psychometrika*, 1936.

[13] S. E. Fienberg and J. McIntyre. Data swapping: Variations on a theme by Dalenius and Reiss. *Journal of Official Statistics*, 21:309–323, 2005.

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

[15] D. A. Freedman, S. P. Klein, M. Ostland, and M. Roberts. On "solutions" to the ecological inference probelm. *Journal of the American Statistical Association*, 93:1518–22, 1999.

[16] D. A. Freedman, S. P. Klein, J. Sacks, C. A. Smyth, and C. G. Everett. Ecological regression and voting rights. *Evaluation Review*, (673-816), 15.

[17] W. A. Fuller. Masking procedures for microdata disclosure limitation. *Journal of Official Statistics*, 9(2):383–406, 1993.

[18] A. Gelman and J. Hill. *Data Analysis using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2007.

[19] H. Goldstein. *Multilevel Statistical Models*. Wiley, 4th edition, 2010.

[20] L. Goodman. Ecological regression and the behavior of individuals. *American Sociological Review*, 18:663–664, 1953.

[21] L. Goodman. Some alternatives to ecological correlation. *American Journal of Socialogy*, 64:610–625, 1959.

[22] W. He, X. Liu, H. Nguyen, K. Nahrstedt, and T. Abdelzaher. PDA: Privacy-preserving data aggregation in wireless sensor netoworks. *IEEE International Conference on Computer Communications*, pages 2045–2053, 2007.

[23] C. R. Johnson. Matrix completion problems: a survey. *Proceedings of Symposia in Applied Mathematics*, 1990.

[24] Kaiser Family Foundation. Hospital adjusted expenses per inpatient day. `http://kff.org/other/state-indicator/expenses-per-inpatient-day/`, 2011.

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

[26] G. King, O. Rosen, and M. A. Tanner. Binomial-beta hierarchical models for ecological inference. *Sociological Methods and Research*, 28:61–90, 1999.

[27] G. King, O. Rosen, and M. A. Tanner. *Ecological Inference*. Cambridge University Press, 2004.

[28] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramanian. l-diversity: Privacy beyond k-anonymity. *Transactions on Knowledge Discovery from Data*, 1, 2007.

[29] C. Ordonez and Z. Chen. Horizontal aggregations in sql to prepare data sets for data mining analysis. *IEEE Transactions on Knowledge and Data Engineering*, pages 678–691, 2012.

[30] Y. Park and J. Ghosh. 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*, pages 27–32, 2011.

[31] Y. Park and J. Ghosh. Cudia: Probabilistic cross-level imputation using individual auxiliary information. *ACM Transactions on Intelligent Systems and Technology*, 2012.

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

[33] President's Concil of Advisors on Science and Technology. Report to the president realizing the full potential of health information technology to improve healthcare for americans: the path forward. Technical report, Office of Science and Technology Policy, the White House, December 2010.

[34] J. Richmond. Aggregation and identification. *International Economic Review*, 17:47–56, 1976.

[35] W. S. Robinson. Ecological correlations and the behavior of individuals. *American Sociological Review*, 15:351–357, 1950.

[36] L. Sweeney. k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10:557–570, October 2002.

[37] M. Templ. Statistical disclosure control for microdata using the r-package sdcMicro. *Transactions on Data Privacy*, pages 67–85, 2008.

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