# Matrix Factorization With Aggregated Observations

Yoshifumi Aimoto and Hisashi Kashima

Department of Mathematical Informatics, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
{Yoshifumi_Aimoto,Kashima}@mist.i.u-tokyo.ac.jp

**Abstract.** Missing value estimation is a fundamental task in machine learning and data mining. It is not only used as a preprocessing step in data analysis, but also serves important purposes such as recommendation. Matrix factorization with low-rank assumption is a basic tool for missing value estimation. However, existing matrix factorization methods cannot be applied directly to such cases where some parts of the data are observed as aggregated values of several features in high-level categories. In this paper, we propose a new problem of restoring original micro observations from aggregated observations, and we give formulations and efficient solutions to the problem by extending the ordinary matrix factorization model. Experiments using synthetic and real data sets show that the proposed method outperforms several baseline methods.

## 1 Introduction

In many real data analysis applications, we often face datasets with missing values due to various reasons such as sensor failures and biased sampling. Since most of the existing data analysis methods are not directly applicable to them, we first need to estimate the missing values before analysis, or we need to develop new methods that can handle data with missing values. With its ubiquitous needs, missing value estimation [9,1,12] has been placed as one of the fundamental tasks in the field of machine learning and data mining, and it has been studied extensively. A typical dataset looks can be represented as a table with missing values (see Table 1). The table shows the numbers of beers of various brands purchased by four customers, where missing values are indicated by "-". The

**Table 1.** An example of purchase data. Typical data are given as a matrix-shaped table. The table shows the numbers of beers of various brands purchased by four customers, and missing values are indicated with "-".

| items\users | Alice | Bob | Carol | Dave |
|---|---|---|---|---|
| Budweiser | 5 | - | 2 | 3 |
| Heineken | 1 | 3 | 2 | - |
| Carlsberg | 1 | - | 1 | 2 |
| Miller | 3 | 1 | 3 | 2 |

J. Pei et al. (Eds.): PAKDD 2013, Part II, LNAI 7819, pp. 521–532, 2013.

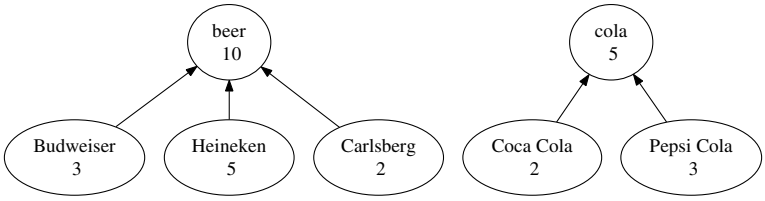


**Fig. 1.** Some portion of micro-level purchase data (e.g., the number of purchased bottles of a particular beer brand) are observed in an aggregated category (e.g., "beer")

table-structured data is mathematically considered as a matrix; hence, matrix analysis techniques are useful for missing value estimation. Matrix factorization (MF), which decomposes matrices by using the low- rank assumption [3,4], is one of the effective approaches to restore missing values in such matrix-shaped data. Missing value estimation using low-rank matrix factorization does not arise only as a preprocessing step, but also as a primary purpose of the analysis. Typical examples include recommender systems [6] and relational learning [10].

In this study, we consider a more complex situation where some parts of data are not completely missing, but are observed at a more abstract category level as aggregated values. Figure 1 shows examples of such cases. In each category (such as "beer" and "cola"), several micro-level counts (such as "Budweiser" and "Heineken") belonging to the category are observed as an aggregated count. To address such situations, we introduce a new variant of the missing value estimation problem, which we call *restoration of micro-level observations from aggregated observation*, where some parts of data are observed as aggregated values of several features. Since the existing techniques for missing value estimation including matrix factorization cannot be applied directly to such cases, we extend the existing low-rank matrix factorization formulation for missing value estimation to our case. We also devise iterative algorithms for solving the optimization problems, where each step consists of the standard singular value decomposition or closed form updates. Finally, using synthetic and real datasets, we show some experimental results on micro-observation restoration, which demonstrates that the proposed approach performs better than baseline methods.

The remainder of this paper is organized as follows. In Section 2, we introduce the *restoration of micro-level observations from aggregated observation* with a motivating example of purchase data analysis. We formulate matrix factorization problems with aggregated observations in Section 3, and give an efficient algorithm to solve the optimization problems in Section 4. In Section 5, we demonstrate the advantage of our approach over baseline approaches. Section 6 summarizes the related work, and Section 7 concludes the paper.

## 2 Problem Definition

In this section, we introduce a new problem that we refer to as the *restoration of micro observations from aggregated observations* using a motivating example

**Table 2.** An example of purchase data with aggregated observations. In addition to micro-level observations $\boldsymbol{Z}$, we have aggregated observations $\boldsymbol{Y}$, whose allocations to micro-level observations are not known. We have to restore the micro-level observation $\boldsymbol{X}$ from $\boldsymbol{Y}$ to obtain $\boldsymbol{Z} + \boldsymbol{X}$ which is the true sales data.

| micro-level observations $\boldsymbol{Z}$ | | | | | aggregated observations $\boldsymbol{Y}$ | | | | |
|---|---|---|---|---|---|---|---|---|---|
| items\users | Alice | Bob | Carol | Dave | items\users | Alice | Bob | Carol | Dave |
| Budweiser | 5 | 0 | 2 | 3 | | | | | |
| Heineken | 1 | 3 | 2 | 1 | beer | 3 | 3 | 4 | 1 |
| Carlsberg | 1 | 0 | 1 | 2 | (aggregated) | | | | |
| Miller | 3 | 1 | 3 | 2 | | | | | |

of purchase data. We consider two cases that differ according to the assumption we make on categorical structures.

### 2.1 Motivating Example

Let us assume that we have purchase data represented as a matrix $\boldsymbol{X}$, where each column corresponds to a customer, each row corresponds to a product (such as a particular brand of beer), and element $X_{ij}$ indicates the number of $i$-th products the $j$-th customer purchased. In many cases, the data has many missing values; for some $(i, j)$, $X_{ij}$ is completely missing, or a part of $X_{ij}$ is missing (for example, only five of eight actual purchases are recorded). Restoration of the true purchases is quite important in sales management and analysis, and various missing value imputation methods [5] are employed for the purpose.

Let us now imagine a more complex situation where a part of $X_{ij}$ is not missing, but is observed at a more abstract category level. In each category, several micro-level counts belonging to the category are observed as an aggregated count. For example, among eight actual purchases of the "Budweiser" brand of beer, only five are observed at the micro level (as five purchases of Budweiser), and the other three are observed in the more abstract "beer" category. The "beer" category might have ten purchases, including other beer brands such as five "Heineken" bottles and two "Carlsberg" bottles (See Figure 1). Now our goal is to restore the original micro level purchases (such as ten Budweiser purchases) from the aggregated observations.

### 2.2 Restoration of Aggregated Observations

**General Problem Definition.** Let us assume that we have two data matrices $\boldsymbol{Z}$ and $\boldsymbol{Y}$. $\boldsymbol{Z}$ is an $I \times J$ matrix that represents micro-level observations. In the previous example, $I$ is the number of product brands, and $J$ is the number of customers. $\boldsymbol{Y}$ is an $L \times J$ matrix which represents category-level observations, where $L$ indicates the number of categories. An example with purchase data is given in Table 2. In addition to $\boldsymbol{Z}$ and $\boldsymbol{Y}$, we also have a correspondence matrix $\boldsymbol{C}$ as side information about product-category relationships. $\boldsymbol{C}$ is an $L \times I$ binary

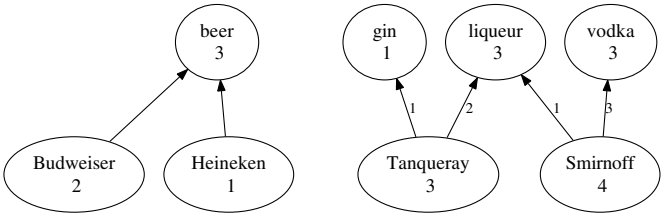


**Fig. 2.** (left) Case 1: each micro-level dimension belong to at most one category. (right) Case 2: each dimension can belong to more than one category.

matrix, whose $(\ell, i)$-th element is 1 if the $i$-th product is included in the $\ell$-th category. Our goal is to restore the hidden micro-level observation matrix $\boldsymbol{X}$ (of size $I \times J$) from $\boldsymbol{Y}$ with the help of $\boldsymbol{C}$ and $\boldsymbol{Z}$.

**Two Different Assumptions on Product-Category Relationships.** In our problem setting, we consider two different assumptions on the correspondence matrix $\boldsymbol{C}$, which results in slightly different formulations of the problem.

The first case is when each dimension of column vectors belongs to only one category (Figure 2 (left)), and the other case is when each dimension can belong to more than one category (Figure 2 (right)). Figure 2 (right) shows that a micro-level product "Tanqueray" belongs to two possible categories "gin" and "liqueur", and another micro-level product "Smirnoff" belongs to both "vodka" and "liqueur". We denote the former case as Case 1, and the latter as Case 2. The difference between the two cases is reflected by the definition of the correspondence matrix $\boldsymbol{C}$. In Case 1, each column of $\boldsymbol{C}$ has at most one value as "1" value and the rest are "0". On the other hand, in Case 2, each column of $\boldsymbol{C}$ can have multiple values as 1.

## 3 Formulation

In this section, we formulate our problem as optimization problems, where we restore micro observations $\boldsymbol{X}$ from aggregated observations $\boldsymbol{Y}$. Our model is an extension of the matrix factorization approach for missing value estimation.

### 3.1 Matrix Factorization Approach for Missing Value Estimation

We first review the existing matrix factorization approach for missing value estimation, where the observed (micro-level) data matrix $\boldsymbol{Z}$ has missing values, i.e., $Z_{ij}$ are missing for some $(i, j)$. Let us assume an observation matrix $\boldsymbol{E}$, where $E_{ij} = 1$ if $Z_{ij}$ is observed; otherwise, $E_{ij} = 0$. To impute the missing values of the matrix, the low-rank assumption is often employed. We consider the following optimization problem of rank-$k$ approximation of the observed matrix.

$$\text{minimize}_{\boldsymbol{A}} \quad \|\boldsymbol{E} * (\boldsymbol{Z} - \boldsymbol{A})\|_{\mathrm{F}}^2 \quad \text{s.t. } \operatorname{rank}(\boldsymbol{A}) \leq k,$$

where the Frobenius norm of a matrix $\boldsymbol{X} \in \mathbb{R}^{I \times J}$ is defined as $\|\boldsymbol{X}\|_{\mathrm{F}} = \sqrt{\sum_{j=1}^{J} \sum_{i=1}^{I} X_{ij}^2}$, and $*$ indicates the element-wise product. If all of the elements of $\boldsymbol{Z}$ are observed, i.e., $E_{ij} = 1$ for $\forall(i, j)$, the optimal solution is obtained by singular value decomposition (SVD). However, since we have missing elements in $\boldsymbol{Z}$, SVD cannot be applied. Furthermore, the optimization problem is not convex, hence numerical optimization methods do not guarantee optimal solutions. Recently, instead of using the rank constraint, the trace-norm constraint is often used, because the trace-norm constraint of a matrix is a convex set (whereas the rank constraint is not) [11,2]. Using the trace-norm constraint, we can formulate the low-rank matrix approximation problem as a convex optimization problem as

$$\text{minimize}_{\boldsymbol{A}} \quad \|\boldsymbol{E} * (\boldsymbol{Z} - \boldsymbol{A})\|_{\mathrm{F}}^2 \quad \text{s.t. } \|\boldsymbol{A}\|_{\mathrm{Tr}} \leq \tau,$$

where the trace norm of a matrix $\boldsymbol{X}$ is defined as $\|\boldsymbol{X}\|_{\mathrm{Tr}} = \mathrm{Tr}(\sqrt{\boldsymbol{X}\boldsymbol{X}^\top})$.

### 3.2 Matrix Factorization with Aggregated Observations

Now we extend the previous formulation to address our problem setting. Similar to the matrix factorization problem for missing value estimation, we also employ the low-rank assumption that our micro-level observations are of low-rank. We consider two slightly different formulations for the two cases we mentioned in the previous section.

**Case 1.** When each row of the micro-observation matrix can belong to at most one category, we need that the linear constraint $\boldsymbol{C}\boldsymbol{X} = \boldsymbol{Y}$, where each column of $\boldsymbol{C}$ has at most one value as "1" and the rest are "0". For example, let us assume that John bought several bottles of beer and cola as in Figure 1, the corresponding column in the constraint $\boldsymbol{C}\boldsymbol{X} = \boldsymbol{Y}$ looks like

$$\begin{matrix} \\ \text{beer} \\ \text{cola} \end{matrix} \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 5 \\ 2 \\ 2 \\ 3 \end{bmatrix} \begin{matrix} \text{Budweiser} \\ \text{Heineken} \\ \text{Carlsberg} \\ \text{Coca Cola} \\ \text{Pepsi Cola} \end{matrix} = \begin{bmatrix} 10 \\ 5 \end{bmatrix} \begin{matrix} \text{beer} \\ \text{cola} \end{matrix}$$

With the constraint $\boldsymbol{C}\boldsymbol{X} = \boldsymbol{Y}$, we formulate the optimization problem as follows.

$$\begin{aligned} \text{minimize}_{\boldsymbol{A},\boldsymbol{X}} \quad & \|\boldsymbol{A} - (\boldsymbol{X} + \boldsymbol{Z})\|_{\mathrm{F}}^2 \\ \text{s.t.} \quad & \|\boldsymbol{A}\|_{\mathrm{Tr}} \leq \tau, \quad \boldsymbol{C}\boldsymbol{X} = \boldsymbol{Y} \end{aligned} \tag{1}$$

Note that we assume that the "true" micro-observations $\boldsymbol{X} + \boldsymbol{Z}$ are of low-rank.

**Case 2.** When each row of the micro-observation matrix can belong to more than one category, aggregation from micro-level observations to category-level observations is not unique; therefore, we divide the micro-level observation matrix $\boldsymbol{X}$ into a sum of multiple matrices $\{\boldsymbol{X}^{(\ell)}\}_{\ell=1}^{L}$ so that $\sum_{\ell=1}^{L} \boldsymbol{X}^{(\ell)} = \boldsymbol{X}$ is satisfied.

In this case, we need that the linear constraints

$$\boldsymbol{C}_{\ell:}\boldsymbol{X}^{(\ell)} = \boldsymbol{Y}_{\ell:} \text{ for } \ell = 1, 2, \ldots, L \tag{2}$$

are satisfied. Note that one constraint is made for each of the $L$ categories. If John bought several bottles of alcoholic beverage as in Figure 2 (right), one column in the constraint (2) looks like

$$\text{liqueur} \begin{bmatrix} 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 2 \\ 1 \end{bmatrix} \begin{matrix} \text{Budweiser} \\ \text{Heineken} \\ \text{Tanqueray} \\ \text{Smirnoff} \end{matrix} = \begin{bmatrix} 3 \end{bmatrix} \text{liqueur}.$$

The optimization problem is defined as follows.

$$\begin{aligned} \text{minimize}_{\boldsymbol{A},\boldsymbol{X}} \quad & \|\boldsymbol{A} - (\boldsymbol{X} + \boldsymbol{Z})\|_{\mathrm{F}}^2 \\ \text{s.t.} \quad & \|\boldsymbol{A}\|_{\mathrm{Tr}} \leq \tau \\ & \boldsymbol{C}_{\ell:}\boldsymbol{X}^{(\ell)} = \boldsymbol{Y}_{\ell:} \text{ for } \ell = 1, \ldots, L, \quad \sum_{\ell=1}^{L} \boldsymbol{X}^{(\ell)} = \boldsymbol{X} \end{aligned} \tag{3}$$

Table 3 summarizes the ordinary formulation of matrix factorization, our formulation for Case 1, and one for Case 2.

## 4 Algorithms

Our optimization problems (1) and (3) are minimization problems of convex functions with respect to both $\boldsymbol{A}$ and $\boldsymbol{X}$. However, the number of variables involved is large, and it is time-consuming to minimize the objective functions with respect to them at once. Therefore, we devise iterative optimization procedures, each of whose step optimizes either of $\boldsymbol{A}$ and $\boldsymbol{X}$. We elaborate the concrete implementations of the estimation steps for both Case 1 and Case 2 below.

### 4.1 Case 1

Our proposed optimization procedure for Case 1 starts with initializing $\boldsymbol{X}$ so that the current $\boldsymbol{X}$ satisfies $\boldsymbol{C}\boldsymbol{X} = \boldsymbol{Y}$. The initialization is discussed in the Experiments section in detail. Then, we iterate the following updates of $\boldsymbol{A}$ and $\boldsymbol{X}$ until convergence.

When we update $\boldsymbol{A}$, we need to solve the optimization problem

$$\boldsymbol{A}^{\mathrm{NEW}} = \operatorname{argmin}_{\boldsymbol{A}} \|\boldsymbol{A} - (\boldsymbol{X} + \boldsymbol{Z})\|_{\mathrm{F}}^2 \text{ s.t. } \|\boldsymbol{A}\|_{\mathrm{Tr}} \leq \tau. \tag{4}$$

**Table 3.** Comparison of the formulation of the ordinary formulation of matrix factorization for missing value imputation and our formulations of restoration of micro observations from aggregated observations (Case 1 and Case 2). The constraints $\boldsymbol{X} \geq \boldsymbol{0}$ and $\boldsymbol{X}^{(\ell)} \geq \boldsymbol{0}$ are the additional non-negativity constraints we employ in Section 4.3.

| | The existing MF | Proposed MF (Case 1) | Proposed MF (Case 2) |
|---|---|---|---|
| **Inputs** | $\boldsymbol{Z} \in \mathbb{R}^{I \times J}, \tau \in \mathbb{R}^{+}$ | $\boldsymbol{Z} \in \mathbb{R}^{I \times J}, \tau \in \mathbb{R}^{+}, \boldsymbol{C} \in \mathbb{R}^{L \times I}, \boldsymbol{Y} \in \mathbb{R}^{L \times J}$ | |
| **Outputs** | $\boldsymbol{A} \in \mathbb{R}^{I \times J}$ | $\boldsymbol{A}$ , $\boldsymbol{X} \in \mathbb{R}^{I \times J}$ | |
| **Objective function** | $\|\boldsymbol{E} * (\boldsymbol{A} - \boldsymbol{Z})\|_{\mathrm{F}}^2$ w.r.t. $\boldsymbol{A}$ | $\|\boldsymbol{A} - (\boldsymbol{X} + \boldsymbol{Z})\|_{\mathrm{F}}^2$ w.r.t. $\boldsymbol{A}, \boldsymbol{X}$ | |
| **Constraints** | $\|\boldsymbol{A}\|_{\mathrm{Tr}} \leq \tau$ | $\|\boldsymbol{A}\|_{\mathrm{Tr}} \leq \tau$ $\boldsymbol{C}\boldsymbol{X} = \boldsymbol{Y}$ $(\boldsymbol{X} \geq \boldsymbol{0})$ | $\|\boldsymbol{A}\|_{\mathrm{Tr}} \leq \tau$ $\boldsymbol{C}_{\ell:}\boldsymbol{X}^{(\ell)} = \boldsymbol{Y}_{\ell:}$ $\sum_{\ell=1}^{L} \boldsymbol{X}^{(\ell)} = \boldsymbol{X}$ $(\boldsymbol{X}^{(\ell)} \geq \boldsymbol{0})$ |

This optimization problem can be solved by applying SVD to $\boldsymbol{X} + \boldsymbol{Z}$ and thresholding the singular values. Let the SVD of $\boldsymbol{X} + \boldsymbol{Z}$ be $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\top}$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices, and $\boldsymbol{\Sigma}$ is a diagonal matrix with the singular values as its diagonals. We eliminate the singular values less than the threshold $\tau$, and denote $\boldsymbol{\Sigma}'$ as the diagonal matrix with diagonal elements greater than or equal to $\tau$. The optimal solution $\boldsymbol{A}^{\mathrm{NEW}}$ of Eq. (4) is obtained as

$$\boldsymbol{A}^{\mathrm{NEW}} = \boldsymbol{U}\boldsymbol{\Sigma}'\boldsymbol{V}^{\top}. \tag{5}$$

Optimization with respect to $\boldsymbol{X}$ is casted as the minimization problem

$$\boldsymbol{X}^{\mathrm{NEW}} = \ \mathrm{argmin}_{\boldsymbol{X}} \|\boldsymbol{X} - (\boldsymbol{A} - \boldsymbol{Z})\|_{\mathrm{F}}^2 \ \text{ s.t. } \boldsymbol{C}\boldsymbol{X} = \boldsymbol{Y}. \tag{6}$$

This is generally a convex quadratic programming problem; however, the optimal solution is given in a simple closed form in this case. Since this problem can be seen as minimization of the Euclidean distance between $\boldsymbol{X}$ and $\boldsymbol{M}$ with the hyper-plane constraint $\boldsymbol{C}\boldsymbol{X} = \boldsymbol{Y}$, the optimal solution is given as the projection of $\boldsymbol{M}$ onto the hyper-plane. Assuming that the micro-level feature $i$ belongs to the aggregated category $\ell$, the optimal solution of $i$-th row $\boldsymbol{X}_{i:}^{\mathrm{NEW}}$ is obtained as

$$\boldsymbol{X}_{i:}^{\mathrm{NEW}} = \boldsymbol{M}_{i:} - \frac{1}{\sum_{i=1}^{I} C_{\ell i}} (\boldsymbol{C}_{\ell:}\boldsymbol{M} - \boldsymbol{Y}_{\ell:}), \tag{7}$$

where we defined $\boldsymbol{M} = \boldsymbol{A} - \boldsymbol{Z}$.

### 4.2 Case 2

In Case 2, noting that $\boldsymbol{X} = \sum_{\ell=1}^{L} \boldsymbol{X}^{(\ell)}$, the update of $\boldsymbol{A}$ is the same as that for Case 1. However, in contrast to Case 1, the update of $\boldsymbol{X}$ cannot be given in a

closed form solution anymore in Case 2, and we have to solve the optimization problem

$$\boldsymbol{X}^{\mathrm{NEW}} = \mathrm{argmin}_{\boldsymbol{X}} \ \|\boldsymbol{X} - (\boldsymbol{A} - \boldsymbol{Z})\|_{\mathrm{F}}^2 \qquad (8)$$
$$\text{s.t. } \boldsymbol{C}_{\ell:}\boldsymbol{X}^{(\ell)} = \boldsymbol{Y}_{\ell:} \quad (\ell = 1, \ldots, L), \quad \boldsymbol{X} = \sum_{\ell=1}^{L} \boldsymbol{X}^{(\ell)}.$$

Although it is a quadratic programming problem, the number of variables involved is rather large; hence, we again resort to iterative optimization, that is, we iterate updates with respect to one of $\{\boldsymbol{X}^{(\ell)}\}_{\ell=1}^{L}$ at once. The optimization problem with respect to only $\boldsymbol{X}^{(\ell)}$ with the other $\{\boldsymbol{X}^{(j)}\}_{j\neq\ell}$ fixed, the problem (8) is written as

$$\boldsymbol{X}^{(\ell)\mathrm{NEW}} = \mathrm{argmin}_{\boldsymbol{X}^{(\ell)}} \ \|\boldsymbol{X}^{(\ell)} - (\boldsymbol{M} - \sum_{j\neq\ell} \boldsymbol{X}^{(j)})\|_{\mathrm{F}}^2 \ \text{ s.t. } \boldsymbol{C}_{\ell:}\boldsymbol{X}^{(\ell)} = \boldsymbol{Y}_{\ell:}.$$

This has the same form as that in Case 1; hence, the closed form update becomes

$$\boldsymbol{X}_{i:}^{(\ell)^{\mathrm{NEW}}} = (\boldsymbol{M}_{i:} - \sum_{j\neq\ell} \boldsymbol{X}^{(j)}) - \frac{1}{\sum_{i=1}^{I} C_{\ell i}}(\boldsymbol{C}_{\ell:}(\boldsymbol{M} - \sum_{i\neq j} \boldsymbol{X}^{(j)}) - \boldsymbol{Y}_{\ell:}). \qquad (9)$$

### 4.3 Non-negativity Constraints

Since our original motivation came from the purchase data example, it is sometimes more reasonable to make a non-negativity assumption on the micro-level observations.

In Case 1, we make an additional constraint that $\boldsymbol{X}$ is non-negative. The resultant optimization problem for Case 1 with respect to $\boldsymbol{X}$ becomes

$$\boldsymbol{X}^{\mathrm{NEW}} = \ \mathrm{argmin}_{\boldsymbol{X}} \ \|\boldsymbol{X} - \boldsymbol{M}\|_{\mathrm{F}}^2 \ \text{ s.t. } \boldsymbol{C}\boldsymbol{X} = \boldsymbol{Y}, \quad \boldsymbol{X} \geq 0.$$

Accordingly, the previous closed form solution (7) is modified to

$$X_{ij}^{\mathrm{NEW}} = \frac{(X_{ij} - s_j)Y_{\ell j}}{(Y_{\ell j} - \sum_{i=1}^{I} C_{\ell i} s_j)}, \qquad (10)$$

where $s_j = \min_{1\leq i\leq I} X_{ij}$.

In Case 2, we make the assumption that each $\boldsymbol{X}^{(\ell)}$ is non-negative. The update (10) is similarly obtained as

$$X_{ij}^{(\ell)^{\mathrm{NEW}}} = \frac{(X_{ij} - s_j)Y_{\ell j}}{Y_{\ell j} - \sum_{i=1}^{I} C_{\ell i} s_j}.$$

Note that the modified optimization problems are still convex; therefore, we obtain optimal solutions when converged.

## 5 Experiments

We show some experimental results using synthetic and real datasets that demonstrate the reasonable performance of the proposed methods to restore micro-level observations from category-level observations. We compare the restoration errors by the proposed methods with those by four baseline methods, and show the advantage of the proposed methods over them.

### 5.1 Datasets

**Synthetic Dataset.** The first dataset is a set of randomly generated matrices. The size of matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ is $1,000 \times 5$, and each element $U_{ir} \in \{0, 1, 2, 3\}$ and $V_{jr} \in \{0, 1, 2\}$ is generated uniformly at random over the ranges. The true micro-level observation matrix is generated as $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{V}^\top$, where 5% of the elements of $\boldsymbol{A}$ are randomly missing.

A $100 \times 1,000$ correspondence matrix $\boldsymbol{C}$ is generated so that the $(\ell, i)$-th element is 1 if and only if $i$ is in $\{10 \times (\ell - 1) + 1, \ldots, 10 \times \ell\}$ for Case 1. For Case 2, starting from the $\boldsymbol{C}$ we created for Case 1, we further sample 300 $(\ell, i)$ pairs to make additional "1" values. To create category-level observations, we employ binomial distributions to divide the true micro-level observations $\boldsymbol{A}$ into the hidden part $\boldsymbol{X}$ and the observed part $\boldsymbol{Z}$. Namely, each element $Z_{ij}$ is determined by $\Pr(Z_{ij} = k) = \binom{A_{ij}}{k} p^k (1-p)^{A_{ij}-k}$, where $p$ controls the likeliness of the observation of each micro-observation at its superordinate category. For example, $p = 1$ corresponds to the perfect observation case with no category-level observations. In our experiments, we varied $p$ in $\{0.1, 0.4, 0.7\}$.

Once the hidden part $\boldsymbol{X}$ is determined, the corresponding category-level observations $\boldsymbol{Y}$ are created using the correspondence relationship $\boldsymbol{C}\boldsymbol{X} = \boldsymbol{Y}$ for Case 1. For Case 2, we set $\boldsymbol{X}^{(\ell)}_{i:} = \boldsymbol{X}_{i:} / \sum_{i=1}^{I} C_{\ell i}$ if $C_{\ell i} = 1$, and aggregate $\boldsymbol{X}^{(\ell)}_{i:}$ to $\boldsymbol{Y}_{\ell:}$ with $\boldsymbol{C}_{\ell:}\boldsymbol{X}^{(\ell)} = \boldsymbol{Y}_{\ell:}$.

**Purchase Dataset for Internet Stores.** Another dataset is a real cross-store purchase dataset collected from 6 internet stores to include for 494 customers, and 150 product brands belonging to 11 categories (such as electronic devices, undergarments, and magazines). Since the granularities of the input sales logs differ from store to store, not all of them provided detailed product names, and gave only category-level information. One product can belong to more than one category in this dataset; hence, this dataset belongs to Case 2. Since we had no ground truth micro-observations, we simulated category-level observations again from the micro-observed data; we assumed that some stores did not provide micro-level sales, and their sales were given as category-level observations.

### 5.2 Comparison Methods

Although there have not been any existing methods that address the restoration problem to the best of our knowledge, we consider four baseline methods as com-

parison methods to evaluate the proposed methods. The first method (that we call "Equal" method) divides each category-level observation into its descendant micro-observation equally. The second method (that we call "Prop" method) divides the category-level observations in proportion to the observed micro-level values (of $\boldsymbol{Z}$) as $X_{ij} = Z_{ij}Y_{\ell j}/\boldsymbol{C}_{\ell:}\boldsymbol{Z}_{:j}$ in Case 1, and $X^{(\ell)}_{ij} = Z_{ij}C_{\ell i}Y_{\ell j}/\boldsymbol{C}_{\ell:}\boldsymbol{Z}_{:j}$ in Case 2. In addition, we applied the matrix factorization method (SVD) to the matrices obtained using the above methods, which results in two additional baseline methods (which we call "Equal+MF" and "Prop+MF").

The micro-level estimations obtained by the simple methods are also used for initialization of $\boldsymbol{X}$ in the proposed method. Although our formulations are convex optimization problems and the solutions do not depend on the initial estimates, our preliminary experiments suggest that initialization with the "Equal" method shows better numerical stability.

### 5.3 Results

Table 4 and 5 show comparison of errors under different methods with varied $p$ among $\{0.1, 0.4, 0.7\}$, where Table 4 shows the results for the synthetic data in Case 1, Table 5 for that in Case 2. Table 6 shows the results for the purchase dataset (in Case 2) where one, two, or four stores out of six are assumed not to provide micro-level sales. As the evaluation metric, we used the difference between the estimated micro-observations $\hat{\boldsymbol{X}}$ and the true micro-observations $\boldsymbol{X}$ defined as $\mathrm{Error}_{\boldsymbol{X}}(\hat{\boldsymbol{X}}) = \|\hat{\boldsymbol{X}} - \boldsymbol{X}\|_{\mathrm{F}}/\|\boldsymbol{X}\|_{\mathrm{F}}$. The ranks for matrix factorization were determined so that the simple MF method (Equal-MF or Prop-MF) performed the best (we reused it for the proposed method); they were 20 for the synthetic dataset (Case 1) and the purchase dataset, and 36 for the synthetic dataset (Case 2). The difference of the error between each comparison method and proposed method is significant in the Wilcoxon signed-rank test at a 0.05 significance level. The results show that the proposed matrix factorization method is superior to the baseline methods. Interestingly, the simple application of matrix factorization (Equal-MF and Prop-MF) sometimes made the results worse than those by Equal and Prop. The simple MF methods roughly correspond to stopping the iterations of the proposed algorithm at the first iteration, and the results show it improved the performance after several iterations.

Finally, we mention the computational cost of the proposed method; the computational cost depends approximately on the number of calls of the SVD routine, which was about five calls to converge.

## 6 Related Work

Dealing with incomplete data has been studied extensively, and widely applied in various fields including machine learning and data mining. Zhu *et al.* [12] categorized strategies to handle missing data into three categories, that are, case deletion, learning without handling of missing data, and data imputation. Case

**Table 4.** Comparison of the micro-observation reconstruction errors by the proposed method and the baseline methods with the artificial dataset in Case 1. $p$ controls how likely each micro-observation is observed at its superordinate category. The proposed method achieves the lowest error for all $p$.

| $p$ | Equal | Prop | Equal+MF | Prop+MF | MFAO |
|---|---|---|---|---|---|
| 0.1 | 0.384 | 0.966 | 0.399 | 0.382 | **0.374** |
| 0.4 | 0.429 | 0.551 | 0.499 | 0.430 | **0.419** |
| 0.7 | 0.521 | 0.552 | 0.812 | 0.772 | **0.519** |

**Table 5.** Comparison of the micro-observation reconstruction errors by the proposed method and the baseline methods with the artificial dataset in Case 2. $p$ controls how likely each micro-observation is observed at its superordinate category. The proposed method achieves the lowest error for all $p$.

| $p$ | Equal | Prop | Equal+MF | Prop+MF | MFAO |
|---|---|---|---|---|---|
| 0.1 | 0.540 | 1.104 | 0.543 | 0.646 | **0.535** |
| 0.4 | 0.570 | 0.708 | 0.592 | 0.561 | **0.560** |
| 0.7 | 0.615 | 0.669 | 0.771 | 0.747 | **0.611** |

**Table 6.** Comparison of the micro-observation reconstruction errors by the proposed method and the baseline methods with the purchase dataset (in Case 2). We assumed that some stores (out of six stores) did not provide the micro-level sales, and their sales were given as category-level observations. The proposed method achieves the lowest error regardless of the number of stores not providing micro-level sales.

| Number of stores not providing micro-level sales | Equal | Prop | Equal+MF | Prop+MF | MFAO |
|---|---|---|---|---|---|
| 1 | **0.947** | 1.308 | 1.006 | 1.340 | **0.947** |
| 2 | 0.964 | 1.100 | 1.461 | 1.529 | **0.939** |
| 4 | 0.975 | 1.151 | 1.712 | 1.800 | **0.947** |

deletion, which ignores missing values, is the simplest method. These kinds of approaches require robust methods to counter incomplete data [9]. Methods in the second category directly work with missing data. Data imputation approaches estimate unobserved values from the observed ones, and this method includes the matrix factorization approach we employed in this research.

Missing value imputation approaches can be classified into two categories, that are, data-driven approach and model-based approach [7]. Our method is categorized into the latter, and employs the matrix factorization model. There are several studies to impute missing values using matrix factorization techniques such as SVD and non-negative matrix factorization [8].

## 7 Conclusion

Missing value estimation is an unavoidable problem in real data analysis. In this study, we introduced an extended matrix factorization for a new missing value

estimation problem, that is, restoration of micro-level observations from category-level aggregated observation. Since the existing methods cannot directly be applied to this problem, we formulated an extended low-rank matrix factorization problem, and devised efficient iterative algorithms for solving the optimization problems. The experimental results using synthetic and real datasets showed that our approach performed better than baseline methods.

**Acknowledgment.** The authors are grateful to Naonori Ueda, Hiroshi Sawada, and Katsuhiko Ishiguro of NTT Communication Science Laboratories, and Noriko Takaya of NTT Cyber Solution Laboratory.

## References

1. Brand, M.: Incremental singular value decomposition of uncertain data with missing values. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002, Part I. LNCS, vol. 2350, pp. 707–720. Springer, Heidelberg (2002)
2. Candes, E.J., Tao, T.: The power of convex relaxation: Near-optimal matrix completion. IEEE Transactions on Information Theory 56(5), 2053–2080 (2010)
3. Eckart, C., Young, G.: The approximation of one matrix by another of lower rank. Psychometrika 1(3), 211–218 (1936)
4. Eriksson, A., Hengel, A.V.D.: Efficient computation of robust low-rank matrix approximations in the presence of missing data using the $L_1$ norm. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 771–778. IEEE, San Francisco (2010)
5. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques, 3rd edn. Morgan Kaufmann (2011)
6. Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. Computer 42(8), 30–37 (2009)
7. Lakshminarayan, K., Harp, S.A., Samad, T.: Imputation of missing data in industrial databases. Applied Intelligence 11, 259–275 (1999)
8. Lee, L., Seung, D.: Algorithms for non-negative matrix factorization. In: Advances in Neural Information Processing Systems 13, pp. 556–562 (2001)
9. Little, R.J., Rubin, D.B.: Statistical Analysis with Missing Data. Wiley (1987)
10. Singh, A.P., Gordon, G.J.: Relational learning via collective matrix factorization. In: ACM SIGKDD, Las Vegas, USA, pp. 650–658 (2008)
11. Srebro, N., Rennie, J., Jaakkola, T.: Maximum-margin matrix factorization. In: Advances in Neural Information Processing Systems 17 (2005)
12. Zhu, X., Zhang, S., Jin, Z., Zhang, Z., Xu, Z.: Missing value estimation for mixed-attribute data sets. IEEE Transactions on Knowledge and Data Engineering 23(1), 110–121 (2011)