Co-occurrence Estimation from Aggregated Data with Auxiliary Information

Tomoharu Iwata, Naoki Marumo
NTT Communication Science Laboratories, Kyoto, Japan
{tomoharu.iwata.gy, naoki.marumo.ec}@hco.ntt.co.jp

Abstract

Complete co-occurrence data are unavailable in many applications, including purchase records and medical histories, because of their high cost or privacy protection. Even with such applications, aggregated data would be available, such as the number of purchasers for each item and the number of patients with each disease. We propose a method for estimating the co-occurrence of items from aggregated data with auxiliary information. For auxiliary information, we use item features that describe the characteristics of each item. Although many methods have been proposed for estimating the co-occurrence given aggregated data, no existing method can use auxiliary information. We also use records of a small number of users. With our proposed method, we introduce latent co-occurrence variables that represent the amount of co-occurrence for each pair of items. We model a probabilistic generative process of the latent co-occurrence variables by a multinomial distribution with Dirichlet priors. The parameters of the Dirichlet priors are parameterized with neural networks that take the auxiliary information as input, where neural networks are shared across different item pairs. The shared neural networks enable us to learn unknown relationships between auxiliary information and co-occurrence using the data of multiple items. The latent co-occurrence variables and the neural network parameters are estimated by maximizing the sum of the likelihood of the latent co-occurrence variables and the likelihood of the small records. We demonstrate the effectiveness of our proposed method using user-item rating datasets.

1 Introduction

Co-occurrence is the basic and important statistics for analyzing categorical data. For example, recommender systems in e-commerce services suggest items that are likely to be purchased by the same user (Sarwar et al. 2001). With text analysis, words that appear in the same document are clustered to discover topics (Blei, Ng, and Jordan 2003). With social network analysis, communities are detected using information about common friends (Girvan and Newman 2002). Medical knowledge about complications is essential for treatment. Complete co-occurrence data are unavailable in many applications, such as purchase records and medical histories, because of their high cost or privacy protection. However, even within such applications, aggregated data would be available, such as the number of purchasers for each item and the number of patients for each disease, since they do not contain any privacy information.

In this paper, we propose a method that estimates co-occurrence from aggregated data with auxiliary information. For auxiliary information, we use item features. In the case of purchase records, item features might include genres, release dates and descriptions. Since item features do not contain privacy information, they are often open to the public. Although many methods have been proposed for estimating co-occurrence counts given aggregated data (Deming and Stephan 1940; Causey 1983; Sheldon and Dietterich 2011), no existing method can use auxiliary information. We also use the records of a few users that are sampled from total users. The sampled records can be obtained by giving incentives to some users if they agree to share their records. We assume that the number of sampled users with their records is much smaller than the total number of users.

With the proposed method, we introduce latent co-occurrence variables, which represent the amount of co-occurrence for each pair of items, or the number of users who purchased both items. The joint probability of the co-occurrence count of two items is assumed to follow a multinomial distribution with Dirichlet priors. The parameters of the Dirichlet priors are modeled by non-linear functions that take the auxiliary information as input. For non-linear functions, we use permutation invariant neural networks since co-occurrence data are invariant to the permutation of two items. The neural network parameters are shared across different pairs of items. Shared neural networks enable us to learn unknown relationships between auxiliary information and the co-occurrence using the data of multiple items.

The latent co-occurrence variables and the neural network parameters are estimated by maximizing the sum of the likelihood of the latent co-occurrence variables and the likelihood of the small sampled records. The parameters of the multinomial distributions are analytically marginalized since we use conjugate Dirichlet priors. By relaxing that the latent co-occurrence variables are non-negative real values instead of integers, we efficiently maximize the objective

function with stochastic gradient descent methods.

2 Related work

Previous work proposed learning from aggregated data, such as imputing individual level records (Park and Ghosh 2012; 2014) and regression from aggregated data (Goodman 1953; Freedman et al. 1991; Bhowmik, Ghosh, and Koyejo 2015; 2016). In this paper, we focus on estimating co-occurrence counts, which can be represented by two-by-two contingency tables. Many methods for estimating contingency tables have been proposed especially in statistics (Ireland and Kullback 1968; Smith 1947; Fienberg 1970; Slavkovic 2010), such as least squares (Deming and Stephan 1940; Stephan 1942), maximum likelihood (Causey 1983) and the Markov chain Monte Carlo (Dobra, Tebaldi, and West 2006), as well as in machine learning (Sheldon and Dietterich 2011; Kumar, Sheldon, and Srivastava 2013; Sheldon et al. 2013; Sun, Sheldon, and Kumar 2015; Nguyen et al. 2016). Most of these methods use the sampled data of a small population and/or the aggregated counts for each item. For other information, samples from different populations (Little and Wu 1991), conditional tables (Slavkovic 2010) and short and long form questionnaires (Greco 2016) have been used. However, these existing methods cannot use auxiliary item features. Although neural collective graphical models (Iwata and Shimizu 2019) estimate people flow from aggregated data using spatio-temporal auxiliary information, they are specialized for people flow, and inapplicable to co-occurrence estimation. The proposed method can be seen as a multi-task learning approach (Caruana 1997) for simultaneously multiple contingency tables by transferring knowledge across different two-by-two contingency tables using shared neural networks.

3 Proposed method

3.1 Problem formulation

For simplicity, we explain the proposed method with an example of item co-occurrence in user purchase records. However, our proposed method is applicable to any kinds of co-occurrences, such as word co-occurrence in documents and disease complications in patients.

Assume I items and U users, where each user has purchased some items but we cannot observe their complete purchase records. We are given marginal counts for each item, $\mathbf{y} = \{y_i\}_{i=1}^{I}$, where y_i is the number of users who have purchased item i. For auxiliary information, we are given item features $\mathbf{S} = \{\mathbf{s}_i\}_{i=1}^{I}$, and item purchase records $\mathbf{R} = \{\mathbf{r}_u\}_{u=1}^{U^*}$ for a small number of users $U^* \ll U$. Here, $\mathbf{s}_i \in \mathbb{R}^D$ is a D-dimensional item feature vector of item i, such as genres, release dates and descriptions, and $\mathbf{r}_u \in \{0,1\}^I$ is an I-dimensional binary vector, where $r_{ui} = 1$ if user u has purchased item i and $r_{ui} = 0$ otherwise.

Our task is to estimate the joint probability of the occurrence of two items, $\boldsymbol{\pi}_{ij} = (\pi_{\bar{i}\bar{j}}, \pi_{\bar{i}j}, \pi_{i\bar{j}}, \pi_{ij})$, for all pairs of items, $i, j \in \{1, \cdots, I\}$, where $\pi_{\bar{i}\bar{j}} = P(r_i = 0, r_j = 0)$ is the probability that a user purchases neither items i nor j, $\pi_{\bar{i}j} = P(r_i = 0, r_j = 1)$ is the probability that a user does not purchase item i but buys item j, $\pi_{i\bar{j}} = P(r_i = 1, r_j = 0)$ is the probability that a user purchases item i but not item j, $\pi_{ij} = P(r_i = 1, r_j = 1)$ is the probability that a user purchases both items i and j, $\sum_{i' \in \{i,\bar{i}\}, j' \in \{j,\bar{j}\}} \pi_{i'j'} = 1$, and $\pi_{i'j'} \geq 0$. Table 1 summarizes our notation.

Table 1: Our notation.

I	number of items
U	number of users
y_i	number of users who have purchased item i
$\mathbf{s}_i$	auxiliary feature vector of item i
$\mathbf{R}$	purchase records of a few sampled users
π_{ij}	probability that a user purchases both items i and j
$\hat{\theta}_i$	empirical probability that item i is purchased
x_{ij}	latent number of users who have purchased both items i and j
$\boldsymbol{\Psi}$	parameters of neural networks

Table 2: Co-occurrence count matrix $\mathbf{x}_{ij}$ between items i and j, number of purchasers y_i and y_j, and total number of users.

Item i \ Item j	Not purchased	Purchased	Total
Not purchased	$x_{\bar{i}\bar{j}}$	$x_{\bar{i}j}$	$U - y_i$
Purchased	$x_{i\bar{j}}$	x_{ij}	y_i
Total	$U - y_j$	y_j	U

3.2 Model

We introduce latent co-occurrence variables $\mathbf{x}_{ij} = (x_{\bar{i}\bar{j}}, x_{\bar{i}j}, x_{i\bar{j}}, x_{ij})$, where $x_{\bar{i}\bar{j}}$ is the number of users who have purchased neither items i nor j, $x_{\bar{i}j}$ is the number of users who have purchased item j but not i, $x_{i\bar{j}}$ is the number of users who have purchased item i but not j, and x_{ij} is the number of users who have purchased both items i and j. Table 2 shows a co-occurrence count matrix that represents the relationship among latent co-occurrence variables $\mathbf{x}_{ij}$, item marginal counts y_i, y_j and the total number of users U. Given x_{ij}, other latent co-occurrence variables $x_{\bar{i}\bar{j}}, x_{\bar{i}j}$ and $x_{i\bar{j}}$ are determined using y_i, y_j and U as follows,

$$x_{\bar{i}j} = y_j - x_{ij}, \quad x_{i\bar{j}} = y_i - x_{ij},$$
$$x_{\bar{i}\bar{j}} = U - y_i - y_j + x_{ij}. \tag{1}$$

Therefore, we do not need to estimate $x_{\bar{i}\bar{j}}, x_{\bar{i}j}, x_{i\bar{j}}$ if x_{ij} is estimated. Using Eq.(1) and the non-negativity of all the latent co-occurrence variables, $x_{\bar{i}\bar{j}}, x_{\bar{i}j}, x_{i\bar{j}}, x_{ij}$, the following constraint on x_{ij} is derived,

$$\max(0, y_i + y_j - U) \leq x_{ij} \leq \min(y_i, y_j), \tag{2}$$

which is known as Fréchet inequalities. Range, $\min(y_i, y_j) - \max(0, y_i + y_j - U)$, indicates information about co-occurrence x_{ij} of aggregated data y_i and y_j. When the range is narrow, the aggregated data are informative for estimating co-occurrence. However, since the range is not generally narrow, the co-occurrence cannot be determined only from the aggregated data.

We assume that the probability of latent co-occurrence variables $\mathbf{x}_{ij}$ is given by the following multinomial distribution with parameters $\boldsymbol{\pi}_{ij}$,

$$p(\mathbf{x}_{ij}|\boldsymbol{\pi}_{ij}) = \frac{U!}{\sum_{i' \in \{i,\bar{i}\}, j' \in \{j,\bar{j}\}} x_{i'j'}!} \prod_{i' \in \{i,\bar{i}\}, j' \in \{j,\bar{j}\}} \pi_{i'j'}^{x_{i'j'}}. \tag{3}$$

Note that a multinomial distribution is derived from the Poisson distributions for each cell when total count U is fixed. We assume the following Dirichlet distribution for the prior of co-occurrence probabilities $\boldsymbol{\pi}_{ij}$,

$$p(\boldsymbol{\pi}_{ij}|\boldsymbol{\beta}_{ij}) = \frac{\Gamma(\sum_{i' \in \{i,\bar{i}\}, j' \in \{j,\bar{j}\}} \beta_{i'j'})}{\prod_{i' \in \{i,\bar{i}\}, j' \in \{j,\bar{j}\}} \Gamma(\beta_{i'j'})} \times \prod_{i' \in \{i,\bar{i}\}, j' \in \{j,\bar{j}\}} \pi_{i'j'}^{\beta_{i'j'}-1}, \tag{4}$$

where $\boldsymbol{\beta}_{ij} = (\beta_{\bar{i}\bar{j}}, \beta_{\bar{i}j}, \beta_{i\bar{j}}, \beta_{ij})$, $\beta_{i'j'} > 0$, and $\Gamma(\cdot)$ is the Gamma function. We model Dirichlet parameters $\boldsymbol{\beta}_{ij}$ as a function of auxiliary information $\mathbf{s}_i$ and $\mathbf{s}_j$ to incorporate them for estimating the co-occurrence probabilities. By modeling the Dirichlet parameters, we can handle the variance of the co-occurrence probabilities as well as their mean, where the mean is $\mathbb{E}[\pi_{i'j'}] = \frac{\beta_{i'j'}}{\beta_{ij0}}$ and the variance is $\mathrm{Var}[\pi_{i'j'}] = \frac{\beta_{i'j'}(\beta_{ij0}-\beta_{i'j'})}{\beta_{ij0}^2(\beta_{ij0}+1)}$, and $\beta_{ij0} = \sum_{i'j'} \beta_{i'j'}$. In particular, we use the following neural network-based models,

$$\begin{aligned}
\beta_{\bar{i}\bar{j}} &= \alpha(1-\hat{\theta}_i)(1-\hat{\theta}_j) + f_0(\mathbf{s}_i, \mathbf{s}_j), \\
\beta_{\bar{i}j} &= \alpha(1-\hat{\theta}_i)\hat{\theta}_j + f_{01}(\mathbf{s}_i, \mathbf{s}_j), \\
\beta_{i\bar{j}} &= \alpha\hat{\theta}_i(1-\hat{\theta}_j) + f_{01}(\mathbf{s}_j, \mathbf{s}_i), \\
\beta_{ij} &= \alpha\hat{\theta}_i\hat{\theta}_j + f_1(\mathbf{s}_i, \mathbf{s}_j),
\end{aligned} \tag{5}$$

where $\hat{\theta}_i = \frac{y_i}{U}$ is the empirical marginal occurrence probability of item i, $\alpha > 0$ is the scalar positive parameter, and $f_0(\cdot)$, $f_{01}(\cdot)$, $f_1(\cdot)$ are functions modeled with neural networks. The first term in Eq.(5) corresponds to the probability when the occurrences of items i and j are independent, and α controls the strength of independence. Since a simple assumption to estimate the joint probabilities is independence when there is no information about the co-occurrence, we include this first term. The second term controls the correlation between the occurrence of two items using auxiliary information. Neural networks enable us to flexibly model the relationships between the correlation and auxiliary information. Even when items i and j are transposed, the prior probability should be invariant. Therefore, we use $f_{01}(\mathbf{s}_j, \mathbf{s}_i)$ for $\beta_{i\bar{j}}$, where input auxiliary information $\mathbf{s}_i$ and $\mathbf{s}_i$ are transposed from $f_{01}(\mathbf{s}_i, \mathbf{s}_j)$ for $\beta_{\bar{i}j}$. Also, $f_0(\cdot)$ and $f_1(\cdot)$ need to be invariant to the permutation of inputs $\mathbf{s}_i$ and $\mathbf{s}_j$,

$$f_0(\mathbf{s}_i, \mathbf{s}_j) = f_0(\mathbf{s}_j, \mathbf{s}_i), \quad f_1(\mathbf{s}_i, \mathbf{s}_j) = f_1(\mathbf{s}_j, \mathbf{s}_i). \tag{6}$$

To satisfy the above invariance, we use the following permutation invariant networks (Zaheer et al. 2017),

$$\begin{aligned}
f_0(\mathbf{s}_i, \mathbf{s}_j) &= \rho_0(\phi_0(\mathbf{s}_i) + \phi_0(\mathbf{s}_j)), \\
f_1(\mathbf{s}_i, \mathbf{s}_j) &= \rho_1(\phi_1(\mathbf{s}_i) + \phi_1(\mathbf{s}_j)),
\end{aligned} \tag{7}$$

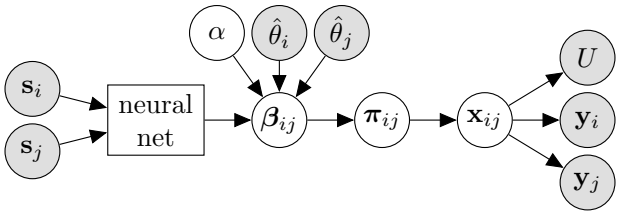


Figure 1: Proposed model: Shaded and unshaded nodes indicate observed and latent variables. Neural networks take item feature vectors $\mathbf{s}_i$ and $\mathbf{s}_j$ as input. Dirichlet parameters $\boldsymbol{\beta}_{ij}$ are determined by α, empirical marginal count probabilities $\hat{\theta}_i$ and $\hat{\theta}_j$, and neural network output. Co-occurrence probabilities $\boldsymbol{\pi}_{ij}$ are generated from Dirichlet distribution with parameters $\boldsymbol{\beta}_{ij}$, and latent co-occurrence variables $\mathbf{x}_{ij}$ are generated from multinomial distribution with parameters $\boldsymbol{\pi}_{ij}$. Total count $U = x_{\bar{i}\bar{j}} + x_{\bar{i}j} + x_{i\bar{j}} + x_{ij}$, marginal counts $y_i = x_{i\bar{j}} + x_{ij}$ and $y_j = x_{\bar{i}j} + x_{ij}$ are determined by latent co-occurrence variables $\mathbf{x}_{ij} = (x_{\bar{i}\bar{j}}, x_{\bar{i}j}, x_{i\bar{j}}, x_{ij})$.

where $\rho_0(\cdot)$, $\phi_0(\cdot)$, $\rho_1(\cdot)$, $\phi_1(\cdot)$ are neural networks. Since addition is invariant to permutation, neural networks with structure of Eq.(7) output the same values even when inputs $\mathbf{s}_i$ and $\mathbf{s}_j$ are permuted as in Eq.(6). Fig. 1 illustrates the proposed model.

3.3 Estimation

First, we count the co-occurrences in small purchase records $\mathbf{R}$, and represent them by $\mathbf{X}^* = \{\mathbf{x}_{ij}^*\}_{i,j=1}^I$, where $\mathbf{x}_{ij}^* = (x_{\bar{i}\bar{j}}^*, x_{\bar{i}j}^*, x_{i\bar{j}}^*, x_{ij}^*)$, and x_{ij}^* is the number of users with $r_{ui} = 1$ and $r_{uj} = 1$ in $\mathbf{R}$, which are the users who have purchased both items i and j in the small purchase records. Then, we estimate latent co-occurrence variables $\mathbf{X} = \{x_{ij}\}_{i,j=1}^I$, parameter α, and parameters $\boldsymbol{\Psi}$ of neural networks $f_0(\cdot)$, $f_1(\cdot)$, $f_{01}(\cdot)$ by maximizing the weighted sum of the likelihoods of $\mathbf{X}$ and $\mathbf{X}^*$.

To automatically satisfy the box constraint in Eq.(2), we parameterize x_{ij} by

$$\begin{aligned}
x_{ij} &= \max(0, y_i + y_j - U) \\
&\quad + \frac{\min(y_i, y_j) - \max(0, y_i + y_j - U)}{1 + \exp(-x'_{ij})},
\end{aligned} \tag{8}$$

where $-\infty < x'_{ij} < \infty$, and optimize x'_{ij}, instead of x_{ij}, without constraints. Also, we parameterized positive parameter α by $\alpha = \exp(\alpha')$, where $-\infty < \alpha' < \infty$, for unconstrained optimization.

Since the Dirichlet distribution is the conjugate prior of the multinomial distribution parameters, we can analytically marginalize co-occurrence probabilities $\boldsymbol{\pi}_{ij}$ in Eqs.(3) and (4). Then the probability of $\mathbf{x}_{ij}$ given $\boldsymbol{\beta}_{ij}$ is as follows,

$$p(\mathbf{x}_{ij}|\boldsymbol{\beta}_{ij}) = \frac{U!\Gamma(\sum_{i',j'} \beta_{i'j'})}{\Gamma(U + \sum_{i',j'} \beta_{i'j'})} \prod_{i',j'} \frac{\Gamma(x_{i'j'} + \beta_{i'j'})}{x_{i'j'}!\Gamma(\beta_{i'j'})}, \tag{9}$$

Algorithm 1: Estimation procedure with the proposed method.

Input: marginal counts $\mathbf{y}$, total count U, auxiliary information $\mathbf{S}$, small records $\mathbf{R}$, hyperparameter λ, batchsize T
Output: estimation of co-occurrence counts $\mathbf{X}$, neural network parameters $\mathbf{\Psi}$, parameter α

Initialize parameters $\mathbf{X}$, $\mathbf{\Psi}$, α;
Calculate empirical marginal occurrence probability $\hat{\theta}_i = \frac{y_i}{U}$ for all items using $\mathbf{y}$;
Calculate co-occurrence counts for small records $\mathbf{X}^*$ using $\mathbf{R}$;
repeat
    Sample a set of T item pairs $\mathbf{Q}$ randomly;
    Calculate the objective function Eq.(10) for the sampled item pairs and its gradients Update parameters $\mathbf{X}$, $\mathbf{\Psi}$, α with a gradient-based optimization method;
until *end condition is satisfied*;

where $i' \in \{i, \bar{i}\}$ and $j' \in \{j, \bar{j}\}$. This is called the Dirichlet compound multinomial distribution or the multivariate Pólya distribution.

The objective function to be maximized is the following weighted sum of the likelihoods of latent co-occurrence variables $\mathbf{X}$ and observed small co-occurrence $\mathbf{X}^*$,

$$L(\mathbf{X}, \mathbf{\Psi}, \alpha) = \lambda \sum_{i=1}^{I} \sum_{j=i+1}^{I} \log p(\mathbf{x}_{ij} | \boldsymbol{\beta}_{ij}) + (1-\lambda) \sum_{i=1}^{I} \sum_{j=i+1}^{I} \log p(\mathbf{x}^*_{ij} | \boldsymbol{\beta}_{ij}), \quad (10)$$

where Dirichlet parameters $\boldsymbol{\beta}_{ij}$ are calculated using neural networks by Eq.(5), λ is the hyperparameter, and $0 \leq \lambda \leq 1$. We used fixed hyperparameter $\lambda = 0.5$ in our experiments. By relaxing that the latent co-occurrence variables are non-negative real values instead of integers, we can efficiently find a local maximum of the objective function with stochastic gradient-based optimization methods. Algorithm 1 shows the estimation procedure with the proposed method.

4 Experiments

4.1 Data

We evaluated the proposed method using sushi and Movielens data sets.

The sushi data were generated from the preferences of 5,000 users among the following ten kinds of sushi: shrimp, sea eel, tuna, squid, sea urchin, salmon roe, egg, fatty tuna, tuna roll, and cucumber roll (Kamishima 2003)[1]. When user u ranked sushi item i within the top five among the ten items, we assumed that the user has purchased the item $r_{ui} = 1$, and generated occurrence data. For the auxiliary item features, we used style (maki or not), major group (seafood or not), minor group, heaviness/oiliness in taste, frequency of consumption, price, and frequency of being sold. We transformed the categorical features, i.e., style, major and minor groups, into one-hot vectors, and normalized the numerical features, i.e., heaviness, frequency, and price, into a range from zero to one. The dimensionality of the item feature vectors was $D = 20$.

Table 3: Comparing methods. Methods with check-marks indicate that they use marginal counts $\mathbf{y}$, small records $\mathbf{R}$, or auxiliary information $\mathbf{S}$.

	IND	ML	EB	CGM	EB+CGM	EB+A	CGM+A	Ours
$\mathbf{y}$	√			√	√		√	√
$\mathbf{R}$		√	√		√	√		√
$\mathbf{S}$						√	√	√

The Movielens data were generated from 100,000 ratings of 943 users of 1,682 movies (Konstan et al. 1997)[2]. When user u rated movie i, we assumed that she has purchased the item $r_{ui} = 1$. For the auxiliary item features, we used the release date and genres, such as action, comedy, and fantasy. We normalized the release dates into a range from zero to one, and transformed the genres into vectors by setting the corresponding element to one when the movie was in the genre, and zero otherwise. The dimensionality of the item feature vectors was $D = 20$.

With both data sets, we generated item count y_i for each item i using all the users. We randomly sampled the purchase records of ten users for the validation data.

4.2 Evaluation measurement

For the evaluation measurement, we used the following absolute error of the co-occurrence probabilities,

$$E = \frac{2}{I(I-1)} \sum_{i=1}^{I} \sum_{j=i+1}^{I} \sum_{i' \in \{i,\bar{i}\}, j' \in \{j,\bar{j}\}} |\hat{\pi}_{i'j'} - \pi^{\mathrm{TRUE}}_{i'j'}|, \quad (11)$$

where $\pi^{\mathrm{TRUE}}_{i'j'}$ is the true empirical co-occurrence probability calculated using all the users, which is not used during training, and $\hat{\pi}_{i'j'}$ is the estimated co-occurrence probability.

4.3 Comparing methods

We compared the proposed method with the following seven methods: IND, ML, EB, CGM, EB+CGM, EB+A and CGM+A. We summarized the characteristics of the comparing methods in Table 3.

The IND method estimates the co-occurrence probability by assuming that items occur independently as follows, $\hat{\pi}_{\bar{i}\bar{j}} = (1-\hat{\theta}_i)(1-\hat{\theta}_j)$, $\hat{\pi}_{\bar{i}j} = (1-\hat{\theta}_i)\hat{\theta}_j$, $\hat{\pi}_{i\bar{j}} = \hat{\theta}_i(1-\hat{\theta}_j)$, $\hat{\pi}_{ij} = \hat{\theta}_i\hat{\theta}_j$.

[1]The original sushi data were obtained from http://www.kamishima.net/sushi/.

[2]The original Movielens data were obtained from https://grouplens.org/datasets/movielens/.

The ML method estimates the probability by the maximum likelihood of the small record data as follows, $\hat{\pi}_{\bar{i}\bar{j}} = \frac{x^*_{\bar{i}\bar{j}}}{U^*}$, $\hat{\pi}_{\bar{i}j} = \frac{x^*_{\bar{i}j}}{U^*}$, $\hat{\pi}_{i\bar{j}} = \frac{x^*_{i\bar{j}}}{U^*}$, $\hat{\pi}_{ij} = \frac{x^*_{ij}}{U^*}$, where U^* is the total number of users in the small record data.

The EB method is the empirical Bayesian method that maximizes the marginalized likelihood of the small record data, $L(\mathbf{B}) = \sum_{i,j} \log p(\mathbf{x}^*_{ij}|\boldsymbol{\beta}_{ij})$, where $\mathbf{B} = \{\boldsymbol{\beta}_{ij}\}_{i,j=1}^{I}$ is the set of Dirichlet parameters, and $\sum_{i,j}$ denotes $\sum_{i=1}^{I}\sum_{j=i+1}^{I}$.

The CGM method is the collective graphical model based method (Kumar, Sheldon, and Srivastava 2013) that maximizes the marginalized likelihood of the aggregated data with the constraints on marginal counts, $L(\mathbf{X}, \mathbf{B}) = \sum_{i,j} \log p(\mathbf{x}_{ij}|\boldsymbol{\beta}_{ij})$.

The EB+CGM method is the combination of the EB and CGM methods, where the weighted sum of the marginalized likelihood of the aggregated and small record data, $L(\mathbf{X}, \mathbf{B}) = \lambda \sum_{i,j} \log p(\mathbf{x}_{ij}|\boldsymbol{\beta}_{ij}) + (1 - \lambda) \sum_{i=1}^{I}\sum_{j=i+1}^{I} \log p(\mathbf{x}^*_{ij}|\boldsymbol{\beta}_{ij})$, is maximized. We used the fixed weight hyperparameter $\lambda = 0.5$ with the EB+CGM and proposed methods. With the EB, CGM and EB+CGM methods, Dirichlet parameters $\boldsymbol{\beta}_{ij}$ are directly optimized without neural networks.

The EB+A method is the empirical Bayesian method that uses the auxiliary data, where neural networks are trained by maximizing the marginalized likelihood of the small record data, $L(\boldsymbol{\Psi}, \alpha) = \sum_{i,j} \log p(\mathbf{x}^*_{ij}|\boldsymbol{\beta}_{ij})$, where Dirichlet parameters $\boldsymbol{\beta}_{ij}$ are parameterized by neural networks in Eq.(5).

The CGM+A is the collective graphical model based method that uses the auxiliary data, where neural networks are trained by maximizing the marginalized likelihood of the small record data, $L(\mathbf{X}, \boldsymbol{\Psi}, \alpha) = \sum_{i,j} \log p(\mathbf{x}_{ij}|\boldsymbol{\beta}_{ij})$. Here, as with the EB+A method, Dirichlet parameters $\boldsymbol{\beta}_{ij}$ are parameterized by neural networks in Eq.(5).

The EB+A, CGM+A and proposed methods use the auxiliary information, and the other methods do not use it. After estimating the latent co-occurrence variables, we estimated the co-occurrence probability by $\hat{\pi}_{ij} = \frac{\hat{x}_{ij}}{U}$ with the CGM, EB+CGM, CGM+A and proposed methods, where $\hat{x}_{ij}$ is the estimate of latent co-occurrence variable x_{ij}. Since the EB and EB+A methods did not estimate co-occurrence x_{ij}, they estimated the co-occurrence probability by the expectation with learned prior $\hat{\pi}_{ij} = \frac{\hat{\beta}_{ij}}{\sum_{i'j'} \hat{\beta}_{i'j'}}$, where $\hat{\beta}_{ij}$ is the estimated Dirichlet parameter.

For neural network $\rho_0(\cdot), \rho_1(\cdot), \phi_o(\cdot), \phi_1(\cdot), f_{01}(\cdot)$ in the EB+A, CGM+A and proposed methods, we used three-layer feed-forward neural networks with 32 hidden units. Since $\beta_{i'j'}$ must be non-negative, we use rectified linear unit $\mathrm{ReLU}(x) = \max(0, x)$ at the end of neural networks $\rho_0(\cdot)$, $\rho_1(\cdot)$ and $f_{01}(\cdot)$. We optimized using ADAM (Kingma and Ba 2015) with learning rate 10^{-2}, weight decay 10^{-2} and dropout rate 0.1. For each batch, we randomly sampled 512 item pairs in training. The validation data were used for early stopping, where the maximum number of training epochs was 1,000. We implemented all the methods based on PyTorch (Paszke et al. 2017).

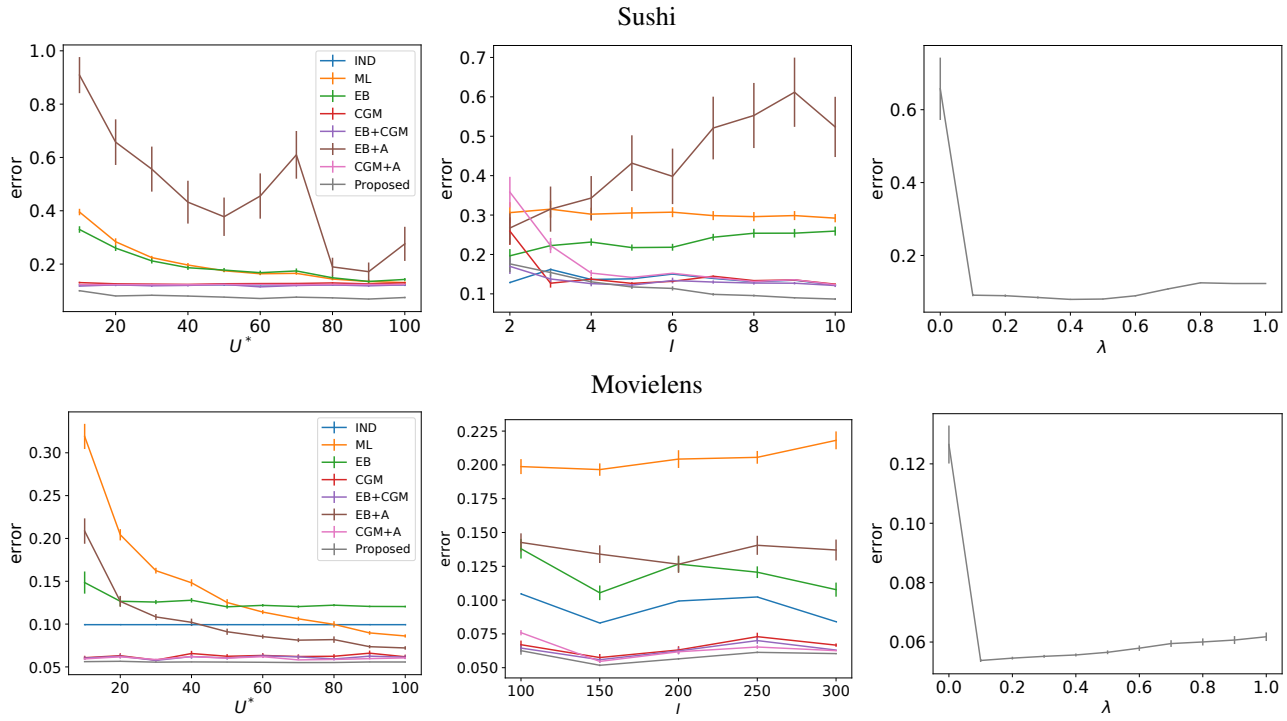
4.4 Results

Figure 2(a) shows the absolute error of the co-occurrence probabilities averaged over 30 experiments with different numbers of sampled users with records U^*, where the number of items was $I = 10$ with the sushi data, and $I = 200$ with the Movielens data. Here, items were randomly sampled from all items with the Movielens data. The proposed method achieved the lowest error for all the cases. The error by the proposed method was statistically lower at the 5% level than that by the other methods by a paired t-test with all the cases in the sushi data and with all the cases except for $U^* = 30$ in the Movielens data. This result indicates that the proposed method improved the estimation performance using all of the aggregated marginal count data, a small number of sampled records, and auxiliary item feature data. The error decreased as the number of observed users U^* increased with the proposed method, since the estimation of the Dirichlet parameters modeled by the neural networks became robust with more training data. Although the performance of the ML, EB, and EB+A methods also improved as the number of observed users increased, their performance was much worse than the proposed method. This result occurred because they did not use aggregated data information. In contrast, the proposed method effectively used the aggregated data information in the first term of the objective function in Eq.(10). The reduction of the error of the proposed method from EB+CGM, which corresponds to the proposed method without auxiliary information, was 34% and 8% on Sushi data and Movielens data with $U^* = 20$, respectively. This result shows the importance to use the auxiliary information for improving performance.

Table 4 shows the average absolute errors for each pair of items on the sushi data with $U^* = 20$ and $I = 10$. The IND method achieved the lowest error with some pairs, where the user preferences were assumed to be independent. The performance of the CGM, EB+CGM, and CGM+A methods resembled the IND method, although they were slightly worse. The proposed method achieved the lowest error with many pairs, which include item pairs whose preferences were not independent. Table 5 shows the rate of item pairs that each method achieved the best on the Movielens data with $U^* = 20$ and $I = 200$. The rate of the the proposed method was highest.

Figure 2(b) shows the absolute error with different numbers of items I, where the number of sampled users was $U^* = 20$ with both the sushi and Movielens data, and we randomly sampled I items from all items. With the sushi data, the performance of the proposed method was not the best when the number of items was small. However, its performance improved as the number of items increased; it was statistically lower at the 5% level than that by the other methods with $I \geq 6$. Since the data for training the neural networks increased with more items, the performance improved. With the Movielens data, the proposed method achieved statistically less error than the other methods for all cases.

Figure 2(c) shows the absolute error with different hyperparameter values λ with the proposed method, where the number of sampled users was $U^* = 20$ with both kinds of



(a) #Users with records U^* (b) #Items I (c) Hyperparameter λ

Figure 2: Absolute error of co-occurrence probabilities averaged over 30 experiments with (a) different numbers of sampled users with records U^*, (b) different numbers of items I, and (c) different hyperparameters λ with proposed method. Error bars show their standard errors.

data, and the number of items was $I = 10$ with the sushi data, and $I = 200$ with the Movielens data. The proposed method corresponds to the EB+A method when $\lambda = 0$, and to the CGM+A method when $\lambda = 1$. The best hyperparameter value was $\lambda = 0.4$ with the sushi data, and $\lambda = 0.1$ with the Movielens data. The co-occurrence data are latent $\mathbf{x}_{ij}$ with the first term of the objective function in Eq.(10), but the co-occurrence data are observed $\mathbf{x}^*_{ij}$ with the second term. Therefore, the best hyperparameter values put more weight on the second term than the first term.

The average computational time with the proposed method for training using the sushi data with $I = 10$, Movielens with $I = 100$, and Movielens with $I = 200$ were 0.12, 5.95 and 24.19 minutes, respectively, on computers with 2.60GHz CPUs, where we fixed $U^* = 20$. The computational time increased quadratically with the number of items since we need to consider all pairs of items. The computational time did not depend on the number of users, U or U^*.

We also evaluated with Movielens-1M data, which contained 6,040 users. We used the genres as auxiliary information, $U^* = 10$ users with records, and $I = 100$ items that were randomly selected from movies rated by more than 300 users. Table 6 shows the results. The proposed method achieved the lowest error with the Movielens-1M data.

5 Conclusion

We proposed a method for estimating co-occurrence with aggregated marginal count data, the records of a small number of users, and auxiliary item feature information. Our proposed method learns unknown relationships between auxiliary information and co-occurrence using neural networks for the parameters of the Dirichlet priors of joint co-occurrence probabilities. Experiments on real-world datasets confirmed that the proposed method achieved better estimation performance for co-occurrence than existing methods. Although our results are encouraging, we must extend our approach in several directions. First, we will apply our framework to general multi-way and/or multivariate contingency tables; we focused on two-by-two contingency tables in this paper. Second, we will investigate tuning the hyperparameters using validation data or a Bayesian framework.

References

Bhowmik, A.; Ghosh, J.; and Koyejo, O. 2015. Generalized linear models for aggregated data. In *Artificial Intelligence and Statistics*, 93–101.

Bhowmik, A.; Ghosh, J.; and Koyejo, O. 2016. Sparse parameter recovery from aggregated data. In *International Conference on Machine Learning*, 1090–1099.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent

Table 4: Absolute error of co-occurrence probabilities averaged over 30 experiments for each pair of items with the sushi data, where the number of sampled users with records was $U^* = 20$ and number of items was $I = 10$. Values in bold typeface are not statistically different (at 5% level) from the best performing method in each row according to a paired t-test. Bottom line shows number of pairs for which the method achieved the best.

Item i	Item j	IND	ML	EB	CGM	EB+CGM	EB+A	CGM+A	Proposed
shrimp	sea eel	0.129	0.285	0.233	0.138	0.120	0.673	0.126	**0.027**
shrimp	tuna	0.136	0.242	0.219	0.098	0.100	0.726	0.133	**0.026**
shrimp	squid	0.035	0.285	0.237	0.062	0.082	0.634	**0.029**	0.165
shrimp	sea urchin	0.161	0.289	0.244	0.164	0.144	0.683	0.158	**0.026**
shrimp	salmon roe	0.190	0.290	0.245	0.207	0.166	0.722	0.186	**0.044**
shrimp	egg	**0.039**	0.305	0.265	0.079	0.075	0.617	0.043	0.052
shrimp	fatty tuna	0.137	0.257	0.235	0.120	0.124	0.769	0.135	**0.101**
shrimp	tuna roll	0.200	0.279	0.231	0.182	0.178	0.658	0.206	**0.098**
shrimp	cucumber roll	**0.002**	0.242	0.240	0.015	0.028	0.524	0.003	0.039
sea eel	tuna	0.221	0.284	0.243	0.172	0.179	0.760	0.218	**0.098**
sea eel	squid	0.157	0.316	0.267	0.146	0.124	0.660	0.163	**0.024**
sea eel	sea urchin	0.017	0.316	0.259	0.020	0.049	0.663	**0.014**	0.147
sea eel	salmon roe	0.126	0.279	0.235	0.137	0.147	0.706	0.123	**0.036**
sea eel	egg	**0.011**	0.313	0.280	0.045	0.054	0.587	0.015	0.080
sea eel	fatty tuna	0.086	0.271	0.252	0.070	0.081	0.772	0.084	**0.039**
sea eel	tuna roll	0.176	0.334	0.287	0.165	0.154	0.653	0.182	**0.048**
sea eel	cucumber roll	**0.071**	0.257	0.249	0.085	0.099	0.519	0.073	0.084
tuna	squid	0.141	0.290	0.279	0.188	0.168	0.723	0.145	**0.041**
tuna	sea urchin	0.203	0.310	0.283	0.154	0.157	0.767	0.201	**0.070**
tuna	salmon roe	0.157	0.295	0.260	0.115	0.109	0.778	0.154	**0.030**
tuna	egg	0.130	0.280	0.280	0.163	0.164	0.693	0.133	**0.069**
tuna	fatty tuna	**0.104**	0.224	0.251	0.120	0.133	0.745	0.106	0.121
tuna	tuna roll	0.102	0.305	0.265	**0.069**	**0.090**	0.668	0.097	0.164
tuna	cucumber roll	**0.088**	0.242	0.273	0.099	0.108	0.632	0.090	**0.110**
squid	sea urchin	0.181	0.324	0.276	0.177	0.160	0.650	0.187	**0.023**
squid	salmon roe	0.205	0.309	0.278	0.188	0.149	0.699	0.210	**0.052**
squid	egg	**0.044**	0.300	0.266	**0.037**	**0.056**	0.544	**0.040**	0.064
squid	fatty tuna	0.175	0.260	0.256	0.193	0.206	0.786	0.177	**0.094**
squid	tuna roll	0.119	0.333	0.272	0.135	0.118	0.598	0.114	**0.044**
squid	cucumber roll	**0.005**	0.258	0.257	0.019	0.032	0.446	0.007	0.033
sea urchin	salmon roe	**0.194**	0.323	0.273	**0.191**	**0.175**	0.660	0.197	0.354
sea urchin	egg	0.275	0.308	0.255	0.311	0.218	0.654	0.279	**0.154**
sea urchin	fatty tuna	**0.057**	0.268	0.254	0.078	0.081	0.733	0.059	0.115
sea urchin	tuna roll	0.273	0.300	0.254	0.269	0.217	0.661	0.278	**0.121**
sea urchin	cucumber roll	0.140	0.277	0.270	0.154	0.155	0.537	0.142	**0.123**
salmon roe	egg	0.170	0.311	0.283	0.212	0.153	0.669	0.174	**0.063**
salmon roe	fatty tuna	**0.007**	0.275	0.264	0.024	0.042	0.774	0.009	0.054
salmon roe	tuna roll	0.227	0.293	0.256	0.211	0.165	0.677	0.233	**0.094**
salmon roe	cucumber roll	**0.102**	0.257	0.265	0.115	0.127	0.583	0.104	**0.104**
egg	fatty tuna	0.169	0.266	0.283	0.183	0.193	0.773	0.171	**0.110**
egg	tuna roll	0.067	0.283	0.243	0.046	0.059	0.530	0.063	**0.033**
egg	cucumber roll	0.084	0.257	0.274	0.094	0.102	0.377	0.086	**0.054**
fatty tuna	tuna roll	**0.027**	0.293	0.261	0.045	0.063	0.736	0.029	0.035
fatty tuna	cucumber roll	**0.142**	0.230	0.296	**0.147**	**0.151**	0.736	**0.143**	**0.135**
tuna roll	cucumber roll	**0.015**	0.256	0.251	0.028	0.041	0.428	0.017	0.027
Best		15	0	0	4	4	0	4	**30**

Table 5: Rate of item pairs that each method achieved the best with the Movielens data, where the number of sampled users with records was $U^* = 20$ and number of items was $I = 200$.

IND	ML	EB	CGM	EB+CGM	EB+A	CGM+A	Proposed
0.121	0.031	0.014	0.170	0.142	0.068	0.211	**0.242**

Table 6: Absolute error of co-occurrence probabilities averaged over 30 experiments with the Movielens-1M data.

IND	ML	EB	CGM	EB+CGM	EB+A	CGM+A	Proposed
0.0169	0.4219	0.0167	0.3952	0.0164	0.0154	0.3597	**0.0147**

Dirichlet allocation. *Journal of Machine Learning Research* 3(Jan):993–1022.

Caruana, R. 1997. Multitask learning. *Machine Learning* 28(1):41–75.

Causey, B. D. 1983. Estimation of proportions for multinomial contingency tables subject to marginal constraints. *Communications in Statistics-Theory and Methods* 12(22):2581–2587.

Deming, W. E., and Stephan, F. F. 1940. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics* 11(4):427–444.

Dobra, A.; Tebaldi, C.; and West, M. 2006. Data augmentation in multi-way contingency tables with fixed marginal totals. *Journal of Statistical Planning and Inference* 136(2):355–372.

Fienberg, S. E. 1970. An iterative procedure for estimation in contingency tables. *The Annals of Mathematical Statistics* 41(3):907–917.

Freedman, D. A.; Klein, S. P.; Sacks, J.; Smyth, C. A.; and Everett, C. G. 1991. Ecological regression and voting rights. *Evaluation Review* 15(6):673–711.

Girvan, M., and Newman, M. E. 2002. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences* 99(12):7821–7826.

Goodman, L. A. 1953. Ecological regressions and behavior of individuals. *American Sociological Review*.

Greco, F. 2016. Estimation of multi-way tables subject to coherence constraints. *Statistica* 76(2):115–125.

Ireland, C. T., and Kullback, S. 1968. Contingency tables with given marginals. *Biometrika* 55(1):179–188.

Iwata, T., and Shimizu, H. 2019. Neural collective graphical models for estimating spatio-temporal population flow from aggregated data. In *AAAI Conference on Artificial Intelligence*.

Kamishima, T. 2003. Nantonac collaborative filtering: recommendation based on order responses. In *Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 583–588. ACM.

Kingma, D. P., and Ba, J. 2015. ADAM: A method for stochastic optimization. In *International Conference on Learning Representations*.

Konstan, J. A.; Miller, B. N.; Maltz, D.; Herlocker, J. L.; Gordon, L. R.; and Riedl, J. 1997. Grouplens: applying collaborative filtering to usenet news. *Communications of the ACM* 40(3):77–87.

Kumar, A.; Sheldon, D.; and Srivastava, B. 2013. Collective diffusion over networks: Models and inference. In *Proceedings of International Conference on Uncertainty in Artificial Intelligence*.

Little, R. J., and Wu, M.-M. 1991. Models for contingency tables with known margins when target and sampled populations differ. *Journal of the American Statistical Association* 86(413):87–95.

Nguyen, T.; Kumar, A.; Lau, H. C.; and Sheldon, D. 2016. Approximate inference using DC programming for collective graphical models. In *Artificial Intelligence and Statistics*, 685–693.

Park, Y., and Ghosh, J. 2012. A probabilistic imputation framework for predictive analysis using variably aggregated, multi-source healthcare data. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, 445–454. ACM.

Park, Y., and Ghosh, J. 2014. Ludia: An aggregate-constrained low-rank reconstruction algorithm to leverage publicly released health data. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 55–64. ACM.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*.

Sarwar, B. M.; Karypis, G.; Konstan, J. A.; and Riedl, J. 2001. Item-based collaborative filtering recommendation algorithms. *WWW* 1:285–295.

Sheldon, D. R., and Dietterich, T. G. 2011. Collective graphical models. In *Advances in Neural Information Processing Systems*, 1161–1169.

Sheldon, D.; Sun, T.; Kumar, A.; and Dietterich, T. G. 2013. Approximate inference in collective graphical models. In *Proceedings of the 30th International Conference on Machine Learning*.

Slavkovic, A. B. 2010. Partial information releases for confidential contingency table entries: Present and future research efforts. *Journal of Privacy and Confidentiality* 1(2).

Smith, J. H. 1947. Estimation of linear functions of cell proportions. *The Annals of Mathematical Statistics* 18(2):231–254.

Stephan, F. F. 1942. An iterative method of adjusting sample frequency tables when expected marginal totals are known. *The Annals of Mathematical Statistics* 13(2):166–178.

Sun, T.; Sheldon, D.; and Kumar, A. 2015. Message passing for collective graphical models. In *Proceedings of the 32nd International Conference on Machine Learning*, 853–861.

Zaheer, M.; Kottur, S.; Ravanbakhsh, S.; Poczos, B.; Salakhutdinov, R. R.; and Smola, A. J. 2017. Deep sets. In *Advances in Neural Information Processing Systems*, 3391–3401.