Leveraging Aggregate Ratings for Better Recommendations

Working paper

CeDER-07-03

Akhmed Umyarov
New York University
Stern School of Business
44 West 4th Street, 8th floor
New York, NY, USA
aumyarov@stern.nyu.edu

Alexander Tuzhilin
New York University
Stern School of Business
44 West 4th Street, 8th floor
New York, NY, USA
atuzhili@stern.nyu.edu

ABSTRACT

The paper presents a method that uses aggregate ratings provided by various segments of users for various categories of items to derive better estimations of unknown individual ratings. This is achieved by converting the aggregate ratings into constraints on the parameters of a rating estimation model presented in the paper. The paper also demonstrates theoretically that these additional constraints reduce rating estimation errors resulting in better rating predictions.

Categories and Subject Descriptors: H.1.2 [Models and Principles]: User/Machine Systems - Human information processing. H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval - Information filtering.

General Terms: Algorithms, Design, Theory

Keywords: Recommender systems, Hierarchical Bayesian models, predictive models, aggregate ratings, OLAP

1. INTRODUCTION

Consider a movie recommender system, such as the one provided by Netflix, and assume that we know an average rating that graduate students provide for action movies from a *reliable external source*. Can we use this type of aggregate rating information to improve quality of individual recommendations? More generally, ratings of individual items provided by individual users can be aggregated into OLAP-based aggregation hierarchies [1], and various aggregate ratings for different groups of users and groups of items at different levels of the OLAP hierarchy can be known to the recommender system. For example, the IMDB database provides average ratings of movies by various categories of users, such as Male vs. Female ratings. In this paper, we describe how this *aggregate* rating information from external sources can be leveraged for providing better recommendations of *individual* items to *individual* users.

We study this problem in the context of the hierarchical regression models, both Bayesian and frequentist, that were independently proposed by statisticians [5] and marketers [3] studying recommender systems. We decided to use this type of hierarchical regression models [12] for the following reasons. First, they constitute *hybrid* models integrating both user and item characteristics into a single recommendation model. Generally, hybrid models tend to outperform collaborative and content-based recommendation methods in many cases [2]. In fact, the Hierarchical Bayesian model presented in [3] outperformed a collaborative filtering model [3]. Second, these models are based on strong statistical theory and have nice statistical properties that can be analyzed theoretically, as is done in this paper. However, the general approach presented in this paper is not limited to this particular type of models and can be generalized to various other statistical and data mining models and approaches. For example, [4] presents a method for using aggregate information about traversal of hypertext pages by a group of users in order to provide better recommendations of hypertext pages to individual members of the group. In contrast to this top-down approach, [11] presents a bottom-up approach in which the goal is to provide recommendations to a group of users. Then these group recommendations are based on the aggregate ratings that are computed based on the individual ratings of the members of the group.

In this paper we show theoretically that the extra knowledge of the external aggregate ratings indeed leads to more accurate recommendations. We also show *how* this aggregate rating information can be converted into additional constraints on model parameters leading to better estimations of individual unknown ratings. Finally, we present a particular semi-parametric frequentist method for estimating parameters of hierarchical regression models and show how the method incorporates aggregate information.

Before presenting the aggregate method, we first describe hierarchical regression models in Sections 2.1 and 3 and how they are used for estimating unknown individual ratings.

RecSys'07, October 19–20, 2007, Minneapolis, Minnesota, USA.
Copyright 2007 ACM 978-1-59593-730-8/07/0010 ...$5.00.

2. HIERARCHICAL BAYESIAN REGRESSION MODEL

2.1 Model specification

As explained in Section 1, [3] describes a hybrid approach to rating estimation that uses the following Hierarchical Bayesian (HB) linear regression model:

$$\begin{cases} r_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\mu} + \boldsymbol{z}'_i\boldsymbol{\gamma}_j + \boldsymbol{w}'_j\boldsymbol{\lambda}_i + \varepsilon_{ij}, \\ \varepsilon_{ij} \sim N(0,\sigma^2), \quad \boldsymbol{\gamma}_j \sim N(\boldsymbol{0},\Gamma), \quad \boldsymbol{\lambda}_i \sim N(\boldsymbol{0},\Lambda), \end{cases} \tag{1}$$

where observed values of the model are ratings r_{ij} assigned by user i for item j, $\boldsymbol{z}_i$ is a vector[1] of attributes of user i, such as age, gender, etc., $\boldsymbol{w}_j$ is a vector of attributes[2] of item j, such as price, weight, etc., and vector $\boldsymbol{x}_{ij} = \boldsymbol{z}_i \otimes \boldsymbol{w}_j$, where $\otimes$ is the Kronecker product. Intuitively, $\boldsymbol{x}_{ij}$ is a long vector containing all possible cross-products between individual elements of $\boldsymbol{z}_i$ and $\boldsymbol{w}_j$.

Vector $\boldsymbol{\mu}$ represents unobserved slope of the regression, vectors $\boldsymbol{\gamma}_j$ and $\boldsymbol{\lambda}_i$ represent unobserved item heterogeneity and user heterogeneity effects respectively. Moreover, the model (1) assumes that vector $\boldsymbol{\gamma}_j \sim N(\boldsymbol{0},\Gamma)$ and $\boldsymbol{\lambda}_i \sim N(\boldsymbol{0},\Lambda)$, where Γ and Λ are unobserved covariance matrices, and that each observation r_{ij} has also an i.i.d. disturbance $\varepsilon_{ij} \sim N(0,\sigma^2)$, where σ is also an unobserved parameter.

Thus, vectors $\boldsymbol{\mu}$, $\{\boldsymbol{\gamma}_j\}$ and $\{\boldsymbol{\lambda}_i\}$, scalar σ, covariance matrices Γ and Λ constitute the unknown parameters of model (1). Prior belief about these parameters is introduced in [3], and the parameters are estimated from the known ratings r_{ij} and known user/item data using Markov Chain Monte Carlo (MCMC) method [6], which constitutes one of the Bayesian estimation techniques for finding the expected value of the posterior distributions of parameters.

[3] compared predictive performance of their Hierarchical Bayesian model (1) against the classical collaborative filtering methods and demonstrated that model (1) outperformed the collaborative filtering considered in [3].

2.2 Why this model?

The natural question to ask is why to use this particular type of model and why the model has the specification it has.

Consider the following steps in deriving this model:

1. Assume we regress movie ratings r_{ij} solely on movie attributes $\boldsymbol{w}_j$.

   $$r_{ij} = \boldsymbol{w}'_j\boldsymbol{\beta}_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0,\sigma^2) \tag{2}$$

   So we run separate regressions for each user i and we get the user-specific vector of coefficients $\boldsymbol{\beta}_i$. Intuitively, j-th element of each vector $\boldsymbol{\beta}_i$ is a (user-specific) "appreciation" to the j-th characteristic of movies. For example, if j-th characteristic of a movie is movie release year, then j-th element of $\boldsymbol{\beta}_i$ will represent average "attitude" of user i towards newer or older movies.

2. Now we say that since the vector of coefficients $\boldsymbol{\beta}_i$ is user-specific, we can try to explain **each element** of it from known user attributes $\boldsymbol{z}_i$.

   $$\boldsymbol{\beta}_i = Z_i\boldsymbol{\mu} + \boldsymbol{\lambda}_i, \quad \boldsymbol{\lambda}_i \sim N(\boldsymbol{0},\Lambda) \tag{3}$$

   where matrix Z_i is constructed from the vector-column of the user attributes $\boldsymbol{z}_i$ as follows:

   $$Z_i = \begin{pmatrix} \boldsymbol{z}'_i & 0 & \cdots & 0 \\ 0 & \boldsymbol{z}'_i & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \ddots & 0 \\ 0 & \cdots & 0 & \boldsymbol{z}'_i \end{pmatrix}$$

   Intuitively, each element of $\boldsymbol{\mu}$ here represents a general "effect" of some user characteristic on his "appreciation" of some movie characteristic. For example, if j-th movie characteristic is movie release year, k-th user characteristic is user age and size of vector $\boldsymbol{z}_i$ is $|\boldsymbol{z}|$. Then the element $\boldsymbol{\mu}_{(j-1)|\boldsymbol{z}|+k}$ can be interpreted as the general effect of user age on his attitude towards movie release year. This interpretation is very similar to the interpretation of regressions with included *interaction terms* that are widely used in social research.

3. Now we substitute eq.(3) into eq.(2) and get

   $$r_{ij} = \boldsymbol{w}'_j\boldsymbol{\beta}_i + \varepsilon_{ij} = \boldsymbol{w}'_j(Z_i\boldsymbol{\mu} + \boldsymbol{\lambda}_i) + \varepsilon_{ij} =$$

   $$= \underbrace{\boldsymbol{w}'_j Z_i}_{\boldsymbol{x}'_{ij}}\boldsymbol{\mu} + \boldsymbol{w}_j\boldsymbol{\lambda}_i + \varepsilon_{ij}$$

   This is how we define the vector $\boldsymbol{x}_{ij}$ and if we examine the vector in detail this vector contains all "interactions" (cross-products) between elements of vectors $\boldsymbol{z}_i$ and $\boldsymbol{w}_j$.

   So right now we got the model

   $$r_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\mu} + \boldsymbol{w}'_j\boldsymbol{\lambda}_i + \varepsilon_{ij} \tag{4}$$

   Now if we repeat the same procedure but at the step 1) we will regress r_{ij} on user attributes $\boldsymbol{z}_i$, instead of movie attributes $\boldsymbol{w}_j$, we will get the model

   $$r_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\mu} + \boldsymbol{z}'_i\boldsymbol{\gamma}_j + \varepsilon_{ij} \tag{5}$$

   (since this task is purely symmetrical of movie attributes and user attributes)

4. Now we unite the two models from eq.(4) and eq.(5) into a single model (just sum them!) and we finally get our exact model

   $$r_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\mu} + \boldsymbol{z}'_i\boldsymbol{\gamma}_j + \boldsymbol{w}'_j\boldsymbol{\lambda}_i + \varepsilon_{ij}$$

2.3 Use of the model

In most practical cases, the number of parameters to be estimated for model (1) is very large. For example, for 1000 users defined by 5 user attributes and 1000 movies defined by 20 movie attributes, we will need to estimate more than 25,000 free parameters in the model. Since we will be dealing with parameter constraints, it is of our interest to examine properties of the model subject to constraints. In general, according to [7], constrained Bayesian estimation techniques are notorious for their computational difficulty, especially in high-dimensional parameter spaces, as is the case with model (1).

[1] We typed vectors in bold font as opposed to matrices and scalars that are typed in regular font.

[2] We also include constant term both in $\boldsymbol{z}_i$ as a user attribute and in $\boldsymbol{w}_j$ as an item attribute.

It is possible to come up with a constrained sampling procedure that *theoretically* eventually converges to its population counterpart. For example, Metropolis-Hastings Algorithm [15] can be used. Metropolis-Hastings Algorithm allows to impose the constraint not on the posterior distribution of parameters that can have a complicated analytical expression, but on more simple jumping distribution. Due to properties of Metropolis-Hastings Algorithm, the resulting sequence will converge to the constrained posterior distribution as if the constraint was actually imposed on the posterior distribution of parameters. However, in practice convergence of this algorithms in the space with tens of thousands dimensions is not feasible.

To address this difficulty, we propose to use a frequentist semi-parametric approach that we present in the next section for solving the aggregate rating problem instead of the Bayesian parametric method defined by (1).

3. GENERALIZED LINEAR REGRESSION MODEL

Consider the same model as in (1), but from a frequentist semi-parametric perspective[3]:

$$\begin{cases} r_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\mu} + \boldsymbol{z}'_i\boldsymbol{\gamma}_j + \boldsymbol{w}'_j\boldsymbol{\lambda}_i + \varepsilon_{ij}, \\ E\left[\varepsilon_{ij}\right] = 0, \quad \mathrm{Var}\left[\varepsilon_{ij}\right] = \sigma^2, \\ E\left[\boldsymbol{\gamma}_j\right] = \mathbf{0}, \quad \mathrm{Var}\left[\boldsymbol{\gamma}_j\right] = \Gamma, \\ E\left[\boldsymbol{\lambda}_i\right] = \mathbf{0}, \quad \mathrm{Var}\left[\boldsymbol{\lambda}_i\right] = \Lambda. \end{cases} \tag{6}$$

For a frequentist, $\boldsymbol{\gamma}_j$ and $\boldsymbol{\lambda}_i$ constitute random effects, so that the model (6) constitutes a mixed-effects model [8].

We introduce the notion of a *compound disturbance* η_{ij} by grouping together all the random effects in (6) as follows

$$r_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\mu} + \underbrace{\boldsymbol{z}'_i\boldsymbol{\gamma}_j + \boldsymbol{w}'_j\boldsymbol{\lambda}_i + \varepsilon_{ij}}_{\eta_{ij}}, \tag{7}$$

thus making it a Generalized Least Squares linear regression model (GLS). Moreover, $\boldsymbol{\mu}$ can be consistently[4] estimated by using ordinary least squares estimator (OLS) if we assume that $\boldsymbol{\gamma}_j$ and $\boldsymbol{\lambda}_i$ are not correlated with $\boldsymbol{x}_{ij}$.

The covariance structure of residuals η_{ij} can be determined from equations (6) and (7) as follows:

$$\begin{cases} E\eta_{ij} = 0, \\ E\eta_{ij}\eta_{kl} = 0, \quad \text{if } i \neq k \text{ and } j \neq l, \\ E\eta_{ij}\eta_{ik} = \boldsymbol{w}'_j\Lambda\boldsymbol{w}_k, \quad \text{if } j \neq k, \\ E\eta_{ij}\eta_{kj} = \boldsymbol{z}'_i\Gamma\boldsymbol{z}_k, \quad \text{if } i \neq k, \\ E\eta^2_{ij} = \sigma^2 + \boldsymbol{z}'_i\Gamma\boldsymbol{z}_i + \boldsymbol{w}'_j\Lambda\boldsymbol{w}_j, \end{cases} \tag{8}$$

where expected value $E(\cdot)$ is taken over ε_{ij}, $\boldsymbol{\lambda}_i$ and $\boldsymbol{\gamma}_j$.

To show why this is the case, consider for example

$$E[\eta_{ij}\eta_{ik}] =$$

$$= E\left[(\boldsymbol{z}'_i\boldsymbol{\gamma}_j + \boldsymbol{w}'_j\boldsymbol{\lambda}_i + \varepsilon_{ij})(\boldsymbol{z}'_i\boldsymbol{\gamma}_k + \boldsymbol{w}'_k\boldsymbol{\lambda}_i + \varepsilon_{ik})\right] =$$

$$= E\left[\boldsymbol{z}'_i\boldsymbol{\gamma}_j\boldsymbol{\gamma}'_k\boldsymbol{z}_i + \boldsymbol{w}'_j\boldsymbol{\lambda}_i\boldsymbol{\gamma}'_k\boldsymbol{z}_i + \varepsilon_{ij}\boldsymbol{\gamma}'_k\boldsymbol{z}_i)+\right.$$

$$+\boldsymbol{z}'_i\boldsymbol{\gamma}_j\boldsymbol{\lambda}'_i\boldsymbol{w}_k + \boldsymbol{w}'_j\boldsymbol{\lambda}_i\boldsymbol{\lambda}'_i\boldsymbol{w}_k + \varepsilon_{ij}\boldsymbol{\lambda}'_i\boldsymbol{w}_k+$$

$$\left.+\boldsymbol{z}'_i\boldsymbol{\gamma}_j\varepsilon_{ik} + \boldsymbol{w}'_j\boldsymbol{\lambda}_i\varepsilon_{ik} + \varepsilon_{ij}\varepsilon_{ik}\right] =$$

$$= E\left[\boldsymbol{w}'_j\boldsymbol{\lambda}_i\boldsymbol{\lambda}'_i\boldsymbol{w}_k\right] = \boldsymbol{w}'_j\Lambda\boldsymbol{w}_k$$

The last equality holds because the expected values of all other terms are zeros, since we assumed that ε_{ij} are i.i.d, $\boldsymbol{\lambda}_i$ are i.i.d, $\boldsymbol{\gamma}_j$ are i.i.d., and also $\boldsymbol{\gamma}_j$, $\boldsymbol{\lambda}_i$ and ε_{ij} are independent $\forall i, j$.

Other equations in (8) are derived similarly.

Let Ω be the covariance matrix of a very long vector of residuals $\boldsymbol{\eta} = ||\eta_{ij}||$; that is $\Omega = \mathrm{Var}(\boldsymbol{\eta})$. From (8), we conclude that Ω depends just on a few unknown parameters: σ, Γ and Λ. Thus σ, Γ and Λ can be consistently estimated from OLS residuals. For example, we can use the following (overdetermined) system of linear equations:

$$\begin{cases} \displaystyle\sum_{\substack{ijk \\ j\neq k, i\in S_U}} \boldsymbol{w}'_j\Lambda\boldsymbol{w}_k = \sum_{\substack{ijk \\ j\neq k, i\in S_U}} e_{ij}e_{ik}, \quad \forall S_U \\ \displaystyle\sum_{\substack{ijk \\ i\neq k, j\in S_I}} \boldsymbol{z}'_i\Gamma\boldsymbol{z}_k = \sum_{\substack{ijk \\ i\neq k, j\in S_I}} e_{ij}e_{kj}, \quad \forall S_I \\ \displaystyle\sigma^2 = \frac{1}{N}\sum_{ij}\left[e^2_{ij} - \boldsymbol{z}'_i\Gamma\boldsymbol{z}_i - \boldsymbol{w}'_j\Lambda\boldsymbol{w}_j\right], \\ \Lambda' = \Lambda, \quad \Gamma' = \Gamma, \end{cases} \tag{9}$$

where e_{ij} is the OLS residual corresponding to observation r_{ij}, N is the total number of observations and S_U and S_I are some subsets of users and items respectively.

Parameter $\boldsymbol{\mu}$ of the model (6) can be estimated asymptotically efficiently using the *Feasible GLS (FGLS)* estimator approach [8] as follows:

$$\hat{\boldsymbol{\mu}} = \left(X'\hat{\Omega}^{-1}X\right)^{-1}X'\hat{\Omega}^{-1}\boldsymbol{r}, \tag{10}$$

where $\boldsymbol{r}$ is a column-vector of observed scalars r_{ij} stacked on top of each other, so the first element of the vector is a scalar $r_{i_1j_1}$, the second element is $r_{i_2j_2}$ and so on. X is a matrix of row-vectors $\boldsymbol{x}'_{ij}$ stacked on top of each other one-by-one; thus the first row of the matrix X is a row-vector $\boldsymbol{x}'_{i_1j_1}$ corresponding to observation $r_{i_1j_1}$, the second row of the matrix X is the row-vector $\boldsymbol{x}'_{i_2j_2}$ and so on. $\hat{\Omega}$ is an estimate of Ω.

Once we estimated consistently parameters σ, Γ and Λ, we can consistently estimate expressions $X'\Omega^{-1}X$ and $X'\Omega^{-1}\boldsymbol{r}$ using the estimates $\hat{\sigma}$, $\hat{\Gamma}$, $\hat{\Lambda}$ and expression (8), and then obtain consistent and asymptotically efficient estimate of $\boldsymbol{\mu}$ using expression (10).

To demonstrate that $\hat{\boldsymbol{\mu}}$ consistently estimates the true value $\boldsymbol{\mu}$ we can proceed as follows. It follows from (7) that

$$\boldsymbol{r} = X\boldsymbol{\mu} + \boldsymbol{\eta}, \text{ where } E\left[\boldsymbol{\eta}\boldsymbol{\eta}'\right] = \Omega$$

Then from equation (10) we have

$$\hat{\boldsymbol{\mu}} = \left(X'\hat{\Omega}^{-1}X\right)^{-1}X'\hat{\Omega}^{-1}(X\boldsymbol{\mu} + \boldsymbol{\eta}) =$$

$$= \boldsymbol{\mu} + \underbrace{\left(X'\hat{\Omega}^{-1}X\right)^{-1}X'\hat{\Omega}^{-1}\boldsymbol{\eta}}_{\text{goes to } 0 \text{ as } N\to\infty}$$

Thus random variable $\hat{\boldsymbol{\mu}}$ converges in probability to the true value $\boldsymbol{\mu}$

$$\hat{\boldsymbol{\mu}} \to \boldsymbol{\mu} \text{ as } N \to \infty$$

[3] Semi-parametric perspective makes no assumptions about the shape of distributions. For example, we don't assume that the residuals are normally distributed. Instead, we make assumptions only about the moments of the residual distribution, not about the shape of the distribution.

[4] Although, not efficiently [8].

The use of "weighting matrix" $\hat{\Omega}^{-1}$ is explained in [8]. Intuitively, $\hat{\Omega}^{-1}$ is a weighting matrix that minimizes the asymptotic variance of $\hat{\boldsymbol{\mu}}$, making the estimator asymptotically efficient.

As long as we assume that for each user $N_i \to \infty$ and for each item $N_j \to \infty$ as $N \to \infty$ for asymptotic analysis, we are able to estimate consistently individual item heterogeneities $\{\boldsymbol{\gamma}_j\}$ and $\{\boldsymbol{\lambda}_i\}$ from the following (overdetermined) system of linear equations

$$\hat{\eta}_{ij} = \boldsymbol{z}_i'\boldsymbol{\gamma}_j + \boldsymbol{w}_j'\boldsymbol{\lambda}_i + \varepsilon_{ij} \quad \forall \text{ observations } (i,j), \qquad (11)$$

where $\hat{\eta}_{ij}$ is a consistent estimator of η_{ij}, for example, it can be an OLS residual $\hat{\eta}_{ij} = e_{ij}$.

System (11) can be interpreted as an ordinary linear regression with dependent variables $\hat{\eta}_{ij}$, regressors $\boldsymbol{z}_i$ and $\boldsymbol{w}_j$, and i.i.d. disturbances ε_{ij}. Since $\hat{\eta}_{ij}$ is a consistent estimator of η_{ij}, the OLS estimators $\hat{\boldsymbol{\lambda}_i}$ and $\hat{\boldsymbol{\gamma}_j}$ consistently estimate $\boldsymbol{\lambda}_i$ and $\boldsymbol{\gamma}_j$ given our assumption about asymptotic behavior of the model. Thus, the frequentist model (6) gives as much of individual heterogeneity information as Bayesian model (1).

As it follows from (10), estimation of $\boldsymbol{\mu}$ requires inverting matrix $\hat{\Omega}$ that is of size $N \times N$, where N is the total number of observations. Matrix $\hat{\Omega}$ is sparse, symmetric and positive-semidefinite and one can use Cholesky decomposition for sparse matrices $\hat{\Omega} = LL'$, where L is the lower-triangular matrix, in order to calculate and store the inverse.

Note that we don't have to store $\hat{\Omega}^{-1}$ itself, we only need to calculate the $X'\hat{\Omega}^{-1}X$ and $X'\hat{\Omega}^{-1}\boldsymbol{r}$. We also notice that $\hat{\Omega}^{-1} = L^{-1'}L^{-1}$ and L^{-1} is itself lower-triangular. Thus,

$$X'\hat{\Omega}^{-1}X = X'L^{-1'}L^{-1}X = (L^{-1}X)'(L^{-1}X), \qquad (12)$$

$$X'\hat{\Omega}^{-1}\boldsymbol{r} = X'L^{-1'}L^{-1}\boldsymbol{r} = (L^{-1}X)'(L^{-1}\boldsymbol{r}). \qquad (13)$$

Unfortunately, matrix $\hat{\Omega}$ is not a band matrix, so the required storage for Cholesky decomposition matrix L can be as large as $O\left(N^2\right)$ of memory, that is too high for large problems. Computational complexity for naive algorithms can be as large as $O(N^3)$. However, the problem is parallelizable. For example, the inversion of triangular matrix takes $O(\log^2 N)$ operations with $O(N^3/\log N)$ processors [10].

Determination of how to invert the sparse matrix $\hat{\Omega}$ more efficiently, and thus making the whole aggregate rating problem scalable, constitutes one of our future research topics.

4. INTRODUCING AGGREGATE RATING

The main research question addressed in this paper is how to use the aggregate ratings in our statistical models to provide better estimators of individual ratings.

Formally, assume that in addition to the classical individual ratings r_{ij}, user data $\boldsymbol{z}_i$ and item data $\boldsymbol{w}_j$ used in equations (1) and (6), we also know the expected value[5] of an average rating across some segment S of user-item pairs. For example, assume that we know for certain from external sources that average rating of all graduate students for all Chaplin movies is 9.1 out of 10. So the segment S in this case is a Cartesian product of the set of graduate students and the set of Chaplin movies.

Assume that there are k total possible user-item pairs in the segment S, thus

$$E_{\varepsilon}\left[\frac{\sum_{i,j} r_{ij}}{k}\right] = a, \qquad (14)$$

where sum is taken over all user-item pairs $(i,j) \in S$. As another example, assume that the expected average rating of some 100 action movies provided by 20 graduate CS students, based on $k = 2000$ possible user-item pairs, is $a = 7.8$.

Substituting the expression for r_{ij} from our model equations (1) or (6), we conclude that

$$E\left[\frac{\sum r_{ij}}{k}\right] = E\left[\frac{\sum\left(\boldsymbol{x}_{ij}'\boldsymbol{\mu} + \boldsymbol{z}_i'\boldsymbol{\gamma}_j + \boldsymbol{w}_j\boldsymbol{\lambda}_i + \varepsilon_{ij}\right)}{k}\right] = \qquad (15)$$

$$= \frac{\sum \boldsymbol{x}_{ij}'}{k}\boldsymbol{\mu} + \frac{\sum \boldsymbol{z}_i'\boldsymbol{\gamma}_j}{k} + \frac{\sum \boldsymbol{w}_j\boldsymbol{\lambda}_i}{k} = a. \qquad (16)$$

Note that both the Bayesian model (1) and the frequentist model (6) have the same expression for r_{ij}, thus the equation (16) has the same form for both. However, interpretation of the equation (16) can be different for the two approaches.

For the Bayesian model (1), the new information from equation (16) about the expected average rating is interpreted as a linear equality constraint on unknown parameters $\boldsymbol{\mu}$, $\{\boldsymbol{\gamma}_j\}$, $\{\boldsymbol{\lambda}_i\}$. For the frequentist model (6), the new information from equation (16) about the expected average rating is interpreted as an additional observation. To see this, denote $\tilde{\boldsymbol{x}} = \frac{\sum \boldsymbol{x}_{ij}}{k}$ and $\tilde{\eta} = \frac{\sum \boldsymbol{z}_i'\boldsymbol{\gamma}_j}{k} + \frac{\sum \boldsymbol{w}_j\boldsymbol{\lambda}_i}{k}$. Then equation (16) is equivalent to having an additional observation in the model

$$a = \tilde{\boldsymbol{x}}'\boldsymbol{\mu} + \tilde{\eta}, \qquad (17)$$

where the residual $\tilde{\eta}$ has a known covariation structure with other residuals η_{ij} defined in (7):

$$E[\tilde{\eta}\eta_{ij}] = \sum_{\substack{t:\\(i,t)\in S}} \frac{\boldsymbol{w}_j'\Lambda\boldsymbol{w}_t}{k} + \sum_{\substack{t:\\(t,j)\in S}} \frac{\boldsymbol{z}_i'\Gamma\boldsymbol{z}_t}{k}, \qquad (18)$$

$$E[\tilde{\eta}^2] = \sum_{\substack{i,j,t:\\(i,j)\in S,\\(i,t)\in S}} \frac{\boldsymbol{w}_j'\Lambda\boldsymbol{w}_t}{k^2} + \sum_{\substack{i,j,t:\\(i,j)\in S,\\(t,j)\in S}} \frac{\boldsymbol{z}_i'\Gamma\boldsymbol{z}_t}{k^2}. \qquad (19)$$

Therefore, the constrained model still fits the GLS paradigm presented in Section 3. Note that for the FGLS estimator, equations (18) and (19) introduce an additional row and a column to matrix Ω corresponding to covariances (18) and (19). That is,

$$\tilde{\Omega} = \left(\begin{array}{c|c} \Omega & * \\ \hline * & * \end{array}\right),$$

where $*$ denotes these additional column and row.

So by including this additional observation we create the corresponding matrix $\tilde{\Omega}$ from the matrix Ω.

5. MULTIPLE AGGREGATE RATINGS

In the previous section, we considered only one true aggregate rating a for one particular segment of ratings. In this section, we assume that there is a whole aggregation hierarchy defined for the ratings matrix. One example would be an OLAP-based hierarchy [9] of aggregate ratings.

OLAP-based hierarchy is a concept that is used here to reflect that both users and items can be classified into some hierarchy. For example, users can be divided into groups

[5] Here we take expected value only over ε, not $\boldsymbol{\gamma}_j$ and $\boldsymbol{\lambda}_i$

"Students" and "Not students". "Students" group can be further divided into "freshmen", "sophomores" etc., so each user has his corresponding "path" in that hierarchy.

Same ideas about hierarchy can be applied to items as well. For example, movies can be divided into "Comedies" and "Not comedies". Each comedy can be divided even further into "Comedies with Chaplin" and "Comedies without Chaplin" etc.

OLAP-based hierarchy is intended to represent the two hierarchies for users and for items in a single concept. The unit of hierarchy is called OLAP cell and represents some unit of user hierarchy connected to some unit of item hierarchy. For example, valid cells here would be "Freshmen"×"Comedies", or "Students"×"Comedies", or "Freshmen"×"Comedies with Chaplin", etc.

Given an OLAP hierarchy for users and items, where ratings constitute measures defined for the OLAP cells [9], consider a particular category of items C_p, a particular segment of users S_q and the cell CELL_{pq} in the OLAP hierarchy corresponding to C_p and S_q. Also let D_{pq} be all the ratings that users in segment S_q provided for items in category C_p, and let R_{pq}^{aggr} be the aggregate rating for CELL_{pq} that was *independently assigned* by the expert to that cell.

Clearly, the expert can assign numerous ratings R^{aggr} to various OLAP cells at different levels of the OLAP hierarchy. Using the results from Section 4, each aggregate rating R^{aggr} produces a constraint of the form (16). This means that various aggregate ratings R_{pq}^{aggr} produce multiple constraints for different values of p and q and that these constraints come from various levels of aggregation in the OLAP hierarchy.

In fact, we may introduce so many such constraints that the estimator itself will be largely determined by the constraints and not the real observation data. The solution to this problem for the aggregation model presented in this paper is that we may have different *levels of confidence* in the aggregate ratings. For example, we may be more sure that the average rating provided by graduate CS students from University of XYZ for action movies is 6.5 than in that the average rating by physics students for drama movies is 7.8.

To model this "degree of confidence" in aggregate ratings, we assume that the aggregate ratings are "noisy," which can be formally represented as:

$$\begin{cases} E_\varepsilon\left[\frac{\sum_{i,j} r_{ij}}{k}\right] = \alpha, \\ a = \alpha + \xi, \quad E\xi = 0, \ \mathrm{Var}(\xi) = \sigma_\xi^2, \end{cases} \tag{20}$$

where ξ is an unknown noise component, α is an unknown true value, a is the observed value for the aggregate rating and σ_ξ^2 is some known parameter.

Including this noise into expression (14) results in the following *fuzzy constraint* rather than the crisp constraint (16):

$$a = \underbrace{\frac{\sum \boldsymbol{x}'_{ij}}{k}}_{\tilde{\boldsymbol{x}}} \boldsymbol{\mu} + \underbrace{\frac{\sum \boldsymbol{z}'_i \boldsymbol{\gamma}_j}{k} + \frac{\sum \boldsymbol{w}_j \boldsymbol{\lambda}_i}{k} + \xi}_{\tilde{\eta}}\,. \tag{21}$$

From the frequentist prospective, the model still can be interpreted as an additional observation of type (17). Therefore, the multiple aggregation model with different degrees of certainty in various aggregate ratings can still be defined with the GLS framework, and the same analysis presented in Sections 3 and 4 still holds. By including this additional observation (21), we create the corresponding matrix $\tilde{\Omega}$ from the matrix Ω defined in Section 4. It can be shown that $\tilde{\Omega}$ is not singular.

Parameter σ_ξ^2 in (20) has the following intuition: it can be interpreted as the weight that we place on the corresponding constraint. It is clear that the larger σ_ξ^2 is, the less the FGLS method will try to satisfy the constraint. Intuitively, σ_ξ^2 represents how strong the noise component is in our observation, so it makes sense to give higher weight to observations with low noise and give lower weight to very noisy observations. FGLS uses this fact for more efficient estimation [8].

Moreover, when we consider multiple constraints, we can put different weights on different constraints by assigning to each constraint i its own "weight" $\sigma_{\xi_i}^2$. In this way, we can accommodate a real situation when some external rating information is more reliable than the other.

The following proposition demonstrates that the constrained models using aggregate ratings, such as FGLS, provide better individual rating estimations than the unconstrained ones.

Proposition 1. The expected mean squared error (MSE) on a test set of the constrained FGLS estimator is smaller than the one of the unconstrained FGLS estimator.

Proof. Intuitively, the proof is based on the idea that specifying an aggregate rating is equivalent to adding a new observation and on the idea that the sample size matters, i.e., the expected MSE on the test set of the estimator trained on the bigger sample size will be smaller than the expected MSE on the test set of the estimator trained on the subset of the sample.

More formally, consider the model as we have it in (7)

$$\boldsymbol{y} = X\boldsymbol{\mu} + \boldsymbol{\eta}, \quad E\boldsymbol{\eta}\boldsymbol{\eta}' = \Omega \tag{22}$$

Denote $\boldsymbol{m}$ — the GLS estimator of $\boldsymbol{\mu}$. This model doesn't take into account additional information, so we call $\boldsymbol{m}$ unrestricted estimator.

Consider also the following model

$$\boldsymbol{y}_* = X_*\boldsymbol{\mu} + \boldsymbol{\eta}_*, \quad E\boldsymbol{\eta}_*\boldsymbol{\eta}'_* = \Omega_*$$

where we just added one observation to eq.(22). So X_* is just X with one additional row corresponding to the observation and Ω_* is just Ω with additional row and column corresponding to covariances of the additional observation with all other observations. That is,

$$\boldsymbol{y}_* = \left(\begin{array}{c} \boldsymbol{y} \\ \hline * \end{array}\right) \quad X_* = \left(\begin{array}{c} X \\ \hline * \end{array}\right)$$

and

$$\Omega_* = \left(\begin{array}{c|c} \Omega & * \\ \hline * & * \end{array}\right)$$

Denote $\boldsymbol{m}_*$ — the estimator for this model. The model takes into account the additional observation, so we call it restricted estimator.

Denote $V = \mathrm{Var}[\boldsymbol{m}]$ and $V_* = \mathrm{Var}[\boldsymbol{m}_*]$.

As we know from [8]

$$\mathrm{Var}[\boldsymbol{m}] = \left(X'\Omega^{-1}X\right)^{-1}$$

and Cholesky decomposition of Ω:

$$\Omega = C'C$$

Thus

$$\Omega^{-1} = C^{-1}\left(C^{-1}\right)'$$

Now do the same thing for Ω_*:

$$\Omega_* = C_*' C_*$$

Actually, C_* is equal to C with an additional column (and an additional row of zeros). That is,

$$C_* = \left(\begin{array}{c|c} C & * \\ \hline 0 & * \end{array}\right)$$

It is a trivial fact since Ω_* differs from Ω just by existance of additional column and additional row. It is also a trivial fact that C_*^{-1} is equal to C^{-1} with an additional column. That is,

$$C_*^{-1} = \left(\begin{array}{c|c} C^{-1} & * \\ \hline 0 & * \end{array}\right)$$

Consider

$$(\mathrm{Var}[\boldsymbol{m}])^{-1} = X'\Omega^{-1}X = X'C^{-1}\left(C^{-1}\right)'X$$

Consider also

$$(\mathrm{Var}[\boldsymbol{m}_*])^{-1} = X_*'\Omega_*^{-1}X_* = X_*'C_*^{-1}\left(C_*^{-1}\right)'X_*$$

As we noted, C_*^{-1} is equal to C^{-1} with an additional column, thus $\left(C_*^{-1}\right)'$ is equal to $\left(C^{-1}\right)'$ with an additional row. It is also easy to notice that $\left(C_*^{-1}\right)'X_*$ differs from $\left(C^{-1}\right)'X$ only by the addition of the last row. Denote this last row as row-vector $\tilde{\boldsymbol{x}}'$. Then,

$$\left(C_*^{-1}\right)'X_* = \begin{pmatrix} \left(C^{-1}\right)'X \\ \tilde{\boldsymbol{x}}' \end{pmatrix}$$

It means that

$$\overbrace{\left(\left(C_*^{-1}\right)'X_*\right)'\left(C_*^{-1}\right)'X_*}^{(\mathrm{Var}[\boldsymbol{m}_*])^{-1}} = \tag{23}$$

$$= \underbrace{\left(\left(C^{-1}\right)'X\right)'\left(C^{-1}\right)'X}_{(\mathrm{Var}[\boldsymbol{m}])^{-1}} + \underbrace{\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}'}_{\text{positive semidefinite}} \tag{24}$$

For positive-semidefinite matrices A and B, we write that $A \succeq B$ if $\exists$ positive-semidefinite matrix C such as

$$A = B + C$$

In terms of these positive-semidefinite inequalities, we can rewrite eq.(23) as follows

$$(\mathrm{Var}[\boldsymbol{m}_*])^{-1} \succeq (\mathrm{Var}[\boldsymbol{m}])^{-1}$$

As we know from theory of positive-semidefinite inequalities [13], it means that

$$\underbrace{\mathrm{Var}[\boldsymbol{m}_*]}_{V^*} \preceq \underbrace{\mathrm{Var}[\boldsymbol{m}]}_{V}$$

So there is a precise sense in which we can say that the covariance matrix of the restricted estimator V^* is actually smaller than the covariance matrix V of the unrestricted one.

Now consider predictions that we make from these two models for some vector of regressors $\boldsymbol{x}$:

$$\begin{cases} \hat{y} = \boldsymbol{x}'\boldsymbol{m}, \quad E[\boldsymbol{x}'\boldsymbol{m}] = \boldsymbol{x}'\boldsymbol{\mu}, \mathrm{Var}[\boldsymbol{x}'\boldsymbol{m}] = \boldsymbol{x}'V\boldsymbol{x} \\ \hat{y}^* = \boldsymbol{x}'\boldsymbol{m}^*, \quad E[\boldsymbol{x}'\boldsymbol{m}] = \boldsymbol{x}'\boldsymbol{\mu}, \mathrm{Var}[\boldsymbol{x}'\boldsymbol{m}] = \boldsymbol{x}'V^*\boldsymbol{x} \end{cases}$$

We know that $V^* \preceq V$. We also assume that $\tilde{\boldsymbol{x}} \not\equiv \boldsymbol{0}$ in eq.(23), that is the constraint is informative. Algebraically, it means that

$$\begin{cases} \forall \boldsymbol{x}: \quad \boldsymbol{x}'V^*\boldsymbol{x} \leq \boldsymbol{x}'V\boldsymbol{x} \\ \exists \boldsymbol{x} \text{ such that } \boldsymbol{x}'V^*\boldsymbol{x} < \boldsymbol{x}'V\boldsymbol{x} \end{cases} \tag{25}$$

Denote y a **true value** at test data point. That is, the test data point itself is going to be a noisy measurement of this true value:

$$y_t = y + \eta$$

Denote $\boldsymbol{x}$, $\boldsymbol{z}$, $\boldsymbol{w}$ corresponding observables. According to the famous equation for expected MSE [14], the MSE between the true value and predicted value for the unrestricted estimator is

$$E[\mathrm{MSE}_U|\boldsymbol{x}] = E[\hat{y} - y]^2 = \mathrm{bias}^2 + \mathrm{Var}[\hat{y}] = \boldsymbol{x}'V\boldsymbol{x}$$

since given our assumption about independence of residuals and regressors, the GLS estimator is unbiased, so bias = 0.

Similarly, expected MSE of the restricted estimator is

$$E[\mathrm{MSE}_R|\boldsymbol{x}] = \boldsymbol{x}'V^*\boldsymbol{x}$$

Taking into account eq.(25), we get that

$$\begin{cases} \forall \boldsymbol{x}: E[\mathrm{MSE}_R|\boldsymbol{x}] \leq E[\mathrm{MSE}_U|\boldsymbol{x}] \\ \exists \boldsymbol{x} \text{ such that } E[\mathrm{MSE}_R|\boldsymbol{x}] < E[\mathrm{MSE}_U|\boldsymbol{x}] \end{cases}$$

So assuming not pathological data generation mechanism for $\boldsymbol{x}$, that is it can possibly generate $\boldsymbol{x}$ such as inequality holds in eq.(25), then it is clear that

$$\underbrace{E_{\boldsymbol{x}}\left[E[\mathrm{MSE}_R|\boldsymbol{x}]\right]}_{E[\mathrm{MSE}_R]} < \underbrace{E_{\boldsymbol{x}}\left[E[\mathrm{MSE}_U|\boldsymbol{x}]\right]}_{E[\mathrm{MSE}_U]}$$

Thus,

$$E[\mathrm{MSE}_R] < E[\mathrm{MSE}_U]$$

So we proved that a single additional observation reduces E[MSE]. We apply this idea inductively and conclude that adding an additional observation can only reduce E[MSE]. Thus, introduction of multiple information on aggregate ratings can only reduce E[MSE].

Q.E.D.

6. CONCLUSIONS

In this paper, we replaced the Bayesian approach previously deployed in [3, 5] with a corresponding frequentist estimation method Feasible GLS (FGLS) and demonstrated how aggregate ratings can be used to produce additional constraints on the parameters of the FGLS model. We also showed that these additional constraints reduce rating estimation errors of the FGLS model resulting in theoretically better rating estimation methods, thus demonstrating how aggregate ratings can improve individual recommendations.

The main issue with the FGLS method is that it works mainly on small to medium-sized problems because of the difficulty with inversion of matrix $\hat{\Omega}$ for large problems. Therefore, as a future research, we plan to work on developing more scalable methods for estimating the FGLS and other types of frequentist estimation models that work well for large problems. Also, the next step in our research would be to test our theoretically-based conclusions about superior performance of the constrained models on real data and try to show that empirical results confirm our theoretical analysis. Finally, we intend to extend Proposition 1 from the FGLS to more general types of estimators.

7. REFERENCES

[1] G. Adomavicius, R. Sankaranarayanan, S. Sen, and A. Tuzhilin. Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Trans. on Inf. Systems*, 23(1), 2005.

[2] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. on Knowledge and Data Engineering*, 17(6), 2005.

[3] A. Ansari, S. Essegaier, and R. Kohli. Internet recommendations systems. *Journal of Marketing Research*, 37(3), 2000.

[4] J. Bollen. Group user models for personalized hyperlink recommendations. *Procs of the International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, 2000.

[5] M. Condliff, D. Lewis, D. Madigan, and C. Posse. Bayesian mixed-effects models for recommender systems. *ACM SIGIR'99 Workshop on Recommender Systems: Algorithms and Evaluation*, 15(5), 1999.

[6] A. Gelfand and A. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410), 1990.

[7] A. Gelfand, A. Smith, and T. Lee. Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association*, 87(418), 1992.

[8] W. Greene. *Econometric Analysis.* Prentice Hall, 2002.

[9] R. Kimball. *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling.* John Wiley and Sons, Inc., 2002.

[10] Y. C. Kwong. *Annual Review of Scalable Computing.* Singapore University Press, 2001.

[11] M. O'Connor, D. Cosley, J. A. Konstan, and J. Riedl. Polylens: A recommender system for groups of users. *Procs of ECSCW*, 2001.

[12] S. W. Raudenbush and A. S. Bryk. *Hierarchical Linear Models: Applications and Data Analysis Methods.* Sage Publications, Inc, 2001.

[13] R. Bhatia. *Positive definite matrices.* Princeton University Press, 2007.

[14] W. Härdle. *Nonparametric and Semiparametric Models.* Springer, 2004.

[15] A. Gelman, J. Carlin, H. Stern, D. Rubin. *Bayesian Data Analysis.* Chapman & Hall/CRC, 2003.