Instant Foodie: Predicting Expert Ratings From Grassroots

Chenhao Tan*
Cornell University
Ithaca, NY
chenhao@cs.cornell.edu

Ed H. Chi
Google Inc.
Mountain View, CA
edchi@google.com

David Huffaker
Google Inc.
Mountain View, CA
huffaker@google.com

Gueorgi Kossinets
Google Inc.
Mountain View, CA
gkossinets@google.com

Alexander J. Smola
Google Inc.
Mountain View, CA
alex@smola.org

ABSTRACT

Consumer review sites and recommender systems typically rely on a large volume of user-contributed ratings, which makes rating acquisition an essential component in the design of such systems. User ratings are then summarized to provide an aggregate score representing a popular evaluation of an item. An inherent problem in such summarization is potential bias due to raters' self-selection and heterogeneity in terms of experiences, tastes and rating scale interpretations. There are two major approaches to collecting ratings, which have different advantages and disadvantages. One is to allow a large number of volunteers to choose and rate items directly (a method employed by e.g. Yelp and Google Places). Alternatively, a panel of raters may be maintained and invited to rate a predefined set of items at regular intervals (such as in Zagat Survey). The latter approach arguably results in more consistent reviews and reduced selection bias, however, at the expense of much smaller coverage (fewer rated items).

In this paper, we examine the two different approaches to collecting user ratings of restaurants and explore the question of whether it is possible to reconcile them. Specifically, we study the problem of inferring the more calibrated Zagat Survey ratings (which we dub "expert ratings") from the user-contributed ratings ("grassroots") in Google Places. To achieve this, we employ latent factor models and provide a probabilistic treatment of the ordinal ratings. We can predict Zagat Survey ratings accurately from ad hoc user-generated ratings by employing joint optimization. Furthermore, the resulting model show that users become more discerning as they submit more ratings. We also describe an approach towards cross-city recommendations, answering questions such as "What is the equivalent of the Per Se[1] restaurant in Chicago?"

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Data Mining; H.3.m [**Information Storage and Retrieval**]: Miscellaneous; J.4 [**Social and Behavioral Sciences**]: Miscellaneous

Keywords

Restaurant ratings; Google Places; Zagat Survey; Latent variables

1. INTRODUCTION

The aggregate or average scores of user ratings carry significant weight with consumers, which is confirmed by a considerable body of work investigating the impact of online reviews on product sales [5, 7, 6, 11, 23, 41, 24]. For instance, Luca [24] finds that a one-star increase in Yelp rating leads to 5-9% increase in revenue; Ye et al. [41] show positive correlation between average rating and hotel online bookings. In a similar vein, Chevalier and Mayzlin [7] demonstrate that the relative market share of a given book across amazon.com and bn.com is related to the differences across the two sites in the average star rating of the reviews.[2]

Due to the high visibility and salience of aggregate scores, acquisition of a large number of reliable user ratings becomes a critical task in building a recommender service. An inherent problem in user-contributed ratings (and the derived item-level aggregate scores) is potential bias due to two sources. First, raters choose items according to their personal interests (e.g. which restaurants they visit) and they also decide which items they rate. As a result, many raters only have experience with a limited set of items, and a particular item may be rated preferentially by a certain, non-independent group of users. Second, users have different tastes and different notions of what constitutes a good item or experience. They also differ in their understanding and familiarity with the evaluation process (e.g. the meaning of "four stars" on the commonly used five-star rating scale can be different to different raters). In this paper, we focus on restaurant ratings, examining two different approaches to acquiring user ratings, and attempt to address the problem of rater bias that is inherent to user-contributed data.

The first popular strategy we consider here is ad hoc data collection, where users voluntarily submit reviews of places they know, which is employed by Yelp[3] and Google Places[4]. An advantage of this approach is that it allows a system to gather a large number of

[*] This work was done while CT was an intern at Google.

[1] Per Se is a famous high-end restaurant in New York City.

[2] We use the terms reviews and ratings interchangeably in this paper.

[3] www.yelp.com

[4] www.google.com/places

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2263-8/13/10 ...$15.00.
http://dx.doi.org/10.1145/2505515.2505712.

ratings for many businesses relatively quickly. The problem with this approach is twofold. First, such voluntarily contributed ratings tend to exhibit a higher *self-selection bias*. Users are more likely to rate places where they have had either a great or a terrible experience, which is known as the brag-and-moan phenomenon [16]. Ratings are typically submitted *ad hoc*, soon after users visit a restaurant, which may amplify rating bias. Second, users have different motivations to submit ratings. Some users see reviews as part of their online persona, which affects which places they rate publicly [13]. In other words, some might choose to only rate high-end restaurants, while others, for example, may prefer and review only burger joints or bars. Compounding these problems is that some businesses have been known to hire people to write fake positive reviews which may further increase rating bias [27].

An alternative approach is to maintain a panel of raters that are invited to review a predetermined set of restaurants at regular intervals. Among review providers that employ this approach are the Michelin Guide[5] and Zagat Survey[6]. Focusing on a predetermined set of restaurants mitigates the effects of user self-selection bias; and maintaining a panel of raters makes the motivations to rate more consistent and objective. Repeated surveys at regular intervals decrease the variance in users' evaluations. However, the panel approach achieves the higher quality of ratings at the expense of much smaller coverage. Usually, an inclusion of a restaurant in Zagat Survey is already a sign of popularity or distinction.

Goals and Contributions

In this paper, we examine two datasets representing two different approaches to collecting user ratings of restaurants. We obtained access to the user-contributed ("grassroots") ratings from Google Places (GP)[7] and the aggregate "expert" scores from Zagat Survey, which are published in the Zagat guides and on the web. We set out to understand and reconcile these different approaches. An interesting research question is whether the aggregate scores from the two approaches correlate with each other and whether it is possible to infer expert scores using "grassroots" data.

It somewhat complicates our task that GP ratings are one-dimensional (reflecting overall quality) whereas Zagat ratings for restaurants distinguish between three different aspects: food, decor, and service. Therefore, a naive approach, consisting of rescaling the average ratings in GP, is clearly insufficient. We need a method for inferring different Zagat quality dimensions from overall GP ratings.

To solve this problem, we employ latent factor models from collaborative filtering. We are inspired by the matrix factorization approach, which is commonly used in state-of-art collaborative filtering. Our strategy is to use the latent features for restaurants, obtained from GP ratings, as feature vectors to regress on Zagat scores. More specifically, we add a virtual Zagat food reviewer, a decor reviewer and a service reviewer to the modeling process. To improve accuracy we perform *joint* optimization over GP ratings and Zagat ratings.

Since GP ratings are *discrete*, we use an exponential family model to handle the discrete ordinal data. Furthermore, we explicitly allow for different users and different places to have varying degrees of rating uncertainty (e.g. some restaurants might be mediocre but consistently so). Finally, we control for the effects of location, cuisine, and price to infer Zagat ratings based on reviews posted on GP.

To summarize, we investigate the problem of predicting three-dimensional Zagat scores from one-dimensional user-contributed ratings in Google Places. We conduct an extensive comparison of different models and validate that joint optimization on expert Zagat ratings and "grassroots" GP ratings improves the performance both in terms of RMSE (root-mean-square error) and correlation measures. We show that the model is able to reconcile the two different approaches reasonably well.

There are a number of other interesting findings from our results: (1) It turns out to be considerably easier to infer food and service scores than decor scores from GP ratings. That suggests that Google Places users tend to care more about food and service than decor when they rate.

(2) The results on user bias suggest that more experienced users tend to be more discerning.

(3) We also find that decor and service scores consistently improve with price level, while the same only holds for food scores in the more expensive price categories, thus pointing to decor and service as the key differentiators between moderate and inexpensive restaurants.

Outline

Section 2 summarizes related work. Section 3 provides a description of the data used in the experiments and formulates the main problem. Section 4 introduces statistical models explored in this paper and describes the associated inference algorithms. Section 5 describes experiment setup and evaluation measures. Finally, Section 6 presents some empirical results; and we conclude with Section 7.

2. RELATED WORK

In this paper, we investigate how to aggregate noisy user-contributed GP ratings to predict Zagat expert ratings. In this context, the two most relevant fields of research are crowdsourced labeling and collaborative filtering.

2.1 Crowdsourced Labeling

Recently, massively distributed data collection techniques have become quite popular thanks to the advances in online commerce and growth of the Internet audience. Crowdsourcing services such as Amazon's Mechanical Turk allow researchers to perform various labeling tasks. Researchers have studied the challenging problem of how to obtain accurate labels from groups of volunteer or paid workers [31, 10, 40]. A popular approach is to model the accuracy of individual labelers, their aptitude for a specific task, and the difficulty of the task.

The problem is inherently related to the literature on standardized tests, particularly Item Response Theory [30]. Dawid and Skene [9] develop a method to handle polychtomous latent class variables. Whitehill et al. [40] simultaneously estimate the true label, item difficulty, and coder expertise. In our problem, the difficulty of rating a place can be interpreted as item difficulty in Item Response Theory, even though the sources of variability might be quite different, i.e. a restaurant might simply serve food of varying quality as opposed to food that is difficult to assess.

[5] www.michelintravel.com/michelin-guides/

[6] www.zagat.com. Zagat restaurant guides were named as "a necessity second only to a valid credit card" by the New York Times. (http://www.nytimes.com/1997/11/12/dining/zagat-s-new-york-survey-entries-up-contributors-off.html)

[7] In May 2012 Google Places was succeeded by a new product called Google+ Local, see http://googleblog.blogspot.com/2012/05/localnow-with-dash-of-zagat-and.html. We performed our analysis using data acquired prior to the launch of Google+ Local and hence we use the former product name i.e. Google Places.

Active learning is also employed to study how to better make use of labelers [45, 35]. Sheng et al. [35] use relabeling to obviate the effects of noise and Vijayanarasimhan et al. [38] identify promising crowdsourcing annotations given a limited budget. However, our setting is somewhat different, since we have a large number of user-contributed ratings from GP. Our main goal is to make better use of the datasets we already have.

Sample selection bias is an important problem in different disciplines, such as economics, sociology and machine learning. The seminal paper by Heckman [14] analyzes a two stage approach. In the first stage, the probability of selection is estimated and then the probability is used as an explanatory variable in second stage. However, it is difficult to estimate the probability of selection without additional features and labels in our setting. We try to consider closeby places in a variation of our model to mitigate the effects of selection bias.

2.2 Collaborative Filtering

Collaborative filtering (CF) is a leading approach to building recommender systems [33]. Koren and Bell [20] give a comprehensive overview on the state-of-art techniques. The data provided by GP falls squarely into the category of recommendation problems.

One of the most successful techniques is latent factor modeling, partially due to strong theoretical justifications for distributions on strongly exchangeable random variables: The Aldous-Hoover theorem [3, 15] states that matrix-valued random variables that are invariant under row and column permutations must follow a latent variable model.

A popular strategy, which we adopt in this paper, is to characterize both items and users as vectors in a space automatically inferred from observed ratings. In our problem, restaurants correspond to “items” in collaborative filtering. Since Zagat ratings are in three dimensions, different parts of the latent space for restaurants could be related to food, decor and service respectively.

One of the most successful realizations of latent factor models is based on matrix factorization [39, 19, 42, 32, 17, 2, 29, 34]. These methods have become popular in recent years because of good scalability with predictive accuracy. It is worth noting that SVD++ in [20] adds smoothing vectors to user vectors to mitigate selection bias. Studies such as [21] explore the ordinal property of user-contributed ratings. Koren [18] also considers collaborative filtering with temporal dynamics.

A few recent studies [22, 28, 44] are concerned with transferring information between several domains of observation. This is related to our goal of inferring Zagat ratings from the user-contributed GP ratings. However, in our problem, the two datasets are collected by different processes (ad hoc rating collection vs. planned surveys) and employ different scales. Moreover, the Zagat rating data effectively amounts to three highly different dimensions, with ratings available as one set of summary scores per place, i.e. as a triple like (food, decor, service). Hence the approach of Aizenberg et al. [2] does not directly apply. In their music recommendation system, they assume access to a very large number of playlists, considerably in excess of the latent representation dimensionality.

Notably, the work of Umyarov and Tuzhilin [36] present a framework for incorporating *aggregate* rating information for improving *individual* recommendation, which addresses the opposite direction of our problem. [2] also presents similar thoughts.

3. PROBLEM DESCRIPTION

After the above informal discussion of previous research and a general introduction to integrating two disparate data streams, we now provide a formal description and discuss properties of the data in greater detail.

3.1 Data

We focus our study on two datasets of ratings of US restaurants obtained respectively from Google Places and Zagat. The GP data is a random sample of 2 million user-contributed reviews collected between 2009 and 2011 in the range of $\{1, 2, 3, 4, 5\}$. Users voluntarily submit ratings for places that they visited. This is common practice in online recommendation sites such as Yelp and IMDb. In this approach the aggregate scores tend to have a larger bias, for two reasons. First, raters self-select into providing the ratings, and thus are more likely to "brag and moan," which leads to bimodal or J-shaped rating distributions [16]. Submitting ratings shortly after the dining experience amplifies the "brag-and-moan" effect and increases variance. Second, there is considerable rater heterogeneity with respect to their motivations to contribute ratings.

As mentioned in the introduction, Zagat takes a different approach by effectively employing a panel of raters and inviting them to rate a predetermined set of restaurants at regular intervals[8]. Since there is a pre-selected set of restaurants[9] evaluated post factum (i.e. not immediately after the dining experience), the bias of self-selection and “brag-and-moan” is somewhat mitigated in the process. Repeated surveys also decrease the variance in users’ experience. This allows for more calibrated and less bimodal scores than user-contributed ratings. Finally, Zagat ratings are collected separately for food, decor and service on a 4-point Likert scale $\{0, 1, 2, 3\}$. For each restaurant, the average of all the ratings for each dimension is computed after filtering out untrustworthy users. This yields scalars in the interval $[0, 3]$, which we try to predict in our experiments. Zagat presents these dimensional scores by multiplying them by 10, i.e. on the characteristic 30-point scale in Zagat guides. Zagat data also contains a prevalence of relatively high food ratings. This is a direct consequence of the fact that an inclusion of a restaurant in Zagat survey is a sign of distinction or popularity. We use the ratings of all restaurants in Zagat survey.

An important aspect of the review data is that the noise is highly heteroscedastic [43]. That is, the variance of ratings is dependent on the actual score and additional covariates. As we will show in the experiments, it is helpful to model the variance of each restaurant explicitly. The variance vanishes for restaurants with low and high aggregate ratings. This is not too surprising given that there are upper and lower limits to the scores that can be assigned. For instance, an excellent restaurant will have all reviewers agreeing on the excellence: Since there is nothing better than 3 or 5 points respectively in Zagat or GP that can be assigned, the variance will vanish. Interestingly, after controlling for scale differences, Zagat user ratings have a smaller standard deviation than the GP user ratings, suggesting that Zagat contributors are indeed more consistent.

3.2 Objective

Our goal is to predict Zagat expert ratings of restaurants based on user-contributed ratings in GP. In the rest of this paper, we use the formal variable definitions in Table 1.

For a given place p, our goal is to estimate the three aspects of a Zagat rating s_{pz} based on the Zagat ratings of *other* places and the available GP ratings s_{pr} for place p. Clearly this task would be impossible if we had no access to Zagat ratings, since the latter are

[8] We had no control over how the Zagat scores were collected.

[9] In this paper, we refer to this set of restaurants as Zagat places, and restaurants not in this set as non-Zagat places.

Table 1: Summarization of variable definitions.

$\{f, d, s\}$	food, decor, service (subscripts)
z	indicator for Zagat
p	index variable for a place
r	index variable for a rater
$s_{pz} = \{s_{pzf}, s_{pzd}, s_{pzs}\}$	Zagat scores for place p
s_{pr}	GP rating for place p by rater r
$\mathcal{P}_z$	Places with Zagat ratings
$\mathcal{P}_r$	Places rated by rater r
$\mathcal{C}_p$	Places close to p (including p)
u_p	Factor for place p
u_{city}	Factor for city
u_{cat}	Factor for category
$u_{\$}$	Factor for price
v_r	Factor for rater r
v_p	Smoothing factor for place p
v_{rated}	Factor if place is Zagat rated
$V_z = \{v_{zf}, v_{zd}, v_{zs}\}$	Zagat food, decor, service factors
b_r	Bias for rater r
b_p	Bias for place p
$b_z = \{b_{zf}, b_{zd}, b_{zs}\}$	Bias for Zagat food, decor, service
b	Common bias for GP
τ	Uniform precision
τ_r	Precision for rater r
τ_p	Precision for place p

needed for calibration purposes. Formally, we aim to estimate:

$$s_{pz} | \{s_{pr}\} \cup \{s_{p'r}\} \cup \{s_{p'z}\} \text{ where } p \notin \mathcal{P}_z \text{ and } p' \in \mathcal{P}_z. \quad (1)$$

Our method of choice is a latent factor model with bias. In a nutshell, such models use inner products and bias $\langle u, v\rangle + b$ to infer the rating of a place. We assess the fidelity of the estimates both by reporting the root-mean-square error (RMSE), i.e. the square root of the average squared deviation between estimates and true ratings. In the experiment on GP ratings only, we also compute the data log-likelihood to see the effects of different variations of models. While the former amounts to a quantifiable prediction error, the latter specifies how good our model is in representing the distribution of the scores, which can lead to a better representation of place vectors.

4. MODEL

The statistical model for inferring ratings for both GP and Zagat consists of three key components:

- A hierarchical inner product model with offsets to capture the latent rating of a place by a user.
- A Gaussian or quadratic ordinal emissions model to reconcile observations with the latent variable.
- A Gaussian prior on the latent variables.

These components are combined to obtain an estimate of the likelihood of the data. We use a maximum-a-posteriori estimate for inference. To address concerns of computational efficiency we use joint stochastic gradient descent, as is common in collaborative factorization methods [20]. In the following section, we give a description of the statistical components and how the model is applied to GP and Zagat ratings.

4.1 Statistical Building Blocks

Inner Product. We assume that for each rating s_{pz} or s_{pr} there exists a latent score y_{pz} or y_{pr} respectively that can be obtained by means of an inner product model between attributes specific to a rater and attributes specific to a place. That is, we assume

$$\begin{aligned} y_{pr} &= \bar{b}_{pr} + \langle \bar{u}_p, \bar{v}_r \rangle && && \text{(2a)} \\ \text{where } \bar{b}_{pr} &= b + b_p + b_r && \text{(bias)} && \text{(2b)} \\ \bar{u}_p &= u_p + u_{\text{city}} + u_{\text{cat}} + u_{\$} && \text{(place factor)} && \text{(2c)} \\ \bar{v}_r &= v_r + |\mathcal{P}_r|^{-\frac{1}{2}} \sum_{p' \in \mathcal{P}_r} v_{p'} && \text{(rater factor).} && \text{(2d)} \end{aligned}$$

Here Eq. (2a) is a standard factorization model with bias. To model the *bias*, Eq. (2b) follows from the assumption that the biases for a given rating are additive between raters and places, and furthermore, we want to take a common rating bias into account. In other words, we effectively perform row-wise and column-wise mean removal on the rating matrix.

The *place factor* $\bar{u}_p$ decomposes hierarchically Eq. (2c) based on the side information available at prediction time. That is, provided that we know the location, category, and price level of a place, it is reasonable to assume that these terms should affect the latent attribute space of a place.

Finally, for the *rater factor*, Eq. (2d) takes selection bias between raters into account by assuming that raters visiting similar places should share common preferences. This is a direct application of the model described by Bell and Koren [20], with an application of hierarchical models from [1].

Emissions model. We now need to connect the latent score y_{pr} to the observed scores s_{pr}. If we ignore the fact that s_{pr} is actually discrete (in the case of Zagat it is the average of a finite number of curated ratings and thus continuous, while for GP it can only take 5 distinct values), we obtain $s_{pr} \sim \mathcal{N}(y_{pr}, \tau^{-1})$, i.e.

$$-\log p(s_{pr}|y_{pr}) = \frac{\tau}{2}(s_{pr} - y_{pr})^2 + \frac{1}{2}\log 2\pi - \frac{1}{2}\log \tau. \quad (3)$$

Note that for notational convenience we parametrize the Gaussian model in terms of its *precision* τ rather than variance $\sigma^2 = \tau^{-1}$. The larger τ, the lower the noise that we assume in the estimation process.

A more principled approach to modeling discrete data is to employ a properly normalized exponential family model. Denote by $\mathcal{Y} = \{1, \ldots, 5\}$ the ordinal range set of ratings. In this case we may replace the normalization in (3) by:

$$-\log p(s_{pr}|y_{pr}) = \frac{\tau}{2}(s_{pr} - y_{pr})^2 - \log \sum_{k=1}^{5} e^{-\frac{\tau}{2}(k-y_{pr})^2}. \quad (4)$$

The advantage of the above model is that it is capable of modeling rating distributions for both excellent and poor places, simply by picking $y_{pr} > 5$ and $y_{pr} < 1$ respectively. These choices concentrate the distribution more towards the extreme ranges of ratings. In the experiment we show that this choice does lead to a better likelihood estimate.

Note that, unlike in (3), the expectation of the renormalized Gaussian distribution does not satisfy $\mathbf{E}[s_{pr}|y_{pr}] = y_{pr}$. Instead, we compute the expectation via:

$$\mathbf{E}[s_{pr}|y_{pr}] = \frac{\sum_{k=1}^{5} k e^{-\frac{\tau}{2}(k-y_{pr})^2}}{\sum_{k=1}^{5} e^{-\frac{\tau}{2}(k-y_{pr})^2}}. \quad (5)$$

This is also what we use to verify the accuracy of the estimates in an RMSE sense. Note that (4) has a unique minimum and can be reparameterized convexly.

Priors. To complete the model specification, the final piece required is to discuss the priors on the latent variables u, v, the biases b, and the precisions τ. We impose a Gaussian prior on the first two latent variables, b_r and b_p, a flat (improper) prior on the bias b, and a Gamma prior on the precisions. These choices are in the spirit of SVD++ [20].

More specifically, we assume that:

$$\begin{aligned} & b_r, b_p, u_p, u_{\text{city}}, u_{\text{cat}}, u_{\$}, v_r, v_p, v_{\text{rated}}, \\ & \qquad v_{zf}, v_{zd}, v_{zs} \sim \mathcal{N}(0, \lambda^{-1}\mathbf{1}). \end{aligned} \tag{6}$$

In other words, we assume that *all* the parameters are drawn from a Normal distribution with precision λ. For simplicity, we use the same precision λ to avoid a large number of parameters.

Furthermore, we model the common bias as a scalar, normally distributed according to $b \sim \mathcal{N}(0, \lambda'^{-1})$. Here λ' denotes the precision of the biases. For an improper prior on the latter set $\lambda' \to 0$, the same as in [20]. With this improper prior, b can be computed by averaging across all the ratings.

Finally, the precisions τ are drawn from a Gamma distribution. Note that one design choice is to model the precision of each place individually. We have:

$$-\log p(\tau_p|\alpha, \beta) = \beta\tau_p - (\alpha - 1)\log\tau_p - \alpha\log\beta + \log\Gamma(\alpha).$$

Note that the hyperparameters α and β are fixed. Thus, for inference purposes we are only concerned with the contribution of $\beta\tau_p - (\alpha - 1)\log\tau_p$ to the log-likelihood.

4.2 Google Places

To predict the rating for a user, we can treat the problem similar to a classical collaborative filtering problem. We address it by combining the components regarding $s_{pr}|y_{pr}$ (Gaussian or renormalized Gaussian), the priors on u, v, b, the priors on the precisions τ, and the choice of whether we model rating uncertainty as being specific to a user or to a place. Such a multitude of choices leads to six different models that we can explore experimentally:

SVD++ or `SVD`. We assume that $s_{pr}|y_{pr}$ is Gaussian and that all precisions are identical (all equal to 1). This is analogous to Koren's work in [17], except that in the latter precision is not modeled explicitly. We are optimizing the total square error in this case.

SVD++ Rater or `SVDRat`. The change is that we now assume that each rater has its own level of precision τ_r with associated Gamma prior.

SVD++ Place or `SVDPla`. Same as above, only now we use τ_p rather than τ_r, so each place has its own level of precision.

SVD++ Renormalized or `SVDRen`. This is the same as SVD++, only that we use the renormalized Gaussian model to take the discrete nature of the ratings into account.

SVD++ Renormalized Rater or `SVDRenRat`. As with SVD++ Rater we now model the precision of raters τ_r in the renormalized model.

SVD++ Renormalized Place or `SVDRenPla`. As above, but with precision τ_p per place in the renormalized Gaussian.

To simplify representation, we refer to these six models as `SVD`, `SVDRat`, `SVDPla`, `SVDRen`, `SVDRenRat`, `SVDRenPla` respectively in the experiments, as denoted above.

For illustration purposes we summarize the objective function for the last model. We use $\mathcal{R}$ to denote the set of reviews. Up to additive constants and perusing the inner product model of (2) the negative log-likelihood $\mathcal{L}$ is given by

$$\mathcal{L} = \sum_{(p,r)\in\mathcal{R}} \left[\frac{\tau_p}{2}(y_{pr} - s_{pr})^2 - \log\sum_{k=1}^{5} e^{-\frac{\tau_p}{2}(k-s_{pr})^2}\right] + \tag{7a}$$

$$\frac{\lambda}{2}\left[\sum_p \|u_p\|^2 + \sum_{\text{city}} \|u_{\text{city}}\|^2 + \sum_{\text{cat}} \|u_{\text{cat}}\|^2 + \sum_{\$} \|u_{\$}\|^2\right] + \tag{7b}$$

$$\frac{\lambda}{2}\left[\sum_r \|v_r\|^2 + \sum_p \|v_p\|^2 + \sum_r b_r^2 + \sum_p b_p^2\right] + \frac{\lambda'}{2}b^2 + \tag{7c}$$

$$\sum_p \left[\beta\tau_p - (\alpha - 1)\log\tau_p\right]. \tag{7d}$$

As a review of our different components, τ_p in (7) corresponds to place precision. Alternatively, we can use τ_r to consider user precision. The first part in (7a) corresponds to the simplified assumption in Eq. (3) (without additive constants), while the second part in (7a) is for the renormalization. (7b), (7c), (7d) correspond to the priors for different variables in the model.

Inference is performed by joint stochastic gradient descent in all parameters of $\mathcal{L}$ as we traverse the set of reviews and ratings $\mathcal{R}$. This is known to yield accurate results [17].

4.3 Zagat

We now proceed to integrating Zagat ratings with those obtained from GP. In this way, the latent attributes for places can be used to model not only GP but also aspects that are important for Zagat, such as food, decor and service. We achieve this goal by adding three virtual raters with associated factors v_{zf}, v_{zd} and v_{zs}. For instance, the equation

$$\langle \bar{u}_p, v_{zf} \rangle + b_f$$

amounts to an estimate of the Zagat food rating.

In learning how these attributes correlate with the observed Zagat ratings we are able to obtain an "instant foodie": A mechanism for translating the idiosyncratic, crowdsourced ratings from GP into the more consistent, better-calibrated ratings[10] of Zagat.

That said, before embarking on a full model of Zagat ratings we need to address the fact that Zagat ratings are not awarded at random. The mere fact of a restaurant being "Zagat rated" promises a modicum of quality. In other words, there is an inherent selection bias, skewing toward good restaurants. Hence we may decompose the distribution into first modeling whether a score is observed $(\text{observed}_p, p \in \mathcal{P}_z)$ and only then capture the actual value of the observation. It is similar to Heckman's correction [14]. We add a softmax term to our objective function. Note that, we only perform this correction for Zagat ratings.[11]

Using an exponential family model yields:

$$-\log p(\text{observed}_p | \{z_p\}) = -z_p + \log\sum_{p'\in\mathcal{C}_p} e^{z_{p'}} \tag{8a}$$

$$\text{where } z_p := \langle \bar{u}_p, v_{\text{rated}} \rangle. \tag{8b}$$

Here $\mathcal{C}_p$ denotes a set of closeby places in the local neighborhood of p (we choose 5 places within a 2 mile radius). This is to ensure

[10] Recall that the Zagat publishes ratings on a $[0, 30]$ scale and we rescale the loss correspondingly.

[11] We could add this to the GP ratings as well. But adding this to all the GP places can induce more noise. Compared to the differences between non-Zagat places and Zagat places, randomly selected closeby places for GP ratings (p, r) is not clearly distinguished from places that users chose to visit.

that we only compare locally-equivalent places regarding their inclusion in Zagat. v_{rated} is an additional vector to evaluate whether a place is rated in Zagat ratings.

This yields two options when including Zagat ratings: a plain version that regresses on the ratings as if they were additional users and a location calibrated version that uses Eq. (8) for debiasing.

For conciseness we use a factorial notation to describe the experiments. In (factorial) combination with the models of the previous section this yields the following grammar:

{'',Clo} × SVD × {'',Ren} × {'',Rat,Pla}.

For instance CloSVDRenPla amounts to a model considering (1) closeby places to Zagat places, (2) rating debiasing using SVD approaches, (3) the renormalized Gaussian model for GP ratings, and (4) a place-dependent precision latent variable.

4.4 Inference

Maximizing the log-posterior is relatively straightforward by employing a stochastic gradient descent procedure. Due to the nonconvexity of the objective (it is convex in factors but not jointly so) we use a rather conservative learning rate adjustment [26] via

$$\eta_t = (a + mt)^{-\frac{1}{2}} \text{ with } m = 0.01. \tag{9}$$

Here t is a counter of how many observations have already been seen. We investigate different values for a in the experiment. Moreover, for the Gamma prior we set $\alpha = \beta = 2$ to adjust shape and rate respectively. All precision variables (τ) are initialized to 1, all latent factors (u, v) are initialized to random values, and all the biases (b_p, b_r) are initialized to 0. We traverse all ratings (p, r) and a set of Zagat ratings (if applicable) to perform stochastic gradient descent updates. To ensure good convergence we traverse the space of observations 20 times (fewer would suffice but this is to ensure that we have accurate results for all settings). This leads to algorithm 1 below.

Algorithm 1 Stochastic gradient descent

```
Input  All the Google ratings and some Zagat ratings.
Output  Latent factors for places.
  Initialize counter $t \leftarrow 0$
  for d in 1:N do
    Shuffle training data;
    for r in training data do
      $\eta_t \leftarrow (a + mt)^{-\frac{1}{2}}$
      Increment $t \leftarrow t + 1$
      Update variables affected by slice $\mathcal{L}_t$ via
          $(b, u, v, \tau) \leftarrow (b, u, v, \tau) - \eta_t \partial_{(b,u,v,\tau)} \mathcal{L}_t$
    end for
  end for
```

Note that the slices $\mathcal{L}_t$ in the algorithm are essentially just parts of the negative log-likelihood $\mathcal{L}$ that are specific to a particular rating instance (p, r, s_{pr}) on GP or a corresponding triplet of Zagat rating.

5. EXPERIMENTS

We give a description of the experimental setup and evaluation measures. Then we show that our models outperform the baseline. Furthermore, we show that a joint model of Zagat ratings and GP ratings improves overall performance. Moreover, we find that place precision and exponential renormalization via a proper generative model helps in the context of predicting Zagat expert ratings. We conclude with a number of analyses and findings from the resulting model.

5.1 Experimental Setup

In the experiments below, we use 100-dimensional latent vectors throughout the experiments.

5.1.1 Estimating GP ratings

First, we want to conduct experiments on GP ratings to see how different versions of collaborative filtering algorithms perform and whether adding different components improves the performance. Also, to assess recommendation performance we need to address the fact that in the absence of additional features it is impossible to learn meaningful place vectors and predict accurately for places with very few ratings. Therefore, we test only on the latest ratings for restaurants with at least 3 reviews.

Baseline. For experiments on GP ratings, our main objective is to see how different models compare with each other. Thus SVD[12] is effectively our baseline.

Cross-validation. For experiments on GP ratings, in order to avoid bias by systematically removing a large number of recent ratings for validation purposes, we perform 5-fold cross-validation in the following way. We randomly partition the latest ratings for restaurants with at least 3 reviews into five partitions, one of which is used for testing, one for validation and three for training. Results are then averaged over the partitions. While there are considerably more advanced model selection tools from statistical learning theory available [37], the above is consistent and considerably easier to implement.

Parameters. The parameter range investigated for the GP Gaussian Priors (λ in Eq. (6)) was $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1\}$, while the range for the initial learning rate (a in Eq. (9)) was $\{5000, 10000, 50000, 100000\}$.

5.1.2 Estimating Zagat ratings

Baseline. For the Zagat ratings, we use two baselines:

Average transformation. This algorithm simply takes the average of GP ratings and then performs a linear transformation from $[1, 5]$ to $[0, 3]$. This is an extremely naive baseline for many reasons. For instance, GP ratings are one-dimensional whereas Zagat ratings are differentiated into three dimensions.

Linear regression transformation. This is a strong baseline that uses the latent attributes inferred on GP using SVD. It then estimates a linear regression for Zagat food, decor and service based on places in the training set.

Cross-validation. For Zagat estimation, we use a classical 5-fold cross-validation since no validation set is needed for parameter selection. We directly use the best parameter in the corresponding GP experiments. The best parameter is mostly consistent between different folds in our experiments on GP. Whenever the best parameters are different for different folds, we simply take the majority choice, i.e., the parameter setting that provides the most best performances in 5 folds.

5.2 Evaluation Measures

We consider three evaluation measures in the experiment. For experiments on GP ratings, we try to predict user ratings for places. We use RMSE as a straightforward metric to measure the loss between predicted scores and actual scores. We also include the log-

[12] SVD refers to SVD++ in our paper.

likelihood as our evaluation metric to see how good the model represents the distribution of user ratings. For experiments on Zagat ratings, we try to predict ratings for each place aggregately. We use RMSE in this case, too. Since we have two lists of place scores, it also makes sense to see how good they correlate with each other, thus we also include Pearson Correlation as a second metric.

In the descriptions below, we use s to denote the true rating and, with some slight abuse of notation (in the context of renormalized Gaussians) we use y for the predicted rating.

RMSE. The Root Mean Square Error describes the average difference between the true rating and the predicted rating. The smaller the RMSE, the better the performance.

$$\mathrm{RMSE} = \sqrt{n^{-1} \sum_{p,r} (s_{pr} - y_{pr})^2}.$$

Note that in many cases it is impossible to achieve $\mathrm{RMSE} = 0$ due to the inherent variance in s_{pr}.

Log-likelihood. This measures more directly the statistical fidelity of the estimate. More specifically, a large log-likelihood is good. It is defined as

$$\mathrm{LL} = \sum_{p,r} \log p(s_{pr}|y_{pr}).$$

Note that by construction LL can never exceed the negative entropy of the distribution, as follows from information theory [8]. Our reason for choosing LL is that it directly measures the fidelity of our data generation model instead of minimizing the RMSE.

Pearson Correlation. This measures the degree of linear dependence between the predicted and observed result. The larger the correlation, the better the performance. It is given by

$$\mathrm{PCorr} = \frac{\sum_i (s_i - \bar{s})(y_i - \bar{y})}{\sqrt{\sum_i (s_i - \bar{s})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}.$$

6. RESULTS

6.1 Estimating Google Places Ratings

We first assess the performance of the estimates on GP ratings, to see whether modeling user precision or place precision improves on SVD (SVD++) and whether discretizing the ratings improves performance.

As we can see from Table 2, SVD provides the best performance in terms of RMSE, while SVDRenRat, i.e. SVD++ with exponential renormalization and user precision provides the best log-likelihood estimates. Employing exponential renormalization improves the log-likelihood but leads to worse performance in RMSE since what we are optimizing in this context is not the total loss but a renormalized likelihood. As expected, the choice of models depends on the choice of objective function.

It seems that considering user precision always works better than place precision. This suggests that variability for a given user is more helpful than variability for a given place in predicting user ratings for places.

6.2 Estimating Zagat Ratings

We now describe estimation of Zagat ratings which constitutes the main task of this paper. Table 4 shows the performance of different models and the two baseline approaches.

6.2.1 *Baseline performance*

Average baseline. As expected, the average score translation from GP ratings provides the worst performance (with RMSE being high at 0.539). The correlation between average score and Zagat decor score is quite low (correlation=0.075). This suggests that when submitting the overall, single-dimensional rating, users tend to care more about food (correlation=0.375) and service (correlation=0.258) than restaurant decoration. And even for the food dimension, which has the highest correlation with the overall GP ratings, the value is low. This indicates that there are indeed substantive differences between the two sets of users and the two methodologies for collecting user evaluations.

Linear regression baseline. The linear regression baseline performs quite well in predicting food and service scores. Note that an RMSE loss of 0.259 in the 3.0-scale is just 2.59 in the usual Zagat 30-point rating. We see a clear drop in accuracy on decor compared with food and service. This confirms that we can mainly learn about food and service related features from GP ratings alone. And the similar performance on food and service actually shows that normal users have the ability to distinguish food quality as well as service quality.

6.2.2 *Joint Modeling Performance*

Jointly modeling Zagat ratings and GP ratings can further improve the performance compared to the linear regression baseline. All of the models that jointly optimize the loss on Zagat ratings and GP ratings improve both RMSE and correlation measures compared to the linear baseline.[13] This is a very encouraging result, suggesting our general approach improves Zagat rating prediction and it is possible to reconcile the two approaches to collecting user ratings.

However, the effects of considering user precision, place precision, exponential renormalization and closeby places are quite complicated:

- First, it seems that considering closeby places to Zagat places hurts the performance in terms of RMSE. This is reasonable since considering closeby places changes the objective function to optimize. However, we note that two of the three best correlation numbers (food with CloSVDRenRat and decor with CloSVDPla) are observed when considering closeby places.
- Second, exponential renormalization seems to help. This is especially clear in RMSE for cases where we consider closeby places. All the models with exponential renormalization (with prefix CloSVDRen) give better or same average RMSE compared to the corresponding model without Ren. The reason might be that though exponential renormalization does not lead to better RMSE results in experiments on GP ratings, it helps better learn the latent features for each place and leads to slightly better performance on Zagat ratings.
- Finally, we also notice that considering place precision helps more than considering rater precision. For instance, the model with the best average RMSE is SVDRenPla that considers place precision. The best numbers in service and decor are also attained when considering place precision. This is different from what we observe in the experiment on GP ratings. The reason might be that we are trying to infer the ratings at

[13] There is one exception on Food RMSE in CloSVDRat, i.e. in the model considering closeby places to Zagat places and rater precision.

Table 2: Google Places accuracy. Algorithm labels as described in Section 4.2. Note that SVD refers to SVD++ in our paper.

	RMSE	log-likelihood
SVD	**1.144**	$-5.64 \cdot 10^4$
SVDRat	1.159	$-5.36 \cdot 10^4$
SVDPla	1.166	$-5.48 \cdot 10^4$
SVDRen	1.147	$-5.55 \cdot 10^4$
SVDRenRat	1.157	$\mathbf{-5.21 \cdot 10^4}$
SVDRenPla	1.166	$-5.28 \cdot 10^4$

Table 3: Zagat RMSE loss, stratified based on price levels.

price	food	decor	service	mean
all	0.274	0.397	0.276	0.321
$	0.241	0.350	0.236	0.281
$$	0.245	0.335	0.225	0.273
$$$	0.227	0.286	0.214	0.244
$$$$	0.208	0.288	0.221	0.242

Table 4: Zagat accuracy. The labels correspond to the 12 models described in Section 4.3. The baselines are as described in Section 5.1. Note that the smaller the RMSE, the better the performance, while the larger the correlation, the better the performance.

	RMSE				Correlation		
model	food	decor	service	average	food	decor	service
average	0.416	0.683	0.482	0.539	0.375	0.075	0.258
regression baseline	0.259	0.383	0.258	0.306	0.427	0.386	0.445
SVD	**0.249**	0.347	0.240	0.283	0.507	0.561	0.567
SVDRat	0.252	0.346	0.239	0.283	0.485	0.560	0.569
SVDPla	0.250	0.348	**0.237**	0.283	0.510	0.561	**0.582**
SVDRen	0.250	0.346	0.240	0.283	0.510	0.562	0.567
SVDRenRat	**0.249**	0.347	0.240	0.283	0.513	0.555	0.566
SVDRenPla	0.250	**0.345**	0.240	**0.282**	0.504	0.558	0.567
CloSVD	0.256	0.353	0.248	0.290	0.502	0.561	0.566
CloSVDRat	0.264	0.358	0.247	0.294	0.481	0.558	0.569
CloSVDPla	0.255	0.353	0.243	0.288	0.503	**0.563**	0.580
CloSVDRen	0.253	0.349	0.244	0.286	0.511	0.560	0.564
CloSVDRenRat	0.253	0.356	0.249	0.290	**0.519**	0.557	0.568
CloSVDRenPla	0.253	0.354	0.243	0.288	0.510	0.558	0.570

place level for Zagat ratings here while we are trying to infer each user rating in the experiment on GP ratings.

As a result of the above observations, the model that provides the best average RMSE is SVDRenPla, i.e. factorization using a renormalized distribution and using place-specific precision. It improves by 0.024 in terms of average RMSE compared to the linear regression baseline, and by 0.257 against the naive average scores from GP ratings. As a reference of scale, in [17], SVD++ improves RMSE by 0.01 compared to SVD in 5 star ratings.[14]

In terms of correlation, the improvements are also significant, especially in decor. But the best correlation in different dimensions is approached by different models. This suggests that we cannot make any conclusive arguments about the power of different models. Note that the improvements relative to the baseline occur in *all three aspects*, which suggests that it is important to do joint optimization on two datasets.

When jointly modeling Zagat and GP ratings we see a larger improvement in decor compared to food and service. This suggests that GP ratings alone convey more information regarding food and service rather than decor, which is consistent with the observation in the simple average baseline.

RMSE By Price Levels. We further check the RMSE performance of estimating the Zagat ratings stratified on price levels for the model SVDRenPla, since it provides the best performance. As we can see from Table 3, there is a clear reduction in error for $$$ and $$$$ places compared to $ and $$ places: we see a drop in RMSE in all the three dimensions and the drop in decor is more significant. For example, comparing the RMSE in $$$ places to $$ places, the RMSE in decor is decreased by 0.049, while the RMSE in food is decreased by 0.018. This indicates that our model does learn that price level is a better signal for decor than for food. On the other hand, since the scores are restricted from above, the estimation problem actually becomes easier (at least on average) for pricier restaurants — they are likely to provide better food, service and decor due to efficiency of the markets.

6.3 Analysis

To investigate implications, we shall focus on the best performing statistical model, SVDRenPla. We investigate place effects such as which cuisines may be, a priori, highly and poorly rated.[15] We also investigate issues such as whether rater bias is a function of rater's experience; and whether the latent space representation is inherently meaningful by an application of cross-city search.

6.3.1 *Place effects*

We now study how place effects such as price levels and cuisine types affect the final scores in Zagat ratings. Recall that our results are based on ratings submitted by users who have chosen to patronize specific places and then submit their ratings. In other words, the findings reported herein reflect dining experience through the eyes of the users who chose to report it.

Price levels. To investigate the effects of pricing we compute the inner products between the latent price attributes $u_\$$ and the associated rating vectors v_{zf}, v_{zd}, v_{zs}. This effectively amounts to price specific offsets. Figure 1 shows how scores in food, decor and service change with price. With the exception of food for the $ or $$ price levels we see a monotone increase in each of the three dimensions. The fact that estimated food score for the $$ places is slightly lower than that in the cheaper restaurants suggests that at the lowest price level, service and decor may be the key differentiators. Alternatively, this may be due to other reporting bias that we did not account for, i.e. customers having relatively lower expectations when rating inexpensive restaurants. We find that the estimated decor score exhibits the largest increase as price increases. The difference in the decor score between the $$$$ and $ places is as high as 1.0, which corresponds to 10.0 on the usual 30-point Zagat scale. It appears that more expensive places are differentiated more by design and service rather than food (again as estimated from the ratings reported by GP users).

Cuisine types. We now check how the estimated scores vary by cuisine type. For that we first compute the inner products between u_{cuisine} and v_{zf}, v_{zd}, and v_{zs}, then sort within each dimension. Table 5 lists the top 5 cuisines types and the bottom 5 cuisine types. We observe Latin American and Korean restaurants among the top rated based on the food score, while hamburger and buffet-style restaurants appear at the bottom. Again, note that this is based on data from users who choose to patronize and rate respective restau-

[14] [1, 5] opposed to [0, 3] in our case.

[15] Some of the findings may be also observed by simply averaging, which can be seen as sanity-check. Here we are more interested in exploring the results of our model.

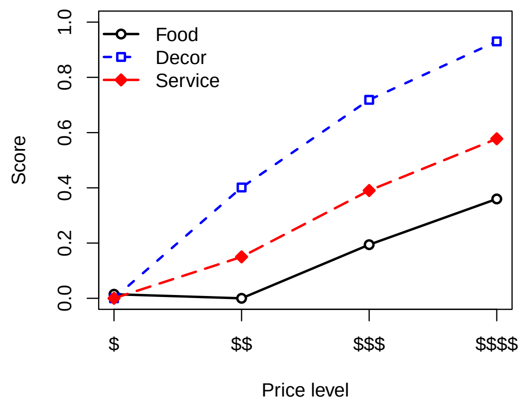

Figure 1: Price Level Effects. The numbers are all relative to the $ category. Note that the food score remains essentially unchanged for the $ vs. the $$ category.

Table 5: Best rated and lowest rated cuisine types based on latent variables.

	food	decor	service
Top tier	latin american	latin american	latin american
	korean	korean	korean
	thai	mediterranean	thai
	vegetarian	fine dining	mediterranean
	mediterranean	thai	vegetarian
Bottom tier	hamburger	hamburger	hamburger
	buffet	buffet	buffet
	seafood	fast food	fast food
	family	chicken	family
	american	chicken wings	seafood

rant types, so the results reflect the perception of those users rather than an objective consensus. For example, it is plausible that vegetarian restaurants would be rated mainly by vegetarians, who might give them higher scores than non-vegetarians. However, the fact that fast food restaurants occupy the bottom tier on decor and service but not on food provides some external validity for our analysis.

6.3.2 *User bias*

Another question on user behavior is whether rater bias is related to rater's experience. Figure 2 shows estimated user bias vs. log of the number of ratings per user.[16] As can be seen, the results are noisy but there is a clear decreasing trend, suggesting that users who give more ratings have a more negative bias. This indicates that as customers rate more places they may become less impressionable. Interestingly, Byers et al. [4] and Godes and Silva [12] respectively observe that average ratings for restaurants and books decrease over time. Our observation is consistent with this trend but from the perspective of raters. Similarly, McAuley and Leskovec [25] find that experts give more "extreme" ratings: they rate the top products more highly, and the bottom products more harshly.

6.3.3 *Most similar places*

Our use of latent variable models also makes possible what we shall call "search by example". This can come in handy when making recommendations to people who are traveling or moving to a different city. Suppose that the user is quite familiar with his or her home city and would like to find an equivalent of a particular place in the destination city. For instance, we can use the latent features to find restaurants most similar to "Gary Danko" (a well known fine dining restaurant in San Francisco) in New York City. To do so, we remove city effects from the model and fix restaurant category. After that, we use the Euclidean distance between latent vectors $u_p + u_\$$ to find the closest restaurants in the destination city to the selected restaurant in the source city.

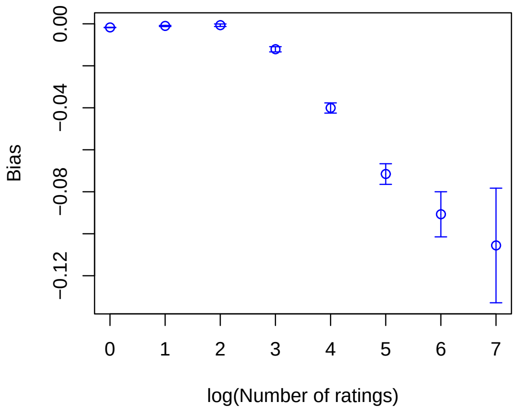

Figure 2: Rater bias as a function of the number of reviews. The error bar corresponds to 95% confidence intervals.

Table 6: Similar places in different cities based on latent variable data. SRC represents source, and DST means destination.

SRC	place	DST	place
SF	Tartine	NYC	Veniero's Pasticceria
			Amy's Bread Chelsea
			Mille-feuille Bakery Cafe
		CHI	Lou Mitchell's
			Starbucks
			Molly's Cupcakes
	Gary Danko	NYC	Jean Georges Restaurant
			Cafe Boulud
			Annisa
		CHI	Les Nomades
			Tru
			Spiaggia
NYC	Per Se	SF	Opaque
			Bix
			Frascati
		CHI	Alinea
			Joe's Seafood Home
			Girl & The Goat
	Shake Shack	SF	Denny's
			Tommy's Joynt
			Acme Burgerhaus
		CHI	Bread Basket
			Hot Dog Express
			Pockets

We show a few such examples in Table 6. We choose a couple of characteristic places from San Francisco and New York, and for each of them find the best match in a different city. We note that the matches intuitively make sense, e.g. with Per Se corresponding to Alinea. Some notable exceptions are Denny's as a purported San Francisco equivalent of Shake Shack (where one should probably expect In-N-Out Burger) and Starbucks suggested by our model as one of Chicago's answers to Tartine (a popular and well-regarded San Francisco bakery).

[16] We use logarithmic binning because the number of ratings per user is heavy-tailed. We remove data points with more than 256 ratings, because the error bars become too large to be meaningful.

7. CONCLUSION

We have discussed the problem of inferring expert Zagat-style three-dimensional restaurant ratings based on noisy user-contributed one-dimensional ratings from Google Places. Inspired by research in collaborative filtering, we employ a latent factor model to link Zagat ratings with GP ratings.

Joint optimization over the two datasets can indeed improve the performance in terms of both RMSE and Pearson correlation compared to the baseline. Curiously, we find that the improvement is more prominent in estimating decor than in food and service scores. This indicates that user-contributed ratings in GP are more likely to reflect the quality of food and service. Without the joint optimization, most latent features do not provide information about decor.

We have explored a number of variations on our model. Exponential renormalization leads to better performance in terms of log-likelihood in GP ratings and also the best performance in terms of average RMSE for Zagat ratings. It validates the effectiveness of considering the ordinal rankings using exponential family models. Based on different evaluation measures, the best performance was generally achieved by incorporating either user precision or place precision. This suggests that in similar applications it may be useful to explicitly model rating variance within users and places.

In general, we have shown that it is possible to reconcile the two quite different approaches to collecting user ratings. Using the noisier user-contributed ratings from Google Places, we are able to infer Zagat-style expert ratings reasonably well. This suggests that it may be a good idea to combine these two approaches (ad hoc rating collection and planned surveys) in the design of recommender systems so that a large number of ratings can be collected and then transformed to aggregate scores of better quality.

Posterior analysis also suggests some interesting research problems in user behavior analysis. For instance, we observe that as users submit more ratings, they tend to become more discerning overall. It is not clear whether this phenomenon holds in different application domains; there may be multiple behavioral explanations. Therefore, a better understanding of the cognitive, social, and technological processes that drive the production of user-contributed ratings is necessary for designing better recommendation platforms.

Acknowledgements. We thank Lillian Lee, Steven Scott, and Andrew Tomkins for helpful comments and discussions. We thank Jessa Mittleman for suggesting the idea of "search by example". Chenhao Tan is partially supported by NSF grant IIS-0910664 outside of the internship.

8. REFERENCES

[1] A. Ahmed, A. J. Smola, C. Teo, and V. Vishwanathan. Fair and balanced: Learning to present news stories. In *WSDM*, 2012.

[2] N. Aizenberg, Y. Koren, and O. Somekh. Build your own music recommender by modeling internet radio streams. In *WWW*, 2012.

[3] D. Aldous. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598, 1981.

[4] J. W. Byers, M. Mitzenmacher, and G. Zervas. The groupon effect on yelp ratings: a root cause analysis. In *EC*, 2012.

[5] P. Chatterjee. Online Reviews - Do Consumers Use Them? *Advances in Consumer Research*, pages 129–134, 2001.

[6] P.-Y. S. Chen, S. Wu, and J. Yoon. The impact of online recommendations and consumer feedback on sales. In *ICIS*, 2004.

[7] J. A. Chevalier and D. Mayzlin. The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, 43(3):345–354, 2006.

[8] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley and Sons, New York, 1991.

[9] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics*, 28:20–28, 1979.

[10] O. Dekel and O. Shamir. Vox populi: Collecting high-quality labels from a crowd. In *COLT*, 2009.

[11] C. Forman, A. Ghose, and B. Wiesenfeld. Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research*, 19:291–313, 2008.

[12] D. Godes and J. C. Silva. Sequential and temporal dynamics of online opinion. *Marketing Science*, 31, 2012.

[13] E. Goffman. *The Presentation of Self in Everyday Life*. Anchor, 1959.

[14] J. J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–61, 1979.

[15] D. Hoover. Relations on probability spaces and arrays of random variables. *Preprint, Institute for Advanced Study, Princeton, NJ*, 1979.

[16] N. Hu, P. A. Pavlou, and J. Zhang. Can online reviews reveal a product's true quality?: empirical findings and analytical modeling of online word-of-mouth communication. In *EC*, 2006.

[17] Y. Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *KDD*, 2008.

[18] Y. Koren. Collaborative filtering with temporal dynamics. In *KDD*, 2009.

[19] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *IEEE Computer*, 42(8):30–37, 2009.

[20] Y. Koren and R. M. Bell. Advances in collaborative filtering. In *Recommender Systems Handbook*, pages 145–186. Springer, 2011.

[21] Y. Koren and J. Sill. Ordrec: an ordinal model for predicting personalized item rating distributions. In *RecSys*, 2011.

[22] B. Li, Q. Yang, and X. Xue. Transfer learning for collaborative filtering via a rating-matrix generative model. In *ICML*, 2009.

[23] Y. Liu. Word-of-mouth for movies: Its dynamics and impact on box office revenue. *Journal of Marketing*, 70, 2006.

[24] M. Luca. Reviews, reputation, and revenue: The case of yelp.com. Harvard business school working papers, Harvard Business School, 2011.

[25] J. McAuley and J. Leskovec. From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews. In *WWW*, 2013.

[26] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[27] M. Ott, C. Cardie, and J. Hancock. Estimating the prevalence of deception in online review communities. In *WWW*, 2012.

[28] W. Pan, E. W. Xiang, N. N. Liu, and Q. Yang. Transfer learning in collaborative filtering for sparsity reduction. In *AAAI*, 2010.

[29] A. Paterek. Improving regularized singular value decomposition for collaborative filtering. In *Proceedings of KDD Cup and Workshop*, 2007.

[30] G. Rasch. *Probabilistic Models for Some Intelligence and Attainment Tests*. Danish Institute for Educational Research, Copenhagen, 1960.

[31] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *Journal of Machine Learning Research*, 11:1297–1322, 2010.

[32] J. Rennie and N. Srebro. Fast maximum margin matrix factoriazation for collaborative prediction. In *Proc. Intl. Conf. Machine Learning*, 2005.

[33] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors. *Recommender Systems Handbook*. Springer, 2011.

[34] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Incremental singular value decomposition algorithms for highly scalable recommender systems. In *Fifth International Conference on Computer and Information Science*, 2002.

[35] V. S. Sheng, F. Provost, and P. G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *KDD*, 2008.

[36] A. Umyarov and A. Tuzhilin. Using external aggregate ratings for improving individual recommendations. *ACM Trans. Web*, 5(1):3:1–3:40, 2011.

[37] V. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, New York, 1998.

[38] S. Vijayanarasimhan, P. Jain, and K. Grauman. Far-sighted active learning on a budget for image and video recognition. In *CVPR*, 2010.

[39] M. Weimer, A. Karatzoglou, Q. Le, and A. J. Smola. Cofi rank - maximum margin matrix factorization for collaborative ranking. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*. MIT Press, Cambridge, MA, 2008.

[40] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. R. Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *NIPS*, 2009.

[41] Q. Ye, R. Law, B. Gu, and W. Chen. The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. *Computers in Human Behavior*, 27(2):634–639, 2011.

[42] K. Yu, S. Zhu, J. Lafferty, and Y. Gong. Fast nonparametric matrix factorization for large-scale collaborative filtering. In *SIGIR*, 2009.

[43] M. Yuan and G. Wahba. Doubly penalized likelihood estimator in heteroscedastic regression. Technical report, University of Winconsin, 2004.

[44] Y. Zhang, B. Cao, and D.-Y. Yeung. Multi-domain collaborative filtering. In *UAI*, 2010.

[45] L. Zhao, G. Sukthankar, and R. Sukthankar. Incremental relabeling for active learning with noisy crowdsourced annotations. In *SocialCom/PASSAT*, 2011.