Additive Co-Clustering of Gaussians and Poissons for Joint Modeling of Ratings and Reviews

Chao-Yuan Wu*
Computer Science Department
University of Texas at Austin, TX
cywu@cs.utexas.edu

Alex Beutel*
Computer Science Department
Carnegie Mellon University, Pittsburgh, PA
abeutel@cs.cmu.edu

Amr Ahmed
Strategic Technologies
Google, Mountain View, CA
amra@google.com

Alexander J. Smola
Carnegie Mellon University, Pittsburgh, PA
Marianas Labs, Mountain View, CA
alex@smola.org

Abstract

Understanding a user's *motivations* provides valuable information beyond the ability to recommend items. Quite often this can be accomplished by perusing both ratings and review texts, since it is the latter where the reasoning for specific preferences is explicitly expressed.

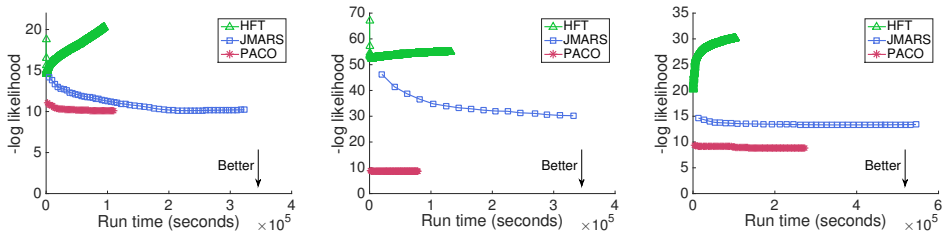
Unfortunately matrix factorization approaches to recommendation result in large, complex models that are difficult to interpret and give recommendations that are hard to clearly explain to users. In contrast, in this paper, we attack this problem through succinct additive co-clustering. We devise a novel Bayesian technique for summing co-clusterings of joint Gaussian and Poisson distributions. With this novel technique we propose a new Bayesian model for joint collaborative filtering of ratings and text reviews through a sum of simple co-clusterings. Our model is non-parametric in the size of the co-clusterings and we offer a novel efficient sampling algorithm for learning the sum of Poisson distributions and the sum of Gaussian distributions. The simple structure of our model yields easily interpretable recommendations. Even with a simple, succinct structure, our model outperforms competitors in terms of predicting both ratings and reviews.

1 Introduction

Recommendation systems often serve a dual purpose — they are expected to generate suggestions that users might like, while simultaneously being able to *explain* why a certain recommendation was made. This increases a user's confidence in a recommender system and it offers valuable insight for debugging a malfunctioning model.

Matrix factorization [3] accomplishes this goal only to a limited extent, since it maps all users and movies into a rather low-dimensional space. This limits attempts to understand the model to principal component analysis and nearest neighbor queries for specific instances. On the other hand, in many cases users are actually happy to provide explicit justification for their preferences in the form of written reviews. JMARS [2] exploited this insight by designing a topic model to capture reviews and ratings jointly, thus offering one of the first works to infer both topics and sentiments without requiring explicit aspect ratings.

*These authors contributed equally.



(a) Amazon fine foods (b) RateBeer (c) Yelp

Figure 1: Negative log likelihood. PACO better jointly predicts ratings and reviews than state-of-the-art JMARS [2] and HFT [4] on Amazon Fine Food, Yelp and RateBeer datasets. The joint predictive power is capture by the normalized negative log likelihood. Lower is better.

A challenge in these approaches is that the model must fit a language model to a rather messy, high dimensional embedding of users and items. In the context of recommender systems ACCAMS [1] addressed this problem by introducing a novel additive non-parametric co-clustering model for matrix completion. This approach yielded excellent predictive accuracy while providing a well-structured, parsimonious model, making the model interpretable. Because the model heavily relied on backfitting and an additive representation of a regression model, it is not possible to combine it with multinomial language models, i.e. a simple bag-of-words representation, since probabilities are not additive — they need to be normalized to 1.

We address this problem by introducing a novel *additive* language description in the form of a sum of *Poisson* distributions rather than a Binomial distribution. This strategy allows us to use backfitting for documents rather than just in a regression setting, and enables a wide variety of new applications. This is possible because the Poisson distribution is closed under addition. This means that sums of Poisson random variables remain Poisson. This property also applies to mixtures of Poisson random variables, i.e. the occurrence of multiple words.

With this approach we make a number of contributions:

- We design an additive co-clustering model, PACO, that is non-parametric in the size of the co-clusterings and can sum of both Gaussian and Poisson distributions. PACO is used to jointly learn a model of reviews and ratings, with the ability to now interpret our model.
- We describe an efficient technique for sampling from a sum of Gaussian and Poisson random variables to facilitate efficient inference.
- We give empirical evidence across multiple datasets that PACO has better prediction accuracy for ratings than competing methods, such as HFT [4] and JMARS [2]. Additionally, our method predicts text reviews better than HFT, and achieves nearly as high quality review prediction as JMARS, while being far faster and simpler. As seen in Figure 1, PACO outperforms both competing models in jointly predicting ratings and reviews.

In summary, we propose a simple and novel model and sampler, capable of characterizing user and item attributes very concisely, while providing excellent accuracy and perplexity.

2 Additive Co-Clustering Model

Our goal is to jointly model documents and review scores. We begin with the problem of matrix approximation, where we have a real valued sparse matrix, where we observe a small percentage of the entries in the matrix and our goal is to predict the missing values. To be precise, we have a matrix $\mathbf{R} \in \mathbb{R}^{N \times M}$ with the set $\mathcal{I}$ of observed entries (u, m). To predict the missing values, we would like to learn a model $\mathcal{M}$ with parameters θ, such that the size of θ is small, $|\theta| \ll \mathbf{R}$, and $\mathcal{M}(\theta)$ approximates $\mathbf{R}$ well:

$$\underset{\theta}{\text{minimize}} \sum_{(u,m)\in\mathcal{I}} (\mathbf{R}_{u,m} - \mathcal{M}(\theta)_{u,m})^2 \tag{1}$$

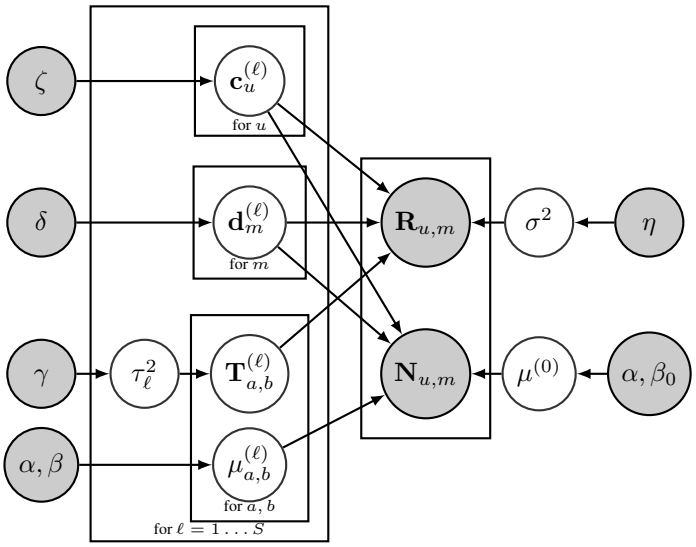


Figure 2: The generative model for PACO to predict both ratings $\mathbf{R}$ and review text $\mathbf{N}$. (Note, for the sake of space we simplify the model slightly by not explicitly separating the different language models associated with each stencil.)

Key to our model is the notion of a stencil, an extremely easy to represent block-wise constant rank-k matrix. A stencil $\mathbf{T}$ assigns each row u, typically a user, to a row cluster a, each column, typically an item, to a column cluster b. The rating that user u gives to item m is thus predicted by the value $\mathbf{T}_{a,b}$. We define $\mathbf{c}_u$ to be the cluster assignment for u and $\mathbf{d}_m$ to be the cluster assignment for m. Therefore, our goal is to find a stencil $(\mathbf{T}, \mathbf{c}, \mathbf{d})$ that gives small approximation error to $\mathbf{R}$.

To improve the approximation accuracy without making the model size grow quickly, we can use *multiple* stencils in an additive fashion:

$$\underset{\left\{\mathbf{T}^{(\ell)}, \mathbf{c}^{(\ell)}, \mathbf{d}^{(\ell)}\right\}}{\text{minimize}} \sum_{(u,m) \in \mathcal{I}} \left(\mathbf{R}_{u,m} - \sum_{\ell=1}^{S} \mathbf{T}^{(\ell)}_{\mathbf{c}_u^{(\ell)}, \mathbf{d}_m^{(\ell)}} \right)^2$$

That is, we learn an additive model of S stencils that minimizes the approximation error to $\mathbf{R}$.

2.1 Modeling Text using an Additive Poisson Model

In addition to mining real-valued matrices of ratings, we also want to be able to model text. We introduce a novel approach to additive co-clustering using Poisson distributions. In a nutshell, we exploit the fact that the Poisson distribution is closed under addition, i.e. for $a \sim \text{Poi}(\lambda)$ and $b \sim \text{Poi}(\gamma)$ we have $a + b \sim \text{Poi}(\lambda + \gamma)$ where λ and γ denote the rates of each of the random variables.

For each user and movie pair (u, m) pair we let $n_{u,m,x}$ denote the count for word x in the review. For each part of the model in each stencil, we learn a Poisson random variable with rate $\mu^{(\ell)}_{c_u, d_m, x}$ for each word x. A review is predicted by summing over the appropriate Poisson random variables for the given user and movie (u, m).

2.2 The Joint Generative Model

Now we are ready to present the full model. We follow the joint objective:

$$\underset{\left\{\mathbf{T}^{(\ell)}, \mathbf{c}^{(\ell)}, \mathbf{d}^{(\ell)}, \lambda\right\}}{\text{minimize}} \sum_{(u,m) \in \mathcal{I}} \left(\mathbf{R}_{u,m} - \sum_{\ell=1}^{S} \mathbf{T}^{(\ell)}_{\mathbf{c}_u^{(\ell)}, \mathbf{d}_m^{(\ell)}} \right)^2 + \sum_{(u,m) \in \mathcal{I}} \frac{1}{|n_{u,m}|} \sum_{x \in \mathcal{W}} \log \text{Poi}(\lambda_{u,m,x})$$

where

$$\lambda_{u,m} = \mu^{(0)} + \mu^{(m)} + \left[\sum_{\ell=1}^{S} \mu^{(\ell)}_{\mathbf{c}_u^{(\ell)}, \mathbf{d}_m^{(\ell)}} + \mu^{(m,\ell)}_{\mathbf{d}_m^{(\ell)}} + \mu^{(u,\ell)}_{\mathbf{c}_u^{(\ell)}} \right] \tag{2}$$

Our goal is learn a set of stencils whose summation minimizes the prediction error on ratings and maximizes the likelihood of generating the text. Following NLP literature [6, 5], we normalize reviews by their length to balance error in both parts of the data.

To model review text, each stencil has three language models: a stencil-specific user language model $\mu^{(\ell)}_{\mathbf{c}_u^{(\ell)}}$, a stencil-specific movie language model $\mu^{(m,\ell)}_{\mathbf{d}_m^{(\ell)}}$, and block language model, $\mu^{(\ell)}_{\mathbf{c}_u^{(\ell)}, \mathbf{d}_m^{(\ell)}}$. The block language model captures the stencil-specific interaction between the movie and the user. In addition, we add a global movie language model, $\mu^{(m)}$, and a global background language model, $\mu^{(0)}$. The text of the review is modeled as a combination of these Poisson language models.

Minimizing the aforementioned objective function, is approximately equivalent to maximizing the log-likelihood of the graphical model in Figure 2. The generative process proceeds as follows:

$$\mathbf{R}_{u,m} \sim \mathcal{N}\left(\sum_{\ell=1}^{S} \mathbf{T}^{(\ell)}_{\mathbf{c}_u^{(\ell)}, \mathbf{d}_m^{(\ell)}}, \sigma^2 \right) \tag{3}$$

$$\mathbf{c}^{(\ell)}, \mathbf{d}^{(\ell)} \sim \mathrm{CRP}(\delta) \text{ for all } (\ell) \tag{4}$$

$$\mathbf{T}^{(\ell)}_{c,d} \sim \mathcal{N}(0, \sigma^2_{(\ell)}) \tag{5}$$

$$n_{u,m,x} \sim \mathrm{Poisson}\left(\lambda_{u,m,x}\right) \tag{6}$$

$$\mu^{(*)}_x \sim \mathrm{Gamma}(\alpha, \beta) \tag{7}$$

where $\mathbf{c}^{(\ell)}_u, \mathbf{d}^{(\ell)}_m$ is the cluster (u, m) assigned to in stencil ℓ. In essence, we model user and movie clusters inside each stencil using a Chinese restaurant process (CRP). Ratings are modeled similar to ACCAMS while Equations (2), (6) and (7) contain the additional text modeling aspects of our model. To ensure no overfitting, we use conjugate Gamma prior for all the vectors μ.

2.3 Inference

We offer an efficient Gibbs sampling procedure to learn the PACO model. The collapsed Gibbs sampler for Gaussian distributions is described in [1]. In our complete paper, we describe the sampling of a given $\mu^{(\ell)}_{a,b,x}$ conditioned on all other language models in PACO. Together, these two samplers enable us to learn the complete PACO model efficiently.

3 Experiments

To extensively test our model, we select four datasets about movies, beer, businesses, and food. All four datasets come from different websites and communities, thus capturing different styles and patterns of online ratings and reviews. We randomly select 80% of data as training set and 20% as testing set. An overview of our results can be seen in Figure 1, and we give an in-depth analysis in the complete paper, including examples of the predicted reviews.

4 Conclusion

We have described our model for taking summations over non-parametric co-clusterings of both real valued and count data. In [1] we describe an efficient collapsed sampler for learning this model over real valued data and demonstrate it offers state of the art performance on Netflix ratings while being a far smaller, simpler model.

In this paper we offer an efficient sampling scheme that also can learn over a summation of Poisson distributions. As shown in Figure 1, this model outperforms competitors, and further experimentation shows that our model yields highly interpretable results over a wide variety of datasets.

References

[1] A. Beutel, A. Ahmed, and A. J. Smola. ACCAMS: Additive Co-Clustering to Approximate Matrices Succinctly. In *Proceedings of the 24th International Conference on World Wide Web*, pages 119–129, 2015.

[2] Q. Diao, M. Qiu, C.-Y. Wu, A. J. Smola, J. Jiang, and C. Wang. Jointly modeling aspects, ratings and sentiments for movie recommendation (jmars). In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 193–202. ACM, 2014.

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

[4] J. McAuley and J. Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172. ACM, 2013.

[5] X. Wang and A. McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433. ACM, 2006.

[6] X. Wang, N. Mohanty, and A. McCallum. Group and topic discovery from relations and text. In *Proceedings of the 3rd international workshop on Link discovery*, pages 28–35. ACM, 2005.