UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

Title
Actively learning a Bayesian matrix fusion model with deep side information

Permalink
https://escholarship.org/uc/item/99b5n8s8

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 46(0)

Authors
Yu, Yangyang
Suchow, Jordan W.

Publication Date
2024

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Actively learning a Bayesian matrix fusion model with deep side information

Yangyang Yu & Jordan W. Suchow
School of Business
Stevens Institute of Technology
Hoboken, NJ 07030
{yyu44,jws}@stevens.edu

Abstract

High-dimensional deep neural network representations of images and concepts can be aligned to predict human annotations of diverse stimuli. However, such alignment requires the costly collection of behavioral responses, such that, in practice, the deep-feature spaces are only ever sparsely sampled. Here, we propose an active learning approach to adaptively sample experimental stimuli to efficiently learn a Bayesian matrix factorization model with deep side information. We observe a significant efficiency gain over a passive baseline. Furthermore, with a sequential batched sampling strategy, the algorithm is applicable not only to small datasets collected from traditional laboratory experiments but also to settings where large-scale crowdsourced data collection is needed to accurately align the high-dimensional deep feature representations derived from pre-trained networks. This provides cost-effective solutions for collecting behavioral data and generating high-quality predictions in large-scale behavioral and cognitive studies.

Keywords: Bayesian Matrix Factorization, Deep Learning, Active Learning

Introduction

In cognitive research, Bayesian probabilistic models typically serve two principal roles: one as a hypothesis positing how individuals draw inferences from their observations of the environment, and the other as a tool enabling scientists to learn from observations of human behavior (Vul et al. [2014], Griffiths et al. [2008], Mamassian et al. [2002]). Our work acts as an intermediate approach that bridges these two uses of Bayesian models. We use Bayesian Probabilistic Matrix Factorization (BPMF) with deep-side information to align a machine representation of entities to human behavioral responses to those entities, such that the model serves as both a model of people's mental representations and as a predictive model of their behavior. It offers a feasible model structure to align machine vision systems with human visual perception. For instance, it can integrate the bimodal information from facial imagery and psychological attributes and yield predictions of people's impressions of human faces.

BPMF has proven effective in consolidating multi-source information and predicting missing responses while constructing confidence intervals (Salakhutdinov and Mnih [2008], Adams et al. [2010]). In BPMF, one is given a participant–stimulus matrix (where rows are participants, columns are stimuli, and cells are the behavioral response or judgment of the row participant to the column stimulus). The method factorizes this matrix into two sets of latent factors, one representing participants and the other representing stimuli. When a pretrained feature representations of participants (e.g., demographic embeddings) or stimuli (e.g., image embeddings) are available, one can use these features as "side information" to guide the factorization via a regression model. In this way, BPMF with deep side information amounts to aligning the pretrained representations to best predict the observed behavioral data (Xiao et al. [2019]).

Even so, implementing BPMF with deep side information in large-scale behavioral prediction tasks with multi-modal object features presents two primary challenges. First, high-quality predictions require advanced machine-generated features to outperform traditional human-defined ones. Machine learning algorithms can generate a vast array of new objects based on prior ones, boosting diversity and realism while minimizing bias among stimuli. However, employing deep learning algorithms to extract informative, high-dimensional features from various objects linearly escalates computational costs with the number of data features. This complicates the integration of BPMF with deep learning methods, restricting it to sparse sampling from the deep-feature space. The second challenge arises from the scarcity of essential information for reliable predictions, owing to the highly sparse response matrix, a scenario exacerbated in human-subjects research. Here, data collection, constrained by budget and resources, needs to deal with large participant populations and potentially extensive question lists. Budgetary limits might restrict the number of questions posed, and elongated experimental instruments could yield inaccurate responses as participants may resort to mental shortcuts (Krosnick [1991]). Consequently, large-scale experiments often only capture responses to a minor portion of the total instrument from each participant. Thus, to effectively employ BPMF with deep-side information on large-scale behavioral prediction tasks, developing a data sampling strategy to effectively target the most informative data points is critical, ensuring satisfactory predictive outcomes despite these constraints.

Active learning is a data acquisition technique that can interactively identify the most informative samples to efficiently create a training data set. Although this training set can be compact, it possesses a powerful predictive capacity. Active learning has been widely employed to tackle problems associated with accuracy in sparse matrix completion (Elahi et al. [2016], Chakraborty et al. [2013a]). One of its key strengths is the capacity to accurately infer the complete response distri-

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.* ©2024 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

bution from a limited selection of samples, obviating the need to query the majority of the response matrix.

Here, we propose an active learning method for a BPMF model using uncertainty (Sugiyama and Ridgeway [2006]) and k-Center Greedy (Sener and Savarese [2017]) sampling strategies, showing enhanced learning efficiency compared to passive learning. We further examine the effect of varying Markov chain Monte Carlo (MCMC) simulation chain lengths on the active sampling performance to optimally integrate active learning into the BPMF model framework. The estimation of posterior uncertainty brings a cost associated with the number of posterior MCMC samples collected, presenting a trade-off between slow precise estimates and quick, less accurate ones. We investigate this trade-off by adjusting the number of MCMC samples for posterior uncertainty estimation in model parameters and measure its impact on algorithm performance within a fixed computational budget.

In this paper, we outline an active learning framework for the deep Bayesian matrix factorization model, train it on a large behavioral dataset — the One Million Impressions dataset (Peterson et al. [2022]) — and validate the model's learning efficiency and performance improvement via an effective active strategy. Lastly, we demonstrate our method's promising predictive results by adaptively querying a subset of the data for training.

Related work

We discuss related work in three separate sections below, on the topics of using Bayesian probabilistic matrix factorization (BPMF) for human behavioral prediction, high-dimensional deep feature representations as the side information of BPMF and active learning.

Bayesian Probabilistic Matrix Factorization In the context of behavioral data, BPMF factorizes a participant–stimulus matrix into latent factors for participants and for stimuli. Through MCMC simulations, Bayesian posteriors are computed for these factors (Mamassian et al. [2002], Kersten et al. [2004]). In the context of data collected from traditional laboratory experiments, it has been successfully applied to predict individuals' perceptual outcomes based on human-interpretable features (Zhang et al. [2020]).

Modeling Perception Using Deep Features Human behavioral tasks often require fusion of multi-modal data (e.g. visual and linguistic inputs). Prior research (Peterson et al. [2022], Zhang et al. [2018]) highlights machine-generated deep features' superiority over traditional human-interpretable attributes (e.g., image color or size) due to their comprehensive high-dimensional representations and self-generating capability for novel predictions. One can use pretrained networks like StyleGAN2 (Karras et al. [2020]) for images and Sentence-BERT (Reimers and Gurevych [2019]) for text to significantly lower the cost of collecting deep feature representations.

Active Learning Taking inspiration from sparse matrix completion research (Settles [2009], Chakraborty et al. [2013b]), we leverage active learning as a practical approach to impute a full response distribution for each participant over diverse objects, given a response matrix with limited entries. This approach prioritizes data points for querying in the response matrix, allowing for a faster alignment of machine inferences with human thinking with fewer queried data points. Active learning accelerates this alignment by adaptively choosing the most informative data samples, a critical approach in large-scale cognitive experiments with constrained resources. In our model, two active strategies are deployed to enhance BPMF performance with a limited training set: 1) Uncertainty sampling (Chakraborty et al. [2013b], Sutherland et al. [2013]), prioritizing data points with the highest predictive uncertainties in each iteration as the most informative; 2) k-Center Greedy sampling (Sener and Savarese [2017]), seeking k data points that maximize the mutual information between them and the remaining unselected data pool in every iteration. This is achieved by pre-setting k centers and identifying data points that minimize the maximum distance of these points to the nearest centers.

Methods

Our approach improves the performance of perceptual learning on limited training cases by implementing the active strategy to the Bayesian matrix factorization with deep side features. The details of this approach and notation are articulated in two sections below.

Deep Bayesian Probabilistic Matrix Factorization

First, in predicting first impressions, a pre-trained deep network is used to create face and trait features. Utilizing indices j for face images, h for traits, and i for participants, Peterson et al. (Peterson et al. [2022]) proposed a method for extracting deep face features f_j using StyleGAN2 generator. Meanwhile, we employ the Sentence-BERT model to create a deep trait feature set t_h. This becomes in two latent spaces for the response matrix R: a 512-dimensional image feature latent space $\mathcal{F}$ for faces and a 300-dimensional linguistic feature space $\mathcal{T}$ for traits. These spaces are represented by conditioned, unit-variance, multivariate normal latent variables: ω_{f_j} for the face space and ω_{t_h} for the trait space. A parallel experiment involved lower-dimensional dense deep features, conducted by further processing pre-trained image and trait features through a multi-layer perceptron (MLP) model (Yu and Suchow [2022]). The goal is to determine whether encapsulating deep features to a lower dimensionality provides a sufficient representation and if it's necessary to include an intermediate dimension reduction operation before inputting side features into the Bayesian model. If so, understanding the appropriate degree of dimension reduction is key to ensuring adequate feature representativeness while minimizing computational cost and simulation time.

Second, under the BPMF setting, the latent variables represent the computational coefficients for the participants' impression ratings R_{jh}. The coefficients are estimates based on two priors of $\omega_{f_j}, \omega_{t_h}$ following spherical Gaussian distributions (Gurkan and Suchow [2022]). In Formulas 1 and 2, the

two-sided features are first condensed to a consistent dimensionality within the latent spaces and then merged through a fusion process. The fusion products are processed via Gaussian likelihood functions. Subsequently, the resulting predictions, denoted as $\hat{R}_{jh}$, are derived through MCMC simulations.

$$\begin{aligned} p(\omega_{f_j}|\sigma) &= \text{Normal}(\omega_{f_j}|\sigma^{-1}\mathbf{I}); \\ p(\omega_{t_h}|\sigma) &= \text{Normal}(\omega_{t_h}|\theta^{-1}\mathbf{I}), \end{aligned} \tag{1}$$

where σ and θ are independent and gamma distributed.

$$\begin{aligned} F_j &= f_j \times \omega_{f_j}^T; & T_h &= t_h \times \omega_{t_h}^T; \\ R_{jh}^* &= F_j \times T_h^T + \varepsilon_{jh}; & \hat{R}_{jh} &= \text{sigmoid}(R_{jh}^*) \times 100, \end{aligned} \tag{2}$$

where σ and θ are independent and Gamma distributed. f_j and t_h are row vectors and $R_{jh}^* \in (-\infty, \infty)$ is projected into predicted ratings as a continuous value $\hat{R}_{jh} \in (0, 100)$.

Third, we further analyze the impact of the simulation process on our deep BPMF model's performance. By running parallel simulation chains of equal length but varying proportions of warm-ups to posterior samples, we identify the optimal mean for impression inference. Furthermore, we employ sequential Markov chain Monte Carlo (MCMC) (Yang and Dunson [2013]) as the approximation method for modeling BPMF across the whole dataset. This choice is guided by the MCMC method family's suitability for large hierarchical model computation (Banerjee et al. [2003]), and its effectiveness in our case for integrating over thousands of parameters associated with deep features. We observe that aggregating parameters from the last ten Bayesian posteriors rather than all posteriors stabilizes model predictions, reducing sampling randomness. Compared to traditional MCMC, sequential MCMC more readily extracts the last few posteriors in each iteration, offering reduced memory usage and faster computation when using Numpyro in Python as the development platform.

Actively Learning The Matrix Factorization Model

By implementing active learning strategies in the training data sampling process of BPMF, we aim to use as few filled matrix entries as possible in training, predict all the rest in the response matrix, and converge the Bayesian model in fewer iterations. We uniformly choose a very small initial training pool S^0 with size L (5, 8, etc.), selected randomly from the whole rated data pool S, which has size N. Notice that only the initial pool is accessible to us. The feature pair (f_{jn}, t_{hn}) of an arbitrary data entry $\mathcal{C}_n$ in S is shortened as $\mathcal{U}_n$, $n \in \{1, ..N\}$, and its rating as R_{in}. The feature pair of a data entry $\mathcal{C}_l$ in the initial pool S^0 is denoted as $\mathcal{U}_l$ with observable rating as R_{il}, $l \in \{1, ..., L\}$. Besides the initial pool, we assume Q opportunities as the budget for asking an oracle for information about an extra p data points and a learning algorithm A_S to guide us in choosing the appropriate points at each time. A_S generates an updated set of parameters $\{\omega_{f_j}^*, \omega_{t_h}^*\}$ based on $S^{0*} = S^0 \cup S^p$. The choice of S^p for the oracle to label is a subset of S, which minimizes the future expected learning loss. Formula 3 shows the future expected loss after q times label querying.

$$\min_{S_q^p} E_{(\mathcal{C}_1, ..., \mathcal{C}_N) \sim S}[\mathcal{L}(\mathcal{C}_n; A_{S^0 \cup ... S_q^p})] \tag{3}$$

Compared to the classic definition of active learning, our active strategy has two significant differences: 1) We set A_S to query a batch of p data points in each round instead of a single data point because previous works (Sener and Savarese [2017], Zhang et al. [2020]) illustrate that with a large dataset and high-dimensional deep features, the performance improvement made by one data point in each round is negligible. 2) As active learning has mainly been shown to be effective for tasks with discrete prediction targets (Settles [2009], Yona et al. [2022]), we further expand its capability of handling continuous prediction cases by reformulating the learning metrics and loss functions. Here, we explore two types of batched active strategies to construct and minimize the loss function for continuous targets and compare their performance. One is pure uncertainty-based sampling, and the other takes the trade-off of sampling diversity and uncertainty of sampling into account.

Uncertainty Sampling One classic approach of uncertainty selection is targeting samples with the least confidence in predictions in each active iteration. Sugiyama and Ridgeway (Sugiyama and Ridgeway [2006]) justified selecting data points with the maximum standard deviations of predictive distribution, $\sigma_{\omega_{\mathcal{U}_l}^*}$ (shorten as σ), as the most uncertain samples in the context of Gaussian linear regression. Here, we further extend this strategy to Bayesian matrix factorization because of two critical similarities: 1) like their work, we have a continuous prediction target, and 2) the prediction distribution is also a univariate Gaussian distribution, $\rho(R^*)$.

In this paper, we validate the performance of uncertainty-based active Bayesian matrix factorization on the impression prediction task.

Algorithm 1 Batched Uncertainty Sample Selection for Deep Bayesian Matrix Factorization

Input: The set of randomly selected d ratings as initial pool S^0; The rating R_{il} in S^0 is given by a participant i for a face image in terms of a certain trait. And budget $Q > 0$.
Output: Predicted ratings for all entries of the response matrix, $\hat{R}$
Extract deep features face images f_{jl} and trait t_{hl}
repeat Q times, initialize $q = 1$;
Update parameters $\omega_{f_j q}^*, \omega_{t_h q}^*$ via MCMC
Compute predicted $\hat{R}_q$, σ_q for the response matrix
Query S_q^p with argmax$\sum_{z=1}^{p} \sigma_{z_q}{}_1$
$S_q^0 \leftarrow S_q^p \cup S_{q-1}^0$
q = q + 1
until $q = Q$
return $\hat{R}_Q$, $S \setminus S_Q^0$

k-Center Greedy Sampling k-Center Greedy sampling is a pool-based active sampling strategy. Sener et al. (Sener and Savarese [2017]) regard active learning loss for classifications using CNN given a batch of p samples, $E_{(C_1,...,C_N)\sim S}(\mathcal{L}(C_n;A_{S^p}))$, as containing three components: generalization loss, training loss, and core-set loss. p denotes the number of central points in the unexplored data pool and also equates to the count of points chosen during each adaptive sampling cycle.

Algorithm 2 k-Center Greedy Sample Selection for Deep Bayesian Matrix Factorization

Input: The set of randomly selected d ratings as initial pool S^0; the rating R_{il} in S^0 is given by a participant i for a face image in terms of a certain trait. And budget $Q > 0$.
Output: Predicted ratings for all the entries of the response matrix, $\hat{R}$
Extract deep features face images f_{jl} and trait t_{hl}
repeat Q times, initialize $q = 1$;
Update parameters $\omega^*_{f_{jq}}, \omega^*_{t_{hq}}$ via MCMC
Set p learning centers
Choose p centers $(C_{1q},...,C_{pq}) = S^p_q$ [2]by
$S^p_q = \underset{C_{nz}\in S\setminus S^p_{zq}}{\text{argmax}} \min_{C_{lz}\in S^p_{zq}} \triangle(\mathcal{U}_{nz},\mathcal{U}_{lz})$,
where set $S^p_q \subseteq S \setminus S^0_{q-1}$
$S^0_q \leftarrow S^0_{q-1} \cup S^p_q$
q = q + 1
until $q = Q$
return $\hat{R}_Q$, $S \setminus S^0_Q$

In our scenario, which involves predicting continuous targets, the first two losses are primarily managed by the Bayesian model, while the active learning strategy concentrates on minimizing core-set loss. Core-set loss is defined as the distance between average empirical loss over the points with known ratings and that over the entire dataset, including entries with unknown ratings. Sener et. al shows its upper bound as $O(d_{S^0\cup S^p}) + O(\sqrt{\frac{1}{N}})$. Consequently, the optimization goal of loss can be converted to minimize the coverage radius from p centers ($d_{S^0\cup S^p}$). The minimization of $d_{S^0\cup S^p}$ is further deduced as $\min_{S^p} \max_{C_n\in S\setminus(S^0\cup S^p)} \min_{C_l\in S^0\cup S^p} \triangle(\mathcal{U}_n,\mathcal{U}_l)$ when p data points are queried. This formula can be computed based on the distances of feature pairs $\mathcal{U}_n = (f_{jn}, t_{hn})$ and $\mathcal{U}_l = (f_{jl}, t_{hl})$ in a two-dimensional coordinate system. The $\min_{S^p} S^0 \cup S^p$ is iteratively recomputed in Algorithm 2 to search for new data points to be queried.

Experiments and Results

We conduct three sets of experiments that apply the active learning strategy. The active learning method's efficacy and efficiency are quantified by the Root Mean Square Error (RMSE) of test data over three experimental repetitions. The experimental dataset is the One Million Impressions dataset (Peterson et al. [2022]), which comprises over 1 million impression ratings on 34 traits for 1,000 machine-generated face images. Given that each image receives ratings from only a subset of the total participant pool, the response matrix is relatively sparse. Rather than integrating the active-learning strategy into the loop of data collection, we use a pregenerated large-scale dataset as an "oracle" so that we can efficiently evaluate each algorithm across multiple runs — the active learning strategy selects which data from full set will be used to train the current model. We assessed the active learning method's effectiveness and efficiency using the Root Mean Square Error (RMSE) averaged across three experimental repetitions.

Choices of Strategy and Batch Size

The first experiment seeks to identify the optimal batch size for an active learning task on a large scale. We employ two active sampling strategies, uncertainty and k-Center Greedy, with deep Bayesian matrix factorization models to predict first-impression ratings. These strategies adaptively query batch samples, with batch size varying per experimental setting. Five active models with distinct strategy and batch size combinations are run concurrently. The initial training pool, equivalent in size to the chosen batch size, is randomly drawn from the unknown data pool. The baseline is the same deep Bayesian matrix factorization model with passive sampling that randomly selects increasing numbers of samples across iterations, without adaptive querying. Pretrained StyleGAN2 image features and Sentence-BERT trait features serve as the deep BPMF's input side information.

Figure 1 presents adaptive sampling across 34,000 samples, where the uncertainty active strategy surpasses both the k-Center Greedy strategy and the passive baseline model. Among the uncertainty sampling batch sizes, a batch size of 8 emerges as the optimal choice, demonstrating the fastest and sharpest decrease in test RMSE, and experiencing minimal fluctuations and the narrowest confidence interval as the training pool size increases. Its advantage is particularly notable when the training pool has fewer than 10,000 samples, representing roughly 0.1% of the entire dataset. In this case, the test RMSE drops sharply from 32.1 to 27.8, approximately 1.0 lower than the passive learning's test RMSE. The uncertainty strategy with a batch size of 10 achieves the second most efficient convergence of the six models but exhibits less stability in test RMSE during the intermediate training pool size in a range of [15,000, 25,000]. The batch size 5 model encounters severe fluctuations when the training pool grows larger than 20,000 and sometimes has noticeably higher predicted RMSEs and a wider confidence interval than the baseline model, indicating it is the worst model in the uncertainty strategy group. On the contrary, regardless of training set size, neither batch-8 nor batch-10 k-Center Greedy performs better than the baseline model. After the training pool reaches 28,000 samples, models with batched uncertainty active sampling gradually

[1] σ_{z_q} denotes the predicted standard deviation for the zth queried data point in a batch p samples during the qth iteration of the uncertainty strategy

[2] S^p_{hq} denotes the kth queried data point in a batch p in the qth iteration of k-Center Greedy active strategy.

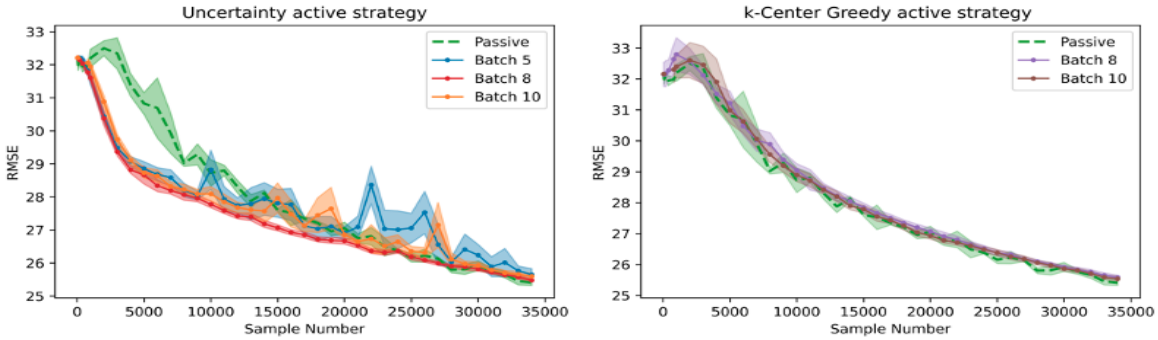


Figure 1: Test RMSEs over sample number with different batch sizes for uncertainty vs. k-Center Greedy active strategies with 95% confidence intervals.[3]

converge with the baseline model at a test RMSE of about 25.8, indicating the training pool has reached a size of containing extensive information. Thus, even random selection works as effectively as the active strategy due to the Bayesian model's estimation and control for future predictive uncertainty. Consequently, we confirm that the batched uncertainty active strategy for deep Bayesian matrix factorization can efficiently decrease predictive RMSE when training data size is limited. Furthermore, the choice of batch size influences the model's prediction stability.

Efficacies of Input Feature Dimensionalities

The second experiment investigated the trade-off between feature dimensionality and predictive accuracy in Bayesian matrix fusion models using uncertainty-based active sampling. It aims to optimize computational resources by identifying the most efficient feature vector size. For our baseline model (Model 0), we use pre-trained deep features (512 for images, 300 for traits). Models 1 to 4 utilize denser feature layers from the bimodal MLP model (Yu and Suchow [2022]) with dimensions reduced to 100, 80, 32, and 16, respectively, before fusion.

Figure 2 demonstrates that Model 0 maintains the lowest test RMSE. Models 1 and 2, with 100 and 80 dimensions, perform similarly well with up to 30,000 samples but struggle with larger sample sizes, showing unstable RMSEs. This instability suggests their limitations in selecting informative samples for further learning. Models 3 and 4, with 32 and 16 dimensions, consistently underperform, unable to identify informative data points across all training sizes. The data indicates that an active strategy with roughly 33% and 20% of the original feature dimensions is optimal for small datasets, but as datasets grow, this strategy becomes less effective, likely due to the loss of detail from dimensionality reduction. The high RMSEs of Models 3 and 4 corroborate this, with information loss becoming more critical as dimensionality decreases.

Impact of Simulation Chain Length

In our third experiment set, we tested three MCMC simulation chain configurations: (1) incrementally increasing warm-ups and posterior samples in a 3:5 ratio, (2) solely increasing

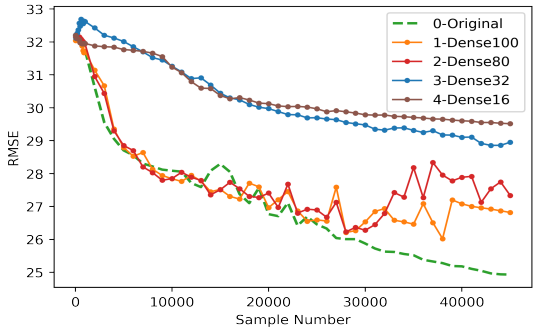
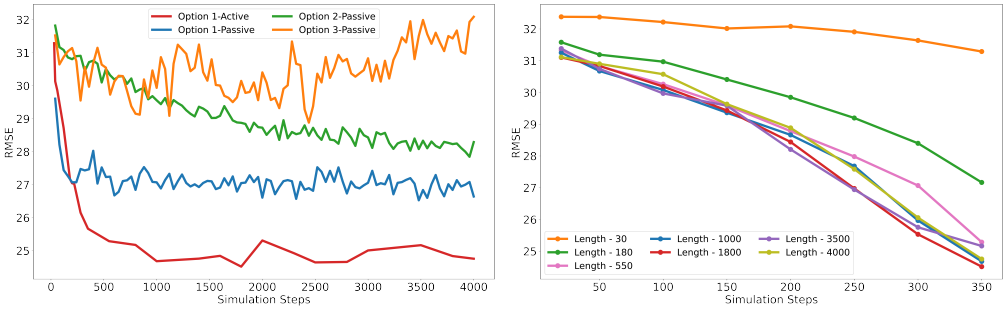


Figure 2: Test RMSEs over sample number for the BPMF with batch-10 uncertainty active strategy under different choices of input features dimensions.

posterior samples by steps of 5, and (3) increasing warm-ups by steps of 5 while keeping posterior samples constant as 5. After determining the best setup, we implemented a batch-2 uncertainty active sampling strategy on the BPMF model. We then compared active to passive sampling to see if longer MCMC chains could match active learning's efficiency. This was tested on a 1000-sample subset from the One Million Impressions dataset, which comprised random selections of ten responses for each face and trait combination. With a 350-sample training pool, we assessed performance by RMSE on the test set of 650. Due to the dataset's reduced size, the active strategy's batch size was also reduced to two.

We first executed three passive-sampling BPMF models, each with different simulation chain options, and compared their test RMSE trends over time (Figure 3(a)). Our observations revealed that the MCMC chains from Option 1 offered the highest reduction efficiency in predictive RMSE as the simulation steps increased. Options 2 and 3 consistently yielded higher test RMSEs than Option 1, demonstrating that a simulation chain with proportional increases in warm-ups and posteriors yields the most powerful predictions for human impression ratings. Moreover, we applied the optimal simu-

[3]To smooth out fluctuations in RMSEs, each dot in Figures 1 and 2 is averaged from 200 samples of the adjacent epochs .



(a) Test RMSEs for different model options over the lengths of the simulation chain with 350 training samples

(b) Test RMSEs for the BPMF using batch-2 active strategy with different simulation chain lengths over training sample numbers.

Figure 3: Result summary of the third experiment. [4]

lation chain (Option 1) to the BPMF model with a batched uncertainty active sampling strategy. Its results in Figure 3(a) showed a significantly better performance with fewer simulation steps than all passive learning models, except during the initial stages (<200 steps). By the time 220 simulation steps were reached, the active model's predictive RMSE dropped to 27.024, already 0.05 lower than the best-performing passive model with much longer simulation chains. When the simulation chain length extended to 1,800 steps, the active model achieved the best test RMSE of 24.515, 2.533 lower than the passive model with the same number of steps. After that point, the active model's test RMSE increased slightly but remained the lowest among all settings. As too few simulation steps lead to incorrect predictions in practice, we demonstrate that deep Bayesian matrix factorization with active strategy retains a sustained advantage over passive sampling BPMF on model performance. The training time analysis Table 1 further illustrates that the advantage of active learning cannot be offset by extensively extending the simulation chain in terms of both training time and predictive performance. With a reasonably short simulation chain length of 220, the active model's RMSE decreases to a level that the passive models can't reach. This training time analysis was run on a single NVIDIA GeForce RTX 4090 GPU with 24 GB memory.

MCMC chain length	Learning type	Running time (mins)	Test RMSE
220	Active	32.46	27.02
280	Active	36.13	26.15
64,000	Passive	29.32	27.50
80,000	Passive	35.81	27.73
88,000	Passive	39.59	27.15

Table 1: Active learning vs. passive learning training time (in minutes) and test RMSE in the third experiment.

Figure 3(b) displays test RMSEs across varying training pool sizes for different simulation chain lengths. The test RMSE for the 1,800-step model remains among the lowest as the training set expands from 50 to 350 samples, confirming that the optimal performance achieved with 350 training samples is not an exceptional case for the model with 1,800-step simulation during the active sampling process.

Conclusion and Discussion

In this work, we propose an active learning method using BPMF with deep neural networks to predict human behavioral data, which selects informative stimulus-attribute pairs based on model parameter uncertainty. Our empirical tests demonstrate superior performance over passive learning, crucial for budget-limited crowdsourced studies and applicable across domains like online recommendations for social media or e-commerce. Key factors affecting performance include sampling strategy, active learning batch size, dimensionality of deep features, and simulation chain lengths. Our experiments highlight the effectiveness of the uncertainty sampling strategy over the k-Center Greedy strategy, especially in large behavioral datasets. Despite a smaller training pool, Bayesian inference efficiently learns from less confident data, reducing predictive error. However, k-Center Greedy sampling fails to capture feature diversity effectively. Our findings suggest that an appropriate active strategy, batch size and feature dimensionality, along with an optimized simulation chain length in the BPMF model, significantly enhance predictive performance, especially under data query budget and computational resource constraints.

The integration of deep Bayesian matrix factorization with the uncertainty active strategy is promising for impression prediction, using deep image and trait features to generate a 2D response matrix. Given Bayesian factorization's capability with 3D tensors (Xiong et al. [2010]), there's interest in exploring our approach's adaptability to more complex behavioral datasets with diverse features.

[4] To smoothing of fluctuations on RMSE values, the lines and do in Figures 3 is averaged from the nearby 5 epochs.

References

Ryan Prescott Adams, George E Dahl, and Iain Murray. Incorporating side information in probabilistic matrix factorization with gaussian processes. *arXiv preprint arXiv:1003.4944*, 2010.

Sudipto Banerjee, Bradley P Carlin, and Alan E Gelfand. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC, Boca Raton, FL, 2003.

Shayok Chakraborty, Jiayu Zhou, Vineeth Balasubramanian, Sethuraman Panchanathan, Ian Davidson, and Jieping Ye. Active matrix completion. In *2013 IEEE 13th International Conference on Data Mining*, pages 81–90, 2013a. 10.1109/ICDM.2013.69.

Shayok Chakraborty, Jiayu Zhou, Vineeth Balasubramanian, Sethuraman Panchanathan, Ian Davidson, and Jieping Ye. Active matrix completion. In *2013 IEEE 13th International Conference on Data Mining*, pages 81–90, 2013b.

Mehdi Elahi, Francesco Ricci, and Neil Rubens. A survey of active learning in collaborative filtering recommender systems. *Computer Science Review*, 20:29–50, 2016.

Thomas L Griffiths, Charles Kemp, and Joshua B Tenenbaum. Bayesian models of cognition. In Ron Sun, editor, *The Cambridge Handbook of Computational Psychology*. Cambridge University Press, New York, 2008.

Necdet Gurkan and Jordan Suchow. Cultural alignment of machine-vision representations. In *NeurIPS 2022 SVRHM Workshop*, 2022.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8107–8116, 2020.

Daniel Kersten, Pascal Mamassian, and Alan Yuille. Object perception as Bayesian inference. *Annual Review of Psychology*, 55:271–304, 2004.

Jon A Krosnick. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5(3):213–236, 1991.

Pascal Mamassian, Michael Landy, and Laurence T Maloney. Bayesian modelling of visual perception. *Probabilistic models of the brain: Perception and neural function*, pages 13–36, 2002.

Joshua C Peterson, Stefan Uddenberg, Thomas L Griffiths, Alexander Todorov, and Jordan W Suchow. Deep models of superficial face judgments. *Proceedings of the National Academy of Sciences*, 119(17):e2115228119, 2022.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. arXiv preprint arXiv:1908.10084, 2019.

Ruslan Salakhutdinov and Andriy Mnih. Bayesian probabilistic matrix factorization using markov chain monte carlo. In *Proceedings of the 25th international conference on Machine learning*, pages 880–887, 2008.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. arXiv preprint arXiv:1708.00489, 2017.

Burr Settles. Active learning literature survey. Technical Report 1648, University of Wisconsin-Madison Department of Computer Sciences, 2009.

Masashi Sugiyama and Greg Ridgeway. Active learning in approximately linear regression based on conditional expectation of generalization error. *Journal of Machine Learning Research*, 7(1):141–166, 2006.

Danica J Sutherland, Barnabás Póczos, and Jeff Schneider. Active learning and search on low-rank matrices. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 212–220, 2013.

Edward Vul, Noah Goodman, Thomas L Griffiths, and Joshua B Tenenbaum. One and done? Optimal decisions from very few samples. *Cognitive Science*, 38(4):599–637, 2014.

Teng Xiao, Shangsong Liang, Weizhou Shen, and Zaiqiao Meng. Bayesian deep collaborative matrix factorization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5474–5481, 2019.

Liang Xiong, Xi Chen, Tzu-Kuo Huang, Jeff Schneider, and Jaime G Carbonell. Temporal collaborative filtering with Bayesian probabilistic tensor factorization. In *Proceedings of the 2010 SIAM International Conference on Data Mining*, pages 211–222, 2010.

Yun Yang and David B Dunson. Sequential Markov chain Monte Carlo. arXiv preprint arXiv:1308.3861, 2013.

Gal Yona, Shay Moran, Gal Elidan, and Amir Globerson. Active learning with label comparisons. arXiv preprint arXiv:2204.04670, 2022.

Yangyang Yu and Jordan Suchow. Deep tensor factorization models of first impressions. In *NeurIPS 2022 SVRHM Workshop*, 2022.

Chelsea Zhang, Sean J Taylor, Curtiss Cobb, and Jasjeet Sekhon. Active matrix factorization for surveys. *The Annals of Applied Statistics*, 14(3):1182–1206, 2020.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.