Collaborative filtering in latent space: a Bayesian approach for cold-start music recommendation

Menglin Kong[1], Li Fan[2], Shengze Xu[3], Xingquan Li[4], Muzhou Hou[1], and Cong Cao[1]✉[0000−0002−6853−6421] *

[1] School of Mathematics and Statistics, Central South University, Changsha, China
{212112025,hmzw,congcao}@csu.edu.cn
[2] School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou
[3] Department of Mathematics, The Chinese University of Hong Kong, Hong Kong
[4] Peng Cheng Laboratory, Shenzhen, China

Abstract. Personalized music recommendation technology is effective in helping users discover desired songs. However, accurate recommendations become challenging in cold-start scenarios with newly registered or limited data users. To address the accuracy, diversity, and interpretability challenges in cold-start music recommendation, we propose CFLS, a novel approach that conducts collaborative filtering in the space of latent variables based on the Variational Auto-Encoder (VAE) framework. CFLS replaces the standard normal distribution prior in VAE with a Gaussian process (GP) prior based on user profile information, enabling consideration of user correlations in the latent space. Experimental results on real-world datasets demonstrate the effectiveness and superiority of our proposed method. Visualization techniques are employed to showcase the diversity, interpretability, and user-controllability of the recommendation results achieved by CFLS.

Keywords: · Music Recommendation · Bayesian Inference · Variational Auto-Encoder · Gaussian Process.

1 Introduction

Music is a popular leisure and entertainment activity in people's daily lives. However, with the increasing problem of information overload, it has become challenging for users to efficiently discover songs of their interest from a vast music library [18]. To address this issue, personalized music recommendation technology has emerged as the most effective method to help users quickly find their desired songs [15]. Personalized music recommendation is a service that utilizes deep learning (DL) and machine learning (ML) technology to offer music suggestions based on users' music preferences, interests, and behavioral data [20]. Its primary objective is to enhance users' music experience, increase user engagement on music platforms, and drive the growth of the music industry[3].

As a typical representative of the more generalized recommender system task of Sequential Recommendation (SR), personalized music recommendation systems encounter similar challenges as SR in practice, including (i) **data sparsity**

[*] Menglin Kong and Li Fan contributed equally to this work. Cong Cao is the corresponding author(congcao@csu.edu.cn).

and cold starts [22,25]. Recommender systems face the cold-start problem when dealing with newly registered users or users with limited historical data. Making accurate personalized recommendations without sufficient information about their preferences becomes difficult. (ii) **balancing diversity and personalization** [7]. Personalized music recommendation systems aim to provide users with recommendations that align with their preferences. While recommendations should match users' interests, it is also important to introduce new music that may differ from their previous preferences, allowing them to explore and discover fresh content. (iii) **interpretability and user controllability** [24]. Complex ML/DL algorithms used for recommendations often hinder users' understanding of why certain recommendations are given. Simultaneously, users desire control over the recommendation process, including the ability to customize or adjust the results. Users and music platforms are most concerned about the cold start problem because it is directly related to the user experience and platform revenue [8].

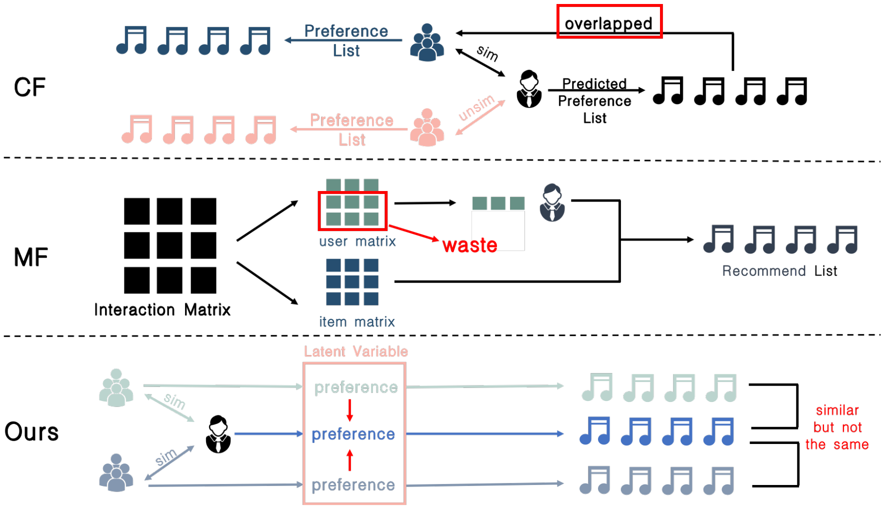


Fig. 1: The basic idea of different cold-start music recommendation methods.

As shown in Figure 1, existing personalized music recommendation systems can be classified into two categories in solving the cold-start problem: **collaborative filtering (CF)**-based approach [15,20] and **matrix factorization (MF)**-based approach [18,3]. The former utilizes the similarity between users to recommend songs from neighboring users' song lists to cold-start users. However, it suffers from poor accuracy in scenarios where data is sparse and users have limited interaction records. Furthermore, it lacks the capability to provide diverse recommendation results. On the other hand, the latter predicts the probability of user-item interactions by employing latent vectors derived from matrix decomposition for users and items. It then selects the item with the highest score as the recommendation. However, this approach fails to fully exploit the rich user behavior information available in the dataset to enhance the recommendation process. Moreover, the latent vectors used for prediction lack explicit semantic information and are not interpretable.

In this paper, we propose a new approach based on Bayesian inference to solve the cold-start problem in music recommendation, taking into account the diversity and interpretability of the recommendation results. The approach involves introducing a latent variable of user preferences, which is determined by the user's profile information that is available to all users in the dataset. Collaborative filtering in the latent space (CFLS) is then performed to infer the preferences of cold-start users. This inference process samples the conditional distributions of the user's behaviors to obtain a candidate set of recommendation results with diversity. Specifically, CFLS employs a neural network (NN) architecture similar to a variational autoencoder (VAE). The NN serves as an encoder, mapping the user's behavioral sequences into the hidden space representing user preference. The decoder NN reconstructs the sampled user preference back into user behavioral sequences. A key difference is that CFLS incorporates a Gaussian Process (GP)-based latent variable prior to account for the similarity among users in the preference hidden space. The covariance function of the GP prior is calculated on the user's profile information. The parameters of the overall model can be optimized using gradient-based methods, enabling accurate and simple out-of-sample predictions. The contributions of this paper are as follows:

- We introduce a new perspective for solving the cold-start problem in music recommendation with additional benefits: Bayesian inference. Sample-based prediction can provide more diversity in recommendation results.
- We specify the prior distribution of user preferences as a GP based on user profile information, which enables us to consider user similarity in the preference latent space and enhances the interpretability of its results.
- By changing the form of the kernel function in the GP prior, different populations can introduce their a priori knowledge about user preferences into the model, leading to user-controllable recommender systems.

2 Related Work and Problem Formulation

Sequential Recommendation (SR): The significance of sequential behaviors in reflecting user preferences has been underscored by studies such as [6] and [25], leading to a surge in interest from both academia and industry in the field of SR. With the emergence of DL, Hidasi et al. [6] pioneered the use of Gated Recurrent Units (GRU) to model sequential behaviors in recommendation systems. Subsequently, researchers have explored a diverse range of deep learning techniques, including Transformers [8], and Large Language Models (LLM) [7,24], to encode interaction sequences. Nevertheless, despite these advancements, it is worth noting that the majority of research in SR has primarily focused on improving recommendation performance, often overlooking the challenging cold-start problems that are inherent in SR [22,25], that is, when the available user-item interactions are very limited, the performance of the SR model decreases dramatically. In this paper, for the cold-start challenge in SR (specifically, music recommendation), we propose a novel method based on Bayesian inference to generate precise recommendation lists for cold-start users by collaborative filtering in the latent space based on the user's profile information.

Cold-Start Recommendation: To address the cold-start problems in recommendations, several primary techniques are commonly employed. Content-based methods aim to incorporate features extracted from side information into CF-based frameworks [21]. Transfer learning methods leverage shared features learned from a source domain to enhance recommendation quality in a target domain [7]. Meta-learning approaches involve a learning-to-learn process across multiple training tasks, optimizing global knowledge to enable rapid adaptation to new recommendation tasks [13]. In this paper, we primarily focus on addressing the cold-start problem in music recommendation, a specific scenario of SR that has been relatively underexplored due to its inherent complexity. While content methods have been proven effective in a large number of recommendation tasks, they struggle to smoothly integrate historical interactions with user profile information in the cold-start scenario [21]. Consequently, researchers have predominantly turned to meta-learning-based methods [25] and LLM-based methods [22] to tackle cold-start challenges in SR, relying on multi-source available user data. However, meta-learning-based methods have high requirements on data volume and data diversity and are more sensitive to the choice of hyperparameters, which increases implementation difficulties [4]; LLM-based methods, on the other hand, lack interpretability of recommendation results [7]. To this end, from the perspective of Bayesian statistics, we propose an improved VAE-based method for generating recommendation lists for cold-start in music recommendation, which effectively fills the gap in existing research by eliminating the need for a large amount of data and possessing good interpretability.

2.1 Problem Formulation

Assume there are $|\mathcal{U}|$ users with historical interaction and profile information (listening habits, length of use, etc.), $|\mathcal{I}|$ songs with item attributes in the dataset, where $\mathcal{U}$ and $\mathcal{I}$ denote the user and item set respectively. For a user $u \in \mathcal{U}$, her profile information is $\mathbf{x}_u \in \mathbb{R}^L$ and interacted sequence is $\mathbf{s}_u = (s_1, s_2, \cdots, s_T)$, where L is the number of available user-related features and T is the length of the sequences. Given a cold-start user u^* with $\mathbf{x}_{u^*}$ and $\mathbf{s}_{u^*}$ (specifically, a full cold-start user when $\mathbf{s}_{u^*} = \emptyset$), we would like to obtain the set of candidate songs $\hat{\mathbf{s}}_{u^*}$ that accurately matches the user's u^* musical preferences, based on $\mathbf{X}$ and $\mathbf{S}$, and $\mathbf{s}_{u^*}, \mathbf{x}_{u^*}$. Where $\mathbf{X} \in \mathbb{R}^{|\mathcal{U}| \times L}$ and $\mathbf{S} = \bigcup_{u \in \mathcal{U}} \mathbf{s}_u$ are the collective information in the dataset.

3 Methodology

3.1 Overview

Our proposed CFLS is similar to the VAE in terms of model structure. Due to the presence of aligned multi-view data (user-item interactions, user profile infor-

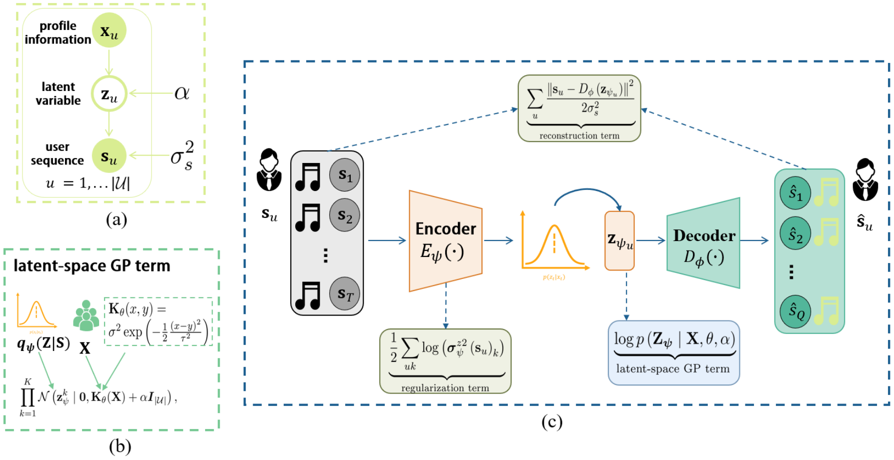

profile information
$\mathbf{x}_u$
latent variable
$\mathbf{z}_u$
α
user sequence
$\mathbf{s}_u$
σ_s^2
$u = 1, \dots |\mathcal{U}|$
(a)
latent-space GP term
$q_\psi(\mathbf{Z}|\mathbf{S})$
$\mathbf{X}$
$\mathbf{K}_\theta(x,y) = \sigma^2 \exp\left(-\frac{1}{2}\frac{(x-y)^2}{\tau^2}\right)$
$\prod_{k=1}^{K} \mathcal{N}\left(\mathbf{z}_\psi^k \mid \mathbf{0}, \mathbf{K}_\theta(\mathbf{X}) + \alpha \boldsymbol{I}_{|\mathcal{U}|}\right),$
(b)
$\mathbf{s}_u$
$\mathbf{s}_1$
$\mathbf{s}_2$
$\mathbf{s}_T$
Encoder $E_\psi(\cdot)$
$p(z_t|x_t)$
$\mathbf{z}_{\psi_u}$
Decoder $D_\phi(\cdot)$
$\hat{s}_1$
$\hat{s}_2$
$\hat{s}_Q$
$\hat{\mathbf{s}}_u$
$\underbrace{\sum_u \frac{\|\mathbf{s}_u - D_\phi(\mathbf{z}_{\psi_u})\|^2}{2\sigma_s^2}}_{\text{reconstruction term}}$
$\underbrace{\frac{1}{2}\sum_{uk} \log\left(\boldsymbol{\sigma}_\psi^{z2}(\mathbf{s}_u)_k\right)}_{\text{regularization term}}$
$\underbrace{\log p\left(\mathbf{Z}_\psi \mid \mathbf{X}, \theta, \alpha\right)}_{\text{latent-space GP term}}$
(c)

Fig. 2: The framework of CFLS. (a) The generative model underlying the proposed CFLS. (b) The formula of latent-space GP term with diagrams. (c) Forward process based on VAE model and loss function in CFLS.

mation, song attributes), in order to fully utilize the information in the dataset to perform cold-start recommendation, we first train a matrix factorization (M-F) [11] model on the original dataset and obtain the pre-trained latent vectors of songs $\mathbf{v}_i \in \mathbb{R}^M$ as the representation of song i. Thus, the interaction sequence of user u can be represented as $\mathbf{s}_u = \frac{1}{|\mathcal{I}_u|}\sum_{j=1}^{|\mathcal{I}_u|}\mathbf{v}_j \in \mathbb{R}^M$, where $\mathcal{I}_u$ denotes the song sets the user u once interacted with.

As can be seen from Figure 2 (c), CFLS consists of two parts: first, an MLP parameterized by ψ (i.e., the encoder $E_\psi(\cdot)$) is used to map user u's sequence of interactions $\mathbf{s}_u$ into the latent space where the user's preference latent variable $\mathbf{z}_u \in \mathbb{R}^K$ are distributed, where $K \ll M$. This process is equivalent to applying a variational posterior $q_\psi(\mathbf{z}_u)$ to approximate the intractable true posterior $p(\mathbf{z}_u|\mathbf{s}_u)$. Subsequently, based on the $\mathbf{z}_u$ sampled from the variational posterior $q_\psi(\mathbf{z}_u)$, a decoder $D_\phi(\cdot)$ parameterized by ϕ is used to map $\mathbf{z}_u$ back to the predicted sequence $\hat{\mathbf{s}}_u$ of of user u . Unlike the original VAE, which specifies the prior distribution of the latent variable $\mathbf{z}_u$ as a standard Gaussian distribution, i.e., the samples are assumed to be independent of each other in the latent space, we specify the prior distribution of $\mathbf{z}_u$ as a GP whose covariance function $\mathbf{K}_\theta(\cdot,\cdot)$'s independent variables are the user's profile information, thus introducing similarity among users into the latent space. Finally, the overall model parameters $\Theta = \{\psi, \phi, \theta\}$ can be optimized by the Stochastic Gradient Variational Bayesian (SGVB) method [23] in an end-to-end manner.

3.2 Statistical model in CFLS

In this section, we describe the data generation process in CFLS in detail from the perspective of a statistical model, including the latent variable of user preferences $\mathbf{z}_u$ based on profile information $\mathbf{x}_u$ and the generation of user behaviors $\mathbf{s}_u$ based on user preferences (Figure 2 (a)).

For a user u, her preference $\mathbf{z}_u$ is generated from profile information $\mathbf{x}_u$:

$$\mathbf{z}_u = f\left(\mathbf{x}_u\right) + \boldsymbol{\eta}_u, \text{ where } \boldsymbol{\eta}_u \sim \mathcal{N}\left(\mathbf{0}, \alpha \boldsymbol{I}_K\right), \tag{1}$$

and sequence $\mathbf{s}_u$ is generated from its preference $\mathbf{z}_u$ as :

$$\mathbf{s}_u = g\left(\mathbf{z}_u\right) + \boldsymbol{\epsilon}_u, \text{ where } \boldsymbol{\epsilon}_u \sim \mathcal{N}\left(\mathbf{0}, \sigma_s^2 \boldsymbol{I}_M\right). \tag{2}$$

CFLS uses a GP prior on f, which enables it to model sample covariances in the latent space as a function of profile information. In this paper, we use an MLP (i.e., decoder $D_\phi(\cdot)$) for g to output the distribution hyperparameters for the user sequence $\mathbf{s}_u$ (i.e., $p_\phi(\mathbf{s}_u|\mathbf{z}_u)$). The formulation is as follows:

$$p\left(\mathbf{S} \mid \mathbf{X}, \phi, \sigma_s^2, \theta, \alpha\right) = \int p\left(\mathbf{S} \mid \mathbf{Z}, \phi, \sigma_s^2\right) p(\mathbf{Z} \mid \mathbf{X}, \theta, \alpha) d\mathbf{Z}, \tag{3}$$

where $\mathbf{S} = \left[\mathbf{s}_1, \ldots, \mathbf{s}_{|\mathcal{U}|}\right]^T \in \mathbb{R}^{|\mathcal{U}| \times M}$, $\mathbf{Z} = \left[\mathbf{z}_1, \ldots, \mathbf{z}_{|\mathcal{U}|}\right]^T \in \mathbb{R}^{|\mathcal{U}| \times K}$, $\mathbf{X} = \left[\mathbf{x}_1, \ldots, \mathbf{x}_{|\mathcal{U}|}\right]^T \in \mathbb{R}^{|\mathcal{U}| \times L}$. Additionally, ϕ denotes the weights and bias of the decoder and θ the learnable GP kernel parameters.

Gaussian Process Prior for $\mathbf{z}_u$. The prior distribution of user preference $\mathbf{z}_u$ is defined as the following multivariate normal distribution:

$$p(\mathbf{Z} \mid \mathbf{X}, \theta, \alpha) = \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{z}^k \mid \mathbf{0}, \mathbf{K}_\theta(\mathbf{X}) + \alpha \boldsymbol{I}_{|\mathcal{U}|}\right), \tag{4}$$

where $\mathbf{z}^k$ is the k-th column of $\mathbf{Z}$. In our implementation, a squared exponential (SE) kernel is selected for $\mathbf{K}_\theta(\cdot,\cdot)$ following the setting in [1,2]. Specifically, for the user u and u', there sample covariance is given by :

$$\mathbf{K}_\theta(\mathbf{X})_{uu'} = \sigma_f^2 \exp\left(-\frac{1}{2}\sum_{l=1}^{L}\frac{\left(x_u^l - x_{u'}^l\right)^2}{\tau_l^2}\right), \tag{5}$$

where $\theta = [\sigma_f, \tau_1, \cdots, \tau_L]$, σ_f is the standard hyperparameter and τ_l is the lengthscale along each individual input direction. When x^l is a sparse feature, we first train a factorization machine (FM) model [17] to obtain the dense low-dimensional hidden vector $\tilde{\mathbf{x}}^l$ of feature l, and then compute the sample covariance based on $\tilde{\mathbf{x}}_u^l$ and $\tilde{\mathbf{x}}_{u'}^l$.

Variational Posterior for $\mathbf{z}_u$. As with a standard VAE, we employ an Encoder $E_\psi(\cdot)$ to output the hyperparameters of the variational posterior $q_\psi(\mathbf{z}_u|\mathbf{s}_u)$, which is optimized to approximate the intractable true posterior $p(\mathbf{s}_u|\mathbf{z}_u)$:

$$q_\psi(\mathbf{Z} \mid \mathbf{S}) = \prod_u \mathcal{N}\left(\mathbf{z}_u \mid \boldsymbol{\mu}_\psi^z\left(\mathbf{s}_u\right), \operatorname{diag}\left(\boldsymbol{\sigma}_\psi^{z2}\left(\mathbf{s}_u\right)\right)\right), \tag{6}$$

where ψ denotes the weights and bias of the MLP for encoder $E_\psi(\cdot)$, $\boldsymbol{\mu}_\psi^z\left(\mathbf{s}_u\right)$ and $\operatorname{diag}\left(\boldsymbol{\sigma}_\psi^z\left(\mathbf{s}_u\right)\right)$ are the hyperparameters of the variational distribution and output by $E_\psi(\mathbf{s}_u)$. Specifically, latent user preference $\mathbf{z}_{\psi_u}$ are sampled using the re-parameterization trick in [10], that is :

$$\mathbf{z}_{\psi_u} = \mu_\psi^z\left(\mathbf{s}_u\right) + \boldsymbol{\delta}_u \odot \sigma_\psi^z\left(\mathbf{s}_u\right), \boldsymbol{\delta}_u \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{I}_{K\times K}\right), \tag{7}$$

where $\odot$ denotes the Hadamard product.

3.3 Optimization

Based on the model and notations defined above, we have the following evidence lower bound (ELBO) for the likelihood in Equation (3):

$$\begin{aligned} \log p\left(\mathbf{S} \mid \mathbf{X}, \phi, \sigma_s^2, \theta\right) \geq & \mathbb{E}_{\boldsymbol{Z}\sim q_\psi}\left[\sum_u \log \mathcal{N}\left(\mathbf{s}_u \mid D_{\boldsymbol{\phi}}\left(\mathbf{z}_u\right), \sigma_s^2 \boldsymbol{I}_M\right) + \log p(\mathbf{Z} \mid \mathbf{X}, \theta, \alpha)\right] + \\ & + \frac{1}{2}\sum_{uk} \log\left(\boldsymbol{\sigma}_\psi^{z2}\left(\mathbf{s}_u\right)_k\right) + \text{ const.} \end{aligned} \tag{8}$$

Inspired by [23], we employ stochastic backpropagation to maximize the above ELBO. By sampling latent variable $\mathbf{z}_{\psi_u}$ from a reparameterized variational posterior $q_\psi(\mathbf{z})$, we approximate the expectation in (8) and obtain a loss function as follows:

$$\begin{aligned} & \mathcal{L}\left(\phi, \psi, \theta, \alpha, \sigma_s^2\right) = \\ & = |\mathcal{U}| M \log \sigma_s^2 + \underbrace{\sum_u \frac{\left\|\mathbf{s}_u - D_\phi\left(\mathbf{z}_{\psi_u}\right)\right\|^2}{2\sigma_s^2}}_{\text{reconstruction term}} - \underbrace{\log p\left(\mathbf{Z}_{\boldsymbol{\psi}} \mid \mathbf{X}, \theta, \alpha\right)}_{\text{latent-space GP term}} + \underbrace{\frac{1}{2}\sum_{uk} \log\left(\boldsymbol{\sigma}_\psi^{z2}\left(\mathbf{s}_u\right)_k\right)}_{\text{regularization term}} \end{aligned} \tag{9}$$

By minimizing (9) via mini-batch stochastic gradient descent (SGD), the optimal parameters $\mathit{\Theta}^* = \left\{\phi^*, \psi^*, \theta^*, \alpha^*, \sigma_s^{2*}\right\}$ can be obtained by end-to-end.

3.4 Prediction

Given a cold-start user u^* with $\mathbf{x}_{u^*}$, we would like to obtain the set of candidate songs $\hat{\mathbf{s}}_{u^*}$ that accurately matches the user's u^* musical preferences, based on $\mathbf{X}$ and $\mathbf{S}$, and $\mathit{\Theta}^*$. The predictive posterior for $\hat{\mathbf{s}}_{u^*}$ is given by:

$$p\left(\hat{\mathbf{s}}_{u^*} \mid \mathbf{x}_{u^*}, \mathbf{S}, \mathbf{X}\right) \approx \int p_{\phi^*}\left(\hat{\mathbf{s}}_{u^*} || \mathbf{z}_{u^*}\right) p_{\theta^*}\left(\mathbf{z}_{u^*} \mid \mathbf{x}_{u^*}, \mathbf{Z}_{\psi^*}, \mathbf{X}\right) q_{\psi^*}(\mathbf{Z} \mid \mathbf{S}) d\mathbf{z}_{u^*} d\mathbf{Z}. \tag{10}$$

The candidate songs $\hat{\mathbf{s}}_{u^*}$ can be obtained in the prediction stage by the following procedure : (i) encode training data to get $\mathbf{Z}_{\psi^*}$ via encoder E_{ψ^*}, (ii) sample $\mathbf{z}_{u^*}$ from the GP predictive posterior $p_{\theta^*}\left(\mathbf{z}_{u^*} \mid \mathbf{x}_{u^*} \mathbf{Z}_{\psi^*}, \mathbf{X}\right)$, (iii) decode $\mathbf{z}_{u^*}$ to get the item embedding of $\hat{\mathbf{s}}_{u^*}$, (iv) search for top Q nearest neighbors of $\hat{\mathbf{s}}_{u^*}$ in the item set $\mathcal{I}$ based on their location in the embdedding space, return the final recommendation list for u^*.

4 Experiments

In this section, we present the industrial dataset used for the experiments, the settings for the experiments, including the evaluation protocols and baseline models, and finally, the results of the experiments and related analysis. Moreover, we demonstrate the diversity, interpretability, and user controllability of CFLS's recommendation results through data visualization.

4.1 Dataset

In this study, we validate the effectiveness of CFLS based on the #**nowplaying-RS** dataset [16]. This dataset comprises listening data from Twitter collected in 2014, including 138,150 users, 346,646 songs, and 11,606,689 listening records. Each dataset's listening record contains rich song features and user profile information. For a more comprehensive understanding of the dataset and its features and detailed statistical information, please refer to [16].

4.2 Experimental settings

Evaluation Protocol. Following previous cold-start recommendation and SR works [25], we utilize the leave-one-out method to evaluate the recommendation performance, and set Precision@K, Recall @K, MAP@K and NDCG@K as the evaluation metrics. In order to verify the superiority and robustness of CFLS in cold-start recommendation scenarios, we introduce the hyperparameter r to denote the proportion of completely cold-start users in the test set to evaluate the performance variation of CFLS and baseline models when r takes different values. We repeated each set of experiments 10 times to ensure the stability of the results and reported their mean values as the final results.

Baseline Models. We selected a bunch of recent DL-based recommendation models as baseline models, including DSSM [9], YoutubeDNN [5], MIND [12], SDM [14] and Bert4Rec [19]. The hyperparameters of all baseline models are consistent with those reported in their paper. In addition, in order to verify the validity of the proposed GP prior, we introduce the CFLS w/o GP prior, a variant that uses the standard normal distribution as the prior distribution of the latent variables of user preferences.

Implementation and Hyperparameters. For CFLS, the hyperparameters are as follows: the training epoch is fixed as 10, the early stopping patience is fixed as 3, the item embedding dimension M is fixed as 128, the latent variable dimension K is fixed as 64, the mini-batch size is 64, the Adam [23] is used for optimization and the learning rate is 0.001. Both encoder and decoder are two-layer MLP with ReLU as an activation function.

Table 1: Performance comparisons of different methods

	Precision@K				**Recall@K**				**MAP@K**				**NDCG@K**			
Model	K=20	K=40	K=60	K=80	K=20	K=40	K=60	K=80	K=20	K=40	K=60	K=80	K=20	K=40	K=60	K=80
DSSM [9]	8.65%	7.75%	7.42%	6.93%	17.30%	31.01%	44.52%	55.44%	9.88%	9.08%	8.47%	8.18%	10.42%	9.32%	8.82%	8.35%
YoutubeDNN [5]	8.83%	7.96%	7.60%	7.05%	17.66%	31.85%	45.61%	56.40%	10.14%	9.42%	8.72%	8.41%	10.60%	9.58%	8.90%	8.54%
MIND [12]	9.11%	8.29%	7.97%	7.23%	18.22%	33.16%	47.82%	57.84%	10.35%	9.63%	8.95%	8.64%	10.82%	9.79%	9.15%	8.75%
SDM [14]	9.23%	8.45%	8.13%	7.37%	18.46%	33.82%	48.78%	58.94%	10.48%	9.77%	9.16%	8.76%	11.08%	10.02%	9.39%	8.93%
Bert4Rec [19]	9.17%	8.35%	8.03%	7.34%	18.35%	33.38%	48.17%	58.70%	10.38%	9.60%	9.00%	8.55%	10.80%	9.85%	9.32%	8.82%
CFLS w/o GP prior	9.05%	8.32%	7.99%	7.33%	18.11%	33.26%	47.92%	58.61%	10.36%	9.56%	8.96%	8.55%	10.85%	9.83%	9.27%	8.74%
CFLS	**9.46%**	**8.62%**	**8.28%**	**7.50%**	**18.92%**	**34.48%**	**49.68%**	**59.99%**	**10.72%**	**10.01%**	**9.28%**	**8.91%**	**11.21%**	**10.12%**	**9.63%**	**9.09%**

4.3 Performance comparisons

In this section, we specify the percentage of completely cold-start users in the test set $r = 0.7$ and give the evaluation results for the four evaluation metrics for K =20, 40, 60, 80. The specific values are shown in Table 1. Bolded numbers are optimal representations, and underlined numbers are sub-optimal. We use the DSSM as a baseline for calculating relative lift to facilitate a comparison of model performance. The following observations can be obtained from Table 1:

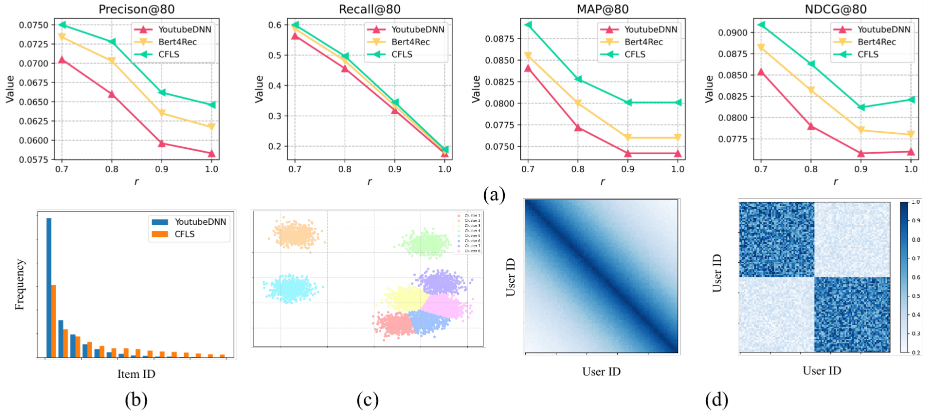
The DSSM model gives the worst performance in all the comparison experiments. This is due to the fact that DSSM only considers user profile information while modeling user preferences and does not utilize user behavioral information, resulting in the inability to achieve personalized music recommendations. YoutubeDNN introduces a heterogeneous subnetwork to model user behavioral information, thereby obtaining more expressive representations of user interests, which improves the NDCG@80 metric by 2.3%. MIND introduces a dynamic routing mechanism to model the user's interest evolution, while SDM uses a session-based approach to separately model the user's interest into long and short periods. Both achieve finer-grained user interest modeling by introducing a new induction bias, resulting in relative performance improvements of 4.79% and 6.95% in NDCG@80, respectively. Among them, SDM achieves the sub-optimal performance in all comparison experiments Bert4Rec utilizes pre-trained Bert to encode information about users and items and introduce knowledge from other domains to represent user interests better. However, when the pre-training is based on data unrelated to the target task, it may lead to performance degradation. This is demonstrated by the increase of 5.63% of NDCG@80, lower than SDM and CFLS.

Comparing the results of CFLS w/o GP prior and CFLS can verify the validity of our proposed GP prior: replacing the standard normal prior for the latent variable of user preferences in the naive VAE with the profile information-based GP prior, our method achieves a 4% improvement on NDCG@80. CFLS relative to DSSM improves by 8.86% and achieves optimal performance among all compared models, demonstrating the superiority of modeling user similarity and performing collaborative filtering in the latent space.

4.4 Influence of different cold-start levels

In this section, in order to verify the robustness of the performance of CFLS to the percentage r of fully cold-started users in the test set, we set up four sets of experiments by setting the values of the hyperparameter r to 0.7, 0.8, 0.9, and 1.0, i.e., gradually increasing the percentage of fully cold-started users in the test set.

As can be seen in Figure 3 (a), with the gradual improvement of r, the performance of YoutubeDNN, Bert4rec, and CFLS on all 4 metrics decreases to different degrees. However, compared with the other two methods, CFLS has a smaller decrease, and the overall change trend is relatively stable and always maintains the optimal performance. This verifies the robustness of our proposed CFLS method, which is less affected by the percentage of completely cold-start users r in the test set.



(a)

(b) (c) (d)

Fig. 3: (a) Robustness of different methods to variations in of r, the percentage of fully cold-started users in the test set. (b) Occurrence frequency of items with different popularity in YoutubeDNN and CFLS recommendation results. (c) Distribution of user preference latent variables encoded by CFLS in 2D space. (d) The covariance matrix of user preference latent variables was obtained by using different kernel functions in the GP prior.

4.5 Diversity, interpretability and user controllability

In this section, we first evaluate the diversity of CFLS's recommendation results. Specifically, we select a part of popular items and a part of cold items, arrange them in descending order of their frequency of appearance in the test set, and then count the frequency of their appearances in the final recommendation results of YoutubeDNN and CFLS, respectively, and the results are shown in Fig. 3 (b). It can be seen that compared with YoutubeDNN, CFLS, which obtains recommendation results based on sampling, has more significant diversity, which is specifically reflected in the fact that cold items can also get some exposure opportunities, and the exposure opportunities of popular items are suppressed, which reduces the popularity bias to a certain extent.

To illustrate the good interpretability of user preference latent variables encoded in CFLS, we first cluster users based on their profile information. The user preference latent variables $\mathbf{z}_\psi$ obtained from the encoder are then subjected to PCA dimensional reduction to visualize their distribution in 2D space. It can be seen from Figure 3 (c) that $\mathbf{z}_\psi$ reflects the similarity between users very well, which justifies our collaborative filtering in the latent space.

Finally, we compare the user covariance matrices obtained using the two kernel functions mentioned in the literature [1] (SE kernel and automatic relevance determination (ARD) kernel) as covariance functions for the GP prior. We choose the SE kernel in our implementation because we believe that the preferences of neighboring users in the feature space should have greater correlation in the latent space and that the change in this correlation is smooth (as shown on the left in Figure 3 (d)); however, if others believe that there are two groups of preference in the users that are more different and have greater intra-group correlation and less inter-group correlation, they can choose the ARD kernel (as in the right of Figure 3 (d)). This validates the user-controllability of CFLS, i.e., users can flexibly adapt the model in both training and prediction stage based on their prior knowledge.

5 Conclusions

In this paper we present a novel Bayesian inference-based approach to address the cold-start problem in music recommendation. Our method incorporates diversity and interpretability considerations, utilizing the VAE framework. We introduce a Gaussian Process (GP) prior for user preferences latent variable, leveraging user profile information to account for user similarity in the preference latent space. Experimental results on real-world datasets demonstrate the efficacy and superiority of our proposed method. Additionally, we employ visualization techniques to showcase the diversity, interpretability, and user-controllability of the recommendation results achieved by our approach.

Acknowledgements This study was supported by Natural Science Foundation of Hunan Province (grant number 2022JJ30673) and by the Graduate Innovation Project of Central South University (2023XQLH032, 2023ZZTS0304).

References

1. N. Botteghi, M. Guo, and C. Brune. Deep kernel learning of dynamical models from high-dimensional noisy data. *Scientific reports*, 12(1):21530, 2022.
2. F. P. Casale, A. Dalca, L. Saglietti, J. Listgarten, and N. Fusi. Gaussian process prior variational autoencoders. *Advances in neural information processing systems*, 31, 2018.
3. K. Chen, B. Liang, X. Ma, and M. Gu. Learning audio embeddings with user listening data for content-based music recommendation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3015–3019. IEEE, 2021.
4. Z. Chu, H. Wang, Y. Xiao, B. Long, and L. Wu. Meta policy learning for cold-start conversational recommendation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 222–230, 2023.
5. P. Covington, J. Adams, and E. Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, pages 191–198, 2016.
6. B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939*, 2015.
7. Y. Hou, Z. He, J. McAuley, and W. X. Zhao. Learning vector-quantized item representation for transferable sequential recommenders. In *Proceedings of the ACM Web Conference 2023*, pages 1162–1171, 2023.
8. Y. Hou, S. Mu, W. X. Zhao, Y. Li, B. Ding, and J.-R. Wen. Towards universal sequence representation learning for recommender systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 585–593, 2022.
9. P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2333–2338, 2013.

10. D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
11. Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.
12. C. Li, Z. Liu, M. Wu, Y. Xu, H. Zhao, P. Huang, G. Kang, Q. Chen, W. Li, and D. L. Lee. Multi-interest network with dynamic routing for recommendation at tmall. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 2615–2623, 2019.
13. Y. Lu, Y. Fang, and C. Shi. Meta-learning on heterogeneous information networks for cold-start recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1563–1573, 2020.
14. F. Lv, T. Jin, C. Yu, F. Sun, Q. Lin, K. Yang, and W. Ng. Sdm: Sequential deep matching model for online large-scale recommender system. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2635–2643, 2019.
15. Y. Mao, G. Zhong, H. Wang, and K. Huang. Music-crn: An efficient content-based music classification and recommendation network. *Cognitive Computation*, 14(6):2306–2316, 2022.
16. A. Poddar, E. Zangerle, and Y.-H. Yang. nowplaying-rs: a new benchmark dataset for building context-aware music recommender systems. In *Proceedings of the 15th Sound & Music Computing Conference*, pages 21–26, 2018.
17. S. Rendle, C. Freudenthaler, and L. Schmidt-Thieme. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 811–820, 2010.
18. J. Shen, M. Tao, Q. Qu, D. Tao, and Y. Rui. Toward efficient indexing structure for scalable content-based music retrieval. *multimedia systems*, 25:639–653, 2019.
19. F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, and P. Jiang. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1441–1450, 2019.
20. A. Van den Oord, S. Dieleman, and B. Schrauwen. Deep content-based music recommendation. *Advances in neural information processing systems*, 26, 2013.
21. L. Wu, C. Quan, C. Li, Q. Wang, B. Zheng, and X. Luo. A context-aware user-item representation learning for item recommendation. *ACM Transactions on Information Systems (TOIS)*, 37(2):1–29, 2019.
22. Y. Wu, R. Xie, Y. Zhu, F. Zhuang, X. Zhang, L. Lin, and Q. He. Personalized prompts for sequential recommendation. *arXiv preprint arXiv:2205.09666*, 2022.
23. H. Yu, T. Nghia, B. K. H. Low, and P. Jaillet. Stochastic variational inference for bayesian sparse gaussian process regression. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.
24. Z. Yuan, F. Yuan, Y. Song, Y. Li, J. Fu, F. Yang, Y. Pan, and Y. Ni. Where to go next for recommender systems? id-vs. modality-based recommender models revisited. *arXiv preprint arXiv:2303.13835*, 2023.
25. Y. Zheng, S. Liu, Z. Li, and S. Wu. Cold-start sequential recommendation via meta learner. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4706–4713, 2021.