We thank the reviewers for the valuable suggestions and appreciate the positive feedback.

1 To both Reviewer #1 & Reviewer #2:

1.1 How is the number of components K decided? What if K differs from the ground truth?

We do not directly set K to the ground truth when evaluating its performance on recommendation tasks (Table 1), so as to ensure a fair comparison with the baselines. Instead, we set K as a sufficiently large value initially and shrink its value during training if the JS divergence between $\{p_{i|k}\}_{i=1}^{M}$ and $\{p_{i|k'}\}_{i=1}^{M}$ for some $k \neq k'$ is negligible compared to a predefined threshold, where $p_{i|k} := p_\theta(c_i = k)/\sum_{i'} p_\theta(c_{i'} = k)$. As for Figure 2 & 3, which are about interpretability, we set K to the ground truth, i.e., $K = 7$. We will revise the paper to ensure these details are included.

We experiment with $K \in \{1, 2, \ldots, 20\}$ on dataset Shop-7C and have the following observations. (1) When K is much smaller than the ground-truth value, the performance of our approach on recommendation tasks degrades and becomes close to or even slightly below that of the baselines. The quality of the micro disentanglement also suffers in the sense that the dimensions become more correlated and less interpretable. Note that Figure 4 from Subsection 3.4 also shows that the learned representations are much less micro-disentangled when $K = 1$ is used in place of $K = 7$. (2) On the other hand, when K is larger than the ground truth, the performance rarely improves, and the degradation is not as severe when it happens. We will update the supplemental material to include the relevant results.

2 To Reviewer #1

2.1 On how our work is fundamentally different from hierarchical recommender systems.

The traditional hierarchical methods usually cluster items and/or users to abstract representations at a higher level of granularity, while our disentangled approach can further decompose an item as well as a user's preference according to the micro factors (e.g., size, color, and price) to obtain fine-grained interpretable representations. The latter direction is a largely unexplored topic in the literature of recommender systems that may enable potential novel applications such as user-controllable recommendation. We adopt the hierarchical design mainly to accommodate the fact that different categories (or macro factors) can be associated with distinctly different sets of micro factors (see Q&A 2.2 below).

2.2 On the relevance of the independence and sufficiency assumptions in real applications.

Actually, it is the real-world application of recommender systems that motivates us to assume the independence and sufficiency. (1) Industrial recommender systems involve highly diverse items, and a user's preference is closely connected to the product categories. A user's preference regarding battery life is applicable to laptops but not to lipsticks. Even for a common property such as price, the user may simultaneously prefer low-priced laptops and high-priced lipsticks. Such complicated user preference can only be effectively discovered if the assumptions of independence and sufficiency at the macro level are imposed. (2) On the other hand, encouraging independence between the dimensions, which is a common idea behind many disentangled representation learning methods, benefits interpretability, because it facilitates the discovery of semantically independent micro factors such as sizes, colors, and prices, which are valuable for explainable recommendation. (3) We can indeed relax the assumptions by instead using a hierarchical prior such as the normal-inverse-Wishart prior. However, we value scalability more than the marginal improvement.

3 To Reviewer #3

3.1 Whether the reported baselines are enough.

Our work focuses more on interpretability and controllability rather than performance. Admittedly, more baselines will necessarily make the results more convincing. However, we must note that the baseline [30] that we do compare with is the state-of-the-art. The recent results of arXiv:1907.06902 show that Liang's method [30] (1) is the strongest neural approach when it comes to top-n recommendation tasks and (2) outperforms or is at least on par with the strongest non-neural ranking methods such as SLIM. Note that SLIM is typically stronger than BPR variants such as BPR-MF and BPR-kNN. Nevertheless, we are willing to include more baselines such as SLIM and BPRs in the revised version.

3.2 Ablation studies of the macro and micro disentanglement, e.g., by setting $K = 1$.

These results can be found in Subsection 3.4 (see Figure 4). Figure 4 shows that (1) $K = 7$ leads to a higher level of micro disentanglement than $K = 1$, which indicates the necessity of macro disentanglement, and that (2) strengthening micro disentanglement often brings better performance. We will revise the paper to make these results more visible.

3.3 Quantitative measurement of the micro disentanglement.

We have quantitatively measured the statistical independence of the dimensions in Subsection 3.4. The results (see Figure 4) show that our approach significantly improves upon the baselines when it comes to micro disentanglement.

3.4 Is there a reason to account for the superior performance, especially on sparse data?

The macro-micro design alleviates data sparsity by allowing a rarely visited item to borrow information from other items of the same category, which is the motivation behind many hierarchical methods [49]. On the other hand, sparse data tend to affect the robustness and stability of an algorithm, and Bengio et al. [3] suggest that disentangled representations are more robust since prediction with such representations tends to be stable when a few nuisance factors vary.

3.5 Whether we plan to release the source code.

Yes, we plan to release the source code, along with all the datasets used, once the paper is publicly published.

Reference [49] Zhang & Koren. Efficient Bayesian hierarchical user modeling for recommendation systems. SIGIR'07.