Reviewers remark our method is intuitive and correct, and opens new directions in sparse clustering, while R1 raised a concern about the extent of our contributions. Most comments mention the paper is well written while providing constructive suggestions to further improve the presentation. We thank the referees for their time and feedback, and provide detailed responses below:

R1:
We thank you for commenting the paper is well-written and for finding a typo. We hope you might reconsider the novelty of our work if you find these responses to sufficiently address your concerns:
− You suggest better baselines for comparison, citing Power *k*-means [37] and matrix + tensor factorization. Our manuscript *does* cite [37], and does compare with this method—see Table 3. Because power *k*-means is not designed for the high-dimensional setting, it performed nearly identically with *k*-means in the simulated experiments where we know the ground truth is sparse, and was omitted due to redundancy. We are happy to include its performance in those comparisons as well, which will convey the same trends.
− Your intuition for low-rank matrix factorization is spot on, but low-rank factorization plus the constraint that one seeks "hard" label assignments becomes **equivalent** to *k*-means, which we consider here. See for instance "*k*-means Clustering Is Matrix Factorization" (Bauckhage 2015). Note that the number of clusters k is analogous to the rank k of the low-rank factor; it remains nontrivial to perform feature selection jointly (i.e. simultaneously seek a sparse number s of informative features) as we do in the proposed method. In light of your comment, we will emphasize the connection to matrix factorization in the revision. We do not compare to tensor factorization as all data we consider are vector-valued and not matrix-valued.
− We respectfully disagree that the contribution is incremental, as the ranking-based feature selection is a marked departure from the existing efforts which largely either rely on generic dimension reduction as a pre-processing step, or penalization via norm-shrinkage. Instead, the proposed method allows for the exact desired number of nonzero features to be specified as input, and yields a scalable approach that is appropriate in high-dimensional settings yet comparable to Lloyd's algorithm in terms of simplicity and speed. As a result, our algorithm looks and functions quite differently than past work on sparse *k*-means, which we have reviewed and compared to our method.

R2:
We thank you for your detailed comments and careful reading of the paper. We will further elaborate on the choice of within-cluster sum of squares score as suggested during the revision.

R3:
We agree that the benchmarks and simulation details can be better described and will provide complete details in the revision. You raise a good point about further competing methods– regarding COSA, we were unable to find the authors' implementation and had implemented our own version whose runtime and performance was far worse than the proposed method. We should also note that COSA was designed for feature weighing rather than selection, and does not typically result in sparse solutions, though we will attempt to add a fair, detailed comparison in the revision. We also note your suggestion regarding filter methods that focus only on feature selection and will include a comparison in the final version.

R4:
You are absolutely correct that "interpretable sparsity" is overloaded here. We will clarify in the revision its twofold meaning: first as you've mentioned, we can inform or "control" the sparsity level via parameter s. In this sense the parameter s is directly interpretable compared to parameters such as λ in existing ℓ_1 approaches. Second, our ranking method selects features among the original dimensions, thus allowing them to retain their original interpretation. For instance, in our mouse protein study it is important that the top ranked features identified by SKFR identify the most relevant genes, as the original features correspond to expression levels along a high-dimensional space of candidate genes. In this sense the *dimension reduction* is interpretable, in contrast to generic dimension reduction such as PCA, where the axes (principal components) in the projected space lose their interpretation as genes. We will improve the exposition to emphasize this, as well as provide further detail in a comparison table with sparse *k*-means focusing on selection in the Supplement of the final draft, and thank you for these constructive comments.