Title: Towards the systematic reconnaissance of seismic signals from glaciers and ice sheets - Part B: Unsupervised learning for source process characterisation
Author(s): Rebecca B. Latto et al.
MS No.: egusphere-2023-1341
MS type: Research article

Please access the discussion at:
https://egusphere.copernicus.org/preprints/2023/egusphere-2023-1341/#discussion

---

Below, the R1 comments are copied in grey. Author Comments continue in blue.

General Comments

The manuscript presents a thoughtful and accessible methodology for performing clustering analysis for glacial seismology. Given increased interest in continuous seismic monitoring of the cryosphere, the study is timely and instructive and would be of value to members of both the seismological and cryosphere communities. Furthermore, the manuscript is well written and strikes an appropriate balance between introducing basic theory and demonstrating careful application and thorough analysis.

Many thanks for this positive appraisal, and an expert review that probes the analysis (including the supplement) very thoroughly. The contribution will benefit significantly from clarifications and additions suggested, and we are happy to follow all suggestions made.

One of the chief concerns I have is the utility of k-means on a parameter space with 30 dimensions. Though the authors took care to appropriately select features for clustering, the procedure was not conceived with dimensionality reduction in mind, nor is dimensionality reduction mentioned once in the literature review, although several papers were cited that make use of it (e.g., autoencoders, PCA). Given the dubious utility of distance-based clustering metrics in high dimensions (Aggarwal et al. 2001; Aggarwal & Reddy 2014), I would at least expect acknowledgement of this limitation. I hesitate to require re-performance of what is already a substantial analysis presented in this manuscript, but a good candidate for future work would be repeating the analysis but with reduced dimensionality, e.g., using the top 5 or 10 components from PCA.

We agree that the reviewer and suggested references articulate important considerations, and we are happy to include a new paragraph in the interpretation and discussion section that addresses these points.

My other main critique is the use of the k-means(++) algorithm. K-means has several important limitations, namely that it performs poorly with overlapping clusters and that it assumes equal variance in all dimensions. It is sensitive to outliers, which is exacerbated by the high dimensionality of the feature space. I would instead suggest using an expectation-maximization (EM) approach to refine the cluster definitions, as provided in Gaussian mixture model (GMM) clustering. GMM handles multivariate distributions and overlapping clusters better than k-means, and is trivial to implement with scikit-learn. However, GMM is still subject to the curse of dimensionality.

We made the choice to use k-means(++) based on its transparency for a reconnaissance workflow. We are happy to note this, and also include the points raised in the interpretation and discussion section, and are grateful for the suggestion to point readers to an alternative unsupervised algorithm (GMM), and its own strengths and limitations.

In summary, the authors should at a minimum revise the manuscript to acknowledge the curse of dimensionality and the limitations of k-means clustering. Reworking the entire analysis to reduce dimensionality and implement GMM clustering, while desirable, is likely beyond the scope of this work.
Agreed, noting that the transparency of the algorithm that we use is a reasonable fit to the scope of the study.

Specific Comments

[Line 62] I would be careful with your use of "high-dimensional" here. As you correctly identify, *k*-means clustering relies on Euclidean distances; however, Euclidean distance becomes a less useful metric for clustering since, with increasing dimensionality, data become sparse and distance less meaningful (the curse of dimensionality). It thus becomes increasingly difficult to distinguish clusters, since all the distances between points appear the same. An informative exploration of this phenomenon is given by Aggarwal et al. (2001, http://link.springer.com/10.1007/3-540-44503-X_27).
We'll re-write this paragraph, removing the phrase from the first sentence, and adding the suggested caution (which we agree with) later in the paragraph. We'll also add the reason for our choice of algorithm and note one alternative (GMM) as suggested above.

[Line 96] I would suggest citing the paper (Jenkins et al. 2021), not the AGU abstract.
Thank-you for picking this up. This reference will be updated.

[Line 102] You cite another AGU abstract by Sawi et al. - if they have a related paper, you should cite it instead.
This reference will be updated to the full paper also.

[Lines 170-176] I appreciate the thorough explanation, but feel it can be said more concisely.
We're happy to edit accordingly.

[Line 175] "...the features demonstrate comparable distributions." Do they? Perhaps you mean to say they are distributed over comparable scales? Because the distributions themselves are quite unique: some are normal, some bimodal, etc.
We're happy to clarify as suggested.

[Line 178-179] Let me first say that I am pleased to see this careful treatment of your input features, acknowledging that not all features are useful for the clustering analysis. However, I think you should explain your reasoning further or at least provide a citation to a useful reference, for the benefit of your readers. What type of bias are you referring to? Is it distinct from the bias you strive to eliminate by standardizing your features? Why is the inclusion of too many features bad?
We will clarify these lines, in the context of the paragraph before them where we define bias in a slightly different use case. We will remove the word *bias* in Line 178 as it is misleading.

[Sec. 3.2] Doesn't k-means as implemented by scikit-learn use random restarts, as well? If so, you should include this in your description and discuss how the centroids are determined accordingly.
We will add a sentence on this point. The clusters may organise in a different order, but we can identify matching clusters in this study through some (manually) identified event groups.

[Lines 207-208] The silhouette score is tricky to describe, and I'm afraid your explanation leaves me confused. Particularly confusing is the phrase, "each cluster set of means." Please clarify the explanation.
We're happy to provide a more condensed explanation.

[Sec. 3.1, Sec. 3.2, Fig. 3] You index clusters by *r*, but also use "r" as the correlation coefficient. I would suggest de-conflicting your notation, e.g., indexing clusters as *k*=1,...,*K* and reserving *r* for the correlation coefficient. Furthermore, I would suggest italicizing the correlation coefficient so it is abundantly clear you are referring to a parameter. In the caption of Fig. 3, I initially interpreted "(r)" as "right-hand."
We will keep the k as used because of its prominence in k-means literature, but we agree to change the r notation as it conflicts with correlation coefficient. We will plan on substituting the integer value of r to g in the k-mean steps detailed in Sect. 3.2. Other options conflict with variables widely used in seismology, but we defer to the editor if there is a better suggestion.

[Line 262 & Suppl. Line 46] Why is *k*=14 a "realistic" maximum?
This is a reasonable number of distinct event types, for the complex seismic wavefield of the WIS. For a valley glacier (for example), without the adjacent ice shelf, we anticipate that a smaller k value would be more appropriate.

[Sec 3.3.2, Sec. S3.3, Fig. 5, Fig. S6] I applaud your effort to thoroughly examine the "evolution" of clusters. However, as currently written, the first paragraph of Sec 3.3.2 leaves me with a sense of confusion at best, or at worst that the analysis may be flawed. How are you able to track the progeny of *k* clusters from from the previous *k*-1 clusters? You mention that k-means has built-in pseudo-randomness (and random restarts), but between Section S3.1 and this section, I fail to understand how (and why) you overcame this. Are you setting the seeds manually, and preventing restarts? Though you refer to Section S3.3 in a previous section (Line 209), I would refer to it again here, as there are crucial details in it that you should relate to this part of the text.
We agree this point needs clarifying and are happy to do so. In our study, we can use event groups to (manually) identify matching clusters (typically they are not in the same order, but have recognisable content) but other studies should consider controlling the re-starts as you mention.

The last thing I'll say about this section & related supplement/figures is that I had to re-read them several times to become convinced that you are not proposing that the composition of clusters at *k*+1 is dependent on the composition at *k*. The discussion in paragraph 2 (Line 261) ultimately settled the question, but I think if you clean up the first paragraph, you can alleviate the confusion. With the question settled, Fig. 5 is a nice analysis of how cluster composition changes.
We are happy to clarify as suggested.

[Suppl. Line 64] This is not *a priori*; it is *a posteriori* since you must run experiments to determine the optimal number of clusters.
We'll delete that word.

[Suppl. Line 66] "that is, ..." This is self-evident and redundant.
We'll delete that half-sentence.

[Suppl. Lines 65-71, Fig. S3b] What justifies ignoring the outliers?
We will clarify the sentence at line 66 as it is confusingly worded. We will also remove the last two sentences as the quantitative information is displayed in Fig. S3.

[Fig. 7] The top row is an interesting analysis, but I'm not convinced of the value of the bottom row. To me, it seems you are comparing apples to oranges. Because the datasets are essentially different due to the inclusion/exclusion of additional features, there are too many factors that affect the cluster assignment of data, including pseudo-random seeding, restarts, and, even if those are controlled, the vastly different dimensionality of the parameter space. All this is to say, clusters in one dimensionality do not look like clusters in another dimensionality, as indicated by the top row of subplots.
We agree in general (noting previous clarifications above), and are happy to qualify the text that refers to the bottom row. We do think that it is useful to include to illustrate the variability of results given choices made in the workflow.

Technical Corrections

[Line 101] A space is missing between "fracture" and the citation.
Space to be added.

[Line 182] Just use "variance."
Wording to be changed.

[Fig. 2] This is a rather trivial comment, but your 4th column is missing x-axis tick labels. My preference (certainly not a prescription) for this type of figure is to have the same y-scale for all plots, and print the axis labels on just one subplot.
Tick labels to be added. We prefer to retain the annotations on all columns, as the tick labels are very simple and give a visual confirmation that a consistence scale is used on the x axis.

[Suppl. Line 58] Change to "where *n*=100 is the number of bins...", etc.
Wording to be changed.