“Estimating the number of clusters using cross-validation”
Summary of changes to the revised manuscript

Wei Fu Patrick O. Perry

January 10, 2019

We thank the editors and reviewers for their time reading our submission, and for providing many helpful suggestions for improvement. We also thank you for the opportunity to revise and resubmit our manuscript. We have revised our manuscript carefully taking into account your comments.

Detailed responses to reviewer comments follow below.

Response to Editor

- *Fix the figure captions.*

  Response: Per your instruction, all figure captions now include (1) what is the plot about, (2) specific details of plot, like what type of display and how variables are mapped, (3) the most important thing that the reader should learn.

  TODO

- *Cite R packages.*

  Response: We added citations for all packages used in our work: TODO

- *Reprodicibility.*

  Response: TODO

Response to Associate Editor

- *Compare to the slope method*

  TODO

Response to Reviewer 1

(No further comments.)

Response to Reviewer 2

- *Bottom of page 4. The authors need to clarify the notation*

  Response: TODO

- *Page 6. Clarify that $n \gg K$, $m \gg L$.*

  Response: TODO

- *Section 2.2. Clarify.*

  Response: TODO

- *Section 3. Note that slope is self consistent.*

  Response: TODO

- *Proposition 1 improvements.*

  Response: TODO

  We chose to keep the proof in the main text, but if there are space constraints we will take your suggestion and move the proof to an appendix.

- *P*roposition 2. Define the indicator function.

  Response: Done.

- *Page 18, line 14. Generalization conditions.*

  Response:

  We clarified what we meant by saying our proof generalized: *In principle the proof of Proposition 3 generalizes to non-normal location families of distributions easily, provided that both cluster distributions are in the same family and expressions for $E(Y|Y > 0)$ and the other related quantities are available (and finite).*

  You rightly point out that it will not generalize to all possible situations, but only to distributions with finite moments and cases where both clusters have the same distribution.

- *Figure 2. Please define correctly the x,y axis*

  Response: Done.

- *Page 20, first paragraph.*

  Response: Your comment was that “Proposition 2 only shows that if $|\rho| < 0.5$ the method works correctly.”Did you mean page 14 instead? We think you may have meant that our statement in the first paragraph of page 14, “Proposition 2 tells us that Gabriel cross-validation fails when there is strong correlation between the variables” is too strong.

- *Page 22, the standardized Euclidean norm k-means could produce good results.*

  Response: You may be correct, but we do not want to overwhelm the reader with two many variations of the same algorithm. The correlation-correction we propose is a slightly more general version of Euclidean standardization, and seems to perform adequately.

- *Compare to slope*

  Response: As you suggested, we added comparisons to the slope method, using the `cstab` package. In our simulations, slope performs very well with noise dimensions and high-dimensional data; its performance degrades with correlated clusters, variance heterogeneity, and heavy-tailed data. Certainly there are some situations where slope is better than our method, but no method uniformly dominates the others. In our simulations, our correlation-corrected Gabriel CV method seems to be a good default choice.