# "Estimating the number of clusters using cross-validation"
# Summary of changes to the revised manuscript

Wei Fu Patrick O. Perry

January 11, 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:** Thank you for the guidance here. Per your instruction, all figure captions now include (1) what is the plot about, (2) specific details of plot, and (3) the most important thing that the reader should learn.

- *Cite R packages.*

  **Response:**

  We added citations for the main packages used in our work, including for the R software environment itself. The packages cited include the following: `cluster`, `cstab`, `e1071` `fpc`, `ggplot2`, `MASS`, `mclust`, and `NbClust`.

- *Reprodicibility.*

  **Response:** TODO

## Response to Associate Editor

- *Compare to the slope method*

  As requested, we added a comparison to the slope method. Slope performs better than our method in Setting 2 (and better than the non-correlation corrected version in Setting 1), but otherwise our method either performs comparably or performs better.

## Response to Reviewer 1

(No further comments.)

## Response to Reviewer 2

- *Bottom of page 4. Clarify meaning of "fold".*

  **Response:** Added the following:

  "For each value $r$ in the range $1, \ldots, K$, we define the $r$th 'fold' as the matrix gotten by permuting the rows of the original data matrix to get $X$, a matrix with the $r$th test subset, the rows $i$ with $r_i = r$ as its trailing rows."

- *Page 6. Clarify that $N \geq K$, $M \geq K$.*

  **Response:**

  Added the sentence: "For this cross-validation to be possible, we must have that $K \leq N$ and $L \leq M$."

- *Section 2.2. Clarify the context in which the method is applied.*

  **Response:**

  Changed the first sentence as follows: "Our version of Gabriel cross-validation for clustering works is designed to find the best number of clusters to use for $N$ observations of a $P$-dimensional random variable with real-valued components."

- *Section 3. Note that other methods are self consistent.*

  **Response:** Per your suggestion, we added this paragraph:

  "Self-consistency by itself does not mean that a procedure performs well in practice. Self-consistency deals only with the no-noise situation. Many other methods for selecting $K$ are likely self-consistent, but not all perform the same in practice. Self-consistency only suggests (but does not guarantee) that a procedure might perform reasonably when the noise level is low."

  We did not specifically single out the slope statistic or Silhouette method specifically here because we have not taken the time to analyze those methods and we are unaware of proofs of their self-consistency or lack thereof.

- *Proposition 1 improvements.*

  **Response:** We fixed the typo you spotted ($CV(k) > CV(K)$), and we added a note that the proposition is purely for theoretical purposes: "The assumptions here are not checkable in practice, but the proposition suggest that Gabriel cross-validation might give a reasonable answer, at least when the noise level is low."

  You also suggested that we omit the proposition entirely. We chose to keep it, 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.