Response to Reviewers

1 Overview

We thank the two reviewers and the Associate Editor for their close reading of the paper and their many suggestions and edits. We have taken these suggesions into account in our revision, and we believe that our paper is greatly improved.

See below for detailed point-by-point responses.

2 Response to Associate Editor

1. What is the exact difference between the proposed method and Owen and Perry (2009) and Gabriel (2002)? I think the three methods are related, and the differences should be stated clearly so to make the contribution of the current paper clear.

We added the following text at the end of Sec 2.1:

> Gabriel's cross validation is analogous to leave-one-out cross-validation, while Owen and Perry's generalization handles more general leave-outs.
>
> Gabriel's cross-validation and the Owen and Perry generalization are designed specifically for choosing the rank of a low-rank matrix approximation like the singular value decomposition or the non-negative matrix factorization. While for some purposes it makes sense to view k-means clustering as a sort of low rank approximation to the data matrix, the discrete nature of clustering (each observation gets assigned to one of k possible clusters) imposes additional structure on the problem that makes the prior procedures not directly applicable.
>
> In the sequel, we extend the version of Gabriel cross-validation described by Owen and Perry to the problem of selecting the number of clusters, k. To do so, we introduce a classification procedure into the cross-validation steps that takes the place of least squares regression in the matrix factorization problem formation. Our new procedure is able to select the number of clusters, k, automatically. For this procedure, we provide theoretical and empirical support analogous to the consistency results proved by Owen and Perry (2009).

2. In numerical comparisons, default parameters settings were used for competing methods. Is this fair?

We added the following to Sec. 6.1:

For the competing methods, some might take issue with the fact that we are using default parameter settings. In many other benchmark comparisons, one would choose the tuning parameters using some form of cross-validation. In our case, this is not an option. We are in an *unsupervised* setting, so there is no clear way to set the tuning parameters. We must rely on the defaults from the implementations of these methods.

3. The correlation correction. How is this done in the high-dimensional setting... It's well known that the sample covariance matrix is unstable.

We added the following to end of Sec. 5:

In many applications, including our application in the sequel to a yeast cell cycle dataset, the number of observations N is below the number of features P. When this happens, the sample covariance matrix $\hat{\Sigma}$ will not have full rank P, and will be a poor estimate of the population covariance matrix. In these situations, though, it will still be the case that a decomposition $\hat{\Sigma} = \Gamma\Lambda\Gamma^T$ exists with Γ having orthonormal columns and Λ being diagonal and positive-definite. It will also be the case that $\Lambda^{-1/2}$ and hence $\tilde{X}$ exist. What fails in these situation is that, since $\hat{\Sigma}$ is no longer a good estimate of Σ, it will no longer be the case that our correction reliably transforms the noise covariance to I. Still, in these situations, applying the correlation correction may give better results than using the original untransformed data. In practice we inspect the resulting estimate $\hat{k}$ both with and without the correlation correction, and if the two answers agree, we can be more confident in our answer.

3 Response to Reviewer 1

Thank you for providing many corrections to our original manuscript. We have mostly made the changes you suggested, except that we preferred keeping the description as "Gabriel cross-validation" rather than "Gabriel's cross-validation".

I missed a comparison with the slope statistics published by Fujita et al.

Regarding your question about why we did not compare to Fujita et al.'s method, we must emphasize that there are literally hundreds of methods for choosing the number of clusters, and it is impossible to compare against them all. The main reason we did not compare against the Fujita et. al method is that there is no publicly-available implementation. Beyond that, the method is not in common usage relative to the others we surveyed (28 Google scholar citations versus, for example, 2992 citations for the Gap statistic). Finally, the Fujita method, being based on inter-point distance, is roughly in the same category as four other methods we surveyed: Gap, Jump, CH, and Hartigan's method, so we do not expect to get radically different results from those four methods.

4 Response to Reviewer 2

4.1 Theoretical justification

(1) In both Propositions 2 and 3, normality assumption is assumed. Can it be relaxed?

1. We added the following comment after the proof of Proposition 2, which describes one way to relax the normality assumption:

   > The proof of Proposition 2 relies on normality assumptions in two crucial ways: (i) assuming that X and Y have symmetric marginal distributions; and (ii) assuming that $E(Y|X) = \rho X$. We could relax the normality assumption by instead supposing that X and Y can be decomposed as $X = Z_1$ and $Y = \rho Z_1 + \sqrt{1-\rho^2} Z_2$ where Z_1 and Z_2 are independent random variables with mean zero, unit variance and symmetric distribution. Any bivariate distribution can be put in this form for uncorrelated Z_1 and Z_2, but only special distributions, including the bivariate normal, can be put in the form with independent Z_1 and Z_2.

2. We also added the following comment after the proof of Proposition 3:

   > In principle the proof of Proposition 3 generalizes to non-normal distributions easily, provided expressions for $E(Y|Y > 0)$ and the other related quantities are available. Different distributions for (X, Y) give rise to different cutoffs for when $k = 1$ cluster is preferred to $k = 2$. The normal distribution is only special here in that the required conditional expectations and the boundary where $CV(1) = CV(2)$ are computable in closed form.

(2) In Proposition 3, only 2 dimensional case is considered. Can it be generalized to multi-dimension as in Proposition 4?

We generalized the result as you requested, and added the following comment after the proof of Proposition 3:

> We have stated and proved Proposition 3 for data in $d = 2$ dimensions, but there is nothing essential about the particulars of the dimension; we generalize the result to arbitrary dimension in Appendex C.

(3) In Proposition 4, one assumption is posed on the first eigenvalue and eigenvector of $\Sigma_Y Y$ for any split, and thus it is difficult to verify. It would be more convincing to assume a verifiable condition on the distribution of (X, Y).

We added a more intuitive condition after the proof of Proposition 4.

The condition that ensures CV(1) < CV(2) is tight but difficult to understand intuitively. We make understand this condition better by noting that it is equivalent to a condition on the correlation between the two linear combinations of X and Y variables:

$$\text{cor}(u_1^T \Sigma_{YX} X, u_1^T Y) < 1/2. \quad (1)$$

A stronger condition that is easier to understand, then, is that the canonical correlation between X and Y is below 1/2; when the latter is true the former follows and CV(1) < CV(2). A necessary but not sufficient guarantee of the stronger condition is that all correlations between variables are below 1/2.

One might object to the condition (1) being stated in terms of a particular (X, Y) split of the data. In practice, one might want the condition to hold for all possible splits of the data. A sufficient guarantee that (1) holds for any (X, Y) split of the data is that the data covariance matrix Σ is diagonal. The ad-hoc adjustment we propose in Sec. 5 attempts to transform the data to this form.

The main point of this theorem is to prove analytically that our method has difficulty in the face of correlated noise. This motivates us to propose the correlation correction in section 5. We then demonstrate empirically through the simulation studies and real data examples that are correlation correction is sufficient to correct many of the types correlation found in practice.

4.2 2. Correlation correction

The adjustment for correlation is based on the assumption of a shared covariance matrix. Yet this assumption is insufficient.

As you note, $\hat{k}_0$ tends to overestimate k when correlation is high, and the covariance matrices for the $\hat{k}_0$ clusters is not the same as the true underlying covariance matrices for the k clusters. We never claim, however, to estimate Σ consistently. We only claim that our adjustment is an ad-hoc procedure that performs well in both our simulations and our real-data examples.

We expaned the discussion at the end of Section 5 to emphasize this point:

Our correlation correction is not backed by a rigorous theoretical justification. Further, since our initial estimate $\hat{k}_0$ tends to over-estimate the true number of clusters in the presence of strong correlation, our covariance estimate $\hat{\Sigma}$ will often be biased. Fortunately, we do not need to estimate Σ with high accuracy. For our correction to be useful, we only need $\tilde{X}$ to have lower within-cluster correlation than the raw data X. In our simulations and in our empirical validation study in Sections 6 and 7 we demonstrate that despite its theoretical shortcoming, our ad-hoc correlation adjustment often performs well in practice.

4.3 3. Numerical experiments

(1) I suspect that the generated clusters are severely overlapped.

Thank you for catching this mistake. Our previous description was incorrect. We changed the description in 6.1 to the following:

> In all simulation settings, we randomly generate cluster centers by drawing from a multivariate normal distribution with covariance matrix τI, conditional on the cluster points being well-separated. By "well-separated" we mean that the distance from any point to its own cluster center is less than the distance of that point to any other cluster center by a margin of at least 1.0 units. In any replicate where the clusters are not well-separated, we discard the replicate and generate another set of points. We choose τ to make the probability the cluster centers being well-separated on the first draw to be equal to approximately 50%. The "well-separated" condition and some of our simulation settings are chosen to mimic those used by Tibshirani et al. (2001) in settings (c) and (d) of their Section 6.

The last example on t-distribution is similar to normal distribution

For large degrees of freedom the two distributions are similar, but when the degrees of freedom approach 2 as in our simulation, the distribution becomes heavy-tailed, with infinite variance. The "well-separated" condition and some of our simulation settings are chosen to mimic those used by Tibshirani et al. (2001) in settings (c) and (d) of their Section 6.

(2) I didn't find any examples the same as in Tibshirani

Added the following in 6.1

> The "well-separated" condition and some o our simulation settings are chosen to mimic those used by Tibshirani et al. (2001) in settings (c) and (d) of their Section 6.

(3) I suggest the authors replace the figures by tables.

This is difficult. Not only do we have 5 simulation setups, for each setup we vary a parameter from 10 settings. This requires us to report 50 tables. We have added a supplementary appendix with the tables, but we have left the figures as-is in the main text of the article.

4.4 4. Some typos

(1) "cross validation" should be "cross-validation"

Fixed.

(2) "of of" should be "of"

Fixed.

(3) Line 1, page 25 misses a period

Fixed.

(4) In Proposition 4, Σ_{XY} is used for two different meanings.

Fixed.

(5) Most figures do not have captions.

Fixed.