We thank all the reviewers for their very helpful comments and suggestions. Please find below our responses.

Literature review

We will rewrite the existing paragraph on Projection Pursuit (PP) and discuss PP-based clustering methods that have direct relevance. In particular, we will review criteria for cluster identification including kurtosis (Peña and Prieto, 2001 and follow-up works), first absolute moment and skewness (Verzelen and Arias-Castro, 2017) and relate them to CURE. We will also discuss algorithms proposed in those papers. In addition, we will review more recent clustering methods for non-spherical data, such as the one (Kushnir et al., 2019) mentioned by Reviewer 4 based on random projections.

Generalization to multi-class nonlinear clustering

While our theories focus on the two-component elliptical mixture model, the idea of CURE generalizes to multi-class scenarios, allowing for nonlinear discriminant functions (i.e. feature mappings). We will revise the brief and somewhat abstract discussion in Section 2.3 and provide more intuitions to the general audience. As is pointed out by Reviewer 2, we will elaborate why the linear CURE approach is a special case. Due to space constraints, we will defer the most of these details to the supplementary.

We will add a new numerical example (to the supplementary) with the first 4 classes in Fashion-MNIST (T-shirt/top, Trouser, Pullover, Dress). All details will be included in the final version. In short, we use Wasserstein-1 distance as the discrepancy measure D for uncoupled regression and compare two classes of feature mappings: linear functions and fully-connected neural networks with one hidden layer that has 100 nodes. The learning curves in Figure 1 shows the advantage of neural network and demonstrates the flexibility of CURE with nonlinear function classes.

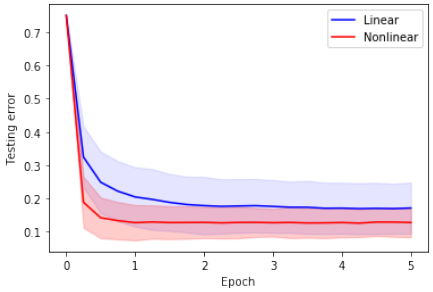

Figure 1: 4-class Fashion-MNIST.

Achieving Bayes optimality in multi-class clustering is indeed very challenging. Under parametric models (e.g. Gaussian mixtures), one may construct suitable loss functions for CURE based on likelihood functions. We will comment on this in the discussion section.

Numerical experiments

A main motivation for this paper is to deal with stretched (elongated) clusters where directions with the largest variabilities of data may not be informative for clustering at all. Instead, one should aim for directions onto which the projected data exhibit cluster structures. This explains why we choose T-shirts/tops and Pullovers in the Fashion-MNIST dataset for demonstration: the bulk of a image corresponds to the belly part of clothing with different grayscales, logos and hence contributes to the most of variability. However, T-shirts and Pullovers are distinguished by sleeves. Hence the two classes can be separated by a linear function that is not related to the leading principle component of data.

We conducted new experiments comparing CURE with more algorithms: [1] Model-based clustering (Mclust) in Fraley and Raftery (1999); [2] Projection Pursuit (PP) in Peña and Prieto (2001); [3] alternations between linear discriminant analysis and K-means (LDA + Kmeans) in Ding and Li (2007); [4] Minimum Density Hyperplane (MDH) in Pavlidis et al. (2016). [1], [2] and [4] are implemented using open-source R packages with default settings. While there is no implementation of [3] available publicly, we did it ourselves following the instructions in the paper. Table 1 shows the results. For randomized algorithms we do 50 independen runs and report means and standard deviations of misclassification rates. Again, CURE outperforms all the competitors. On a Macbook Pro it takes less than 10 seconds to converge while others usually require one minute or more. Currently, we are also exploring other datasets.

Table 1: Misclassification rates of CURE and other methods.

Method \ $N_1 : N_2$	$1:1$	$2:1$	$3:1$	$4:1$
CURE (ours)	$5.2 \pm 0.2\%$	$7.1 \pm 0.4\%$	$9.3 \pm 0.7\%$	$11.3 \pm 1.1\%$
Mclust [1]	$48.7 \pm 1.3\%$	$39.1 \pm 4.8\%$	$34.1 \pm 8.0\%$	$28.2 \pm 7.8\%$
Projection Pursuit [2]	$36.9 \pm 9.8\%$	$37.4 \pm 9.6\%$	$39.7 \pm 6.9\%$	$40.6 \pm 7.3\%$
LDA + Kmeans [3]	45.9%	49.0%	45.6%	44.3%
MDH [4]	48.6%	43.1%	38.3%	35.2%

It is worth pointing out that CURE learns a classification rule that readily predicts labels for any new data. This is an advantage over many existing approaches for clustering and embedding, including spectral methods and t-SNE where out-of-sample extensions are not so straightforward. We will highlight this in the paper.

Other issues. For the balanced two-class problem we have an explicit construction of f and set the regularization parameter λ to be 1. Our theory goes through as long as $\lambda \geq 1$, and our experimental results are not sensitive to the choice of λ. So we choose $\lambda = 1$ to reduce the need of tuning and simplify statements of theoretical results.