Testing and Tuning SkinnyDip: Noise-Robust Clustering

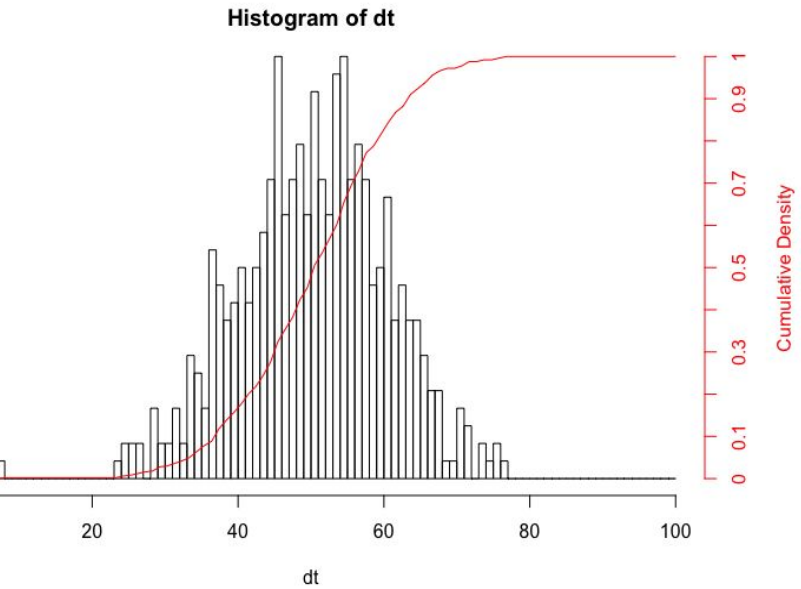
Presenter: Grant King

SkinnyDip

- Non-parametric, deterministic clustering
- Robust to very noisy data
- Based on the “dip-test” for modality

Modality and Empirical Cumulative Distribution Function

- ECDF is a monotonically increasing function



- Unimodal: ECDF is concave up to mode and convex after

https://i.stack.imgur.com/czgsR.png

Dip Test

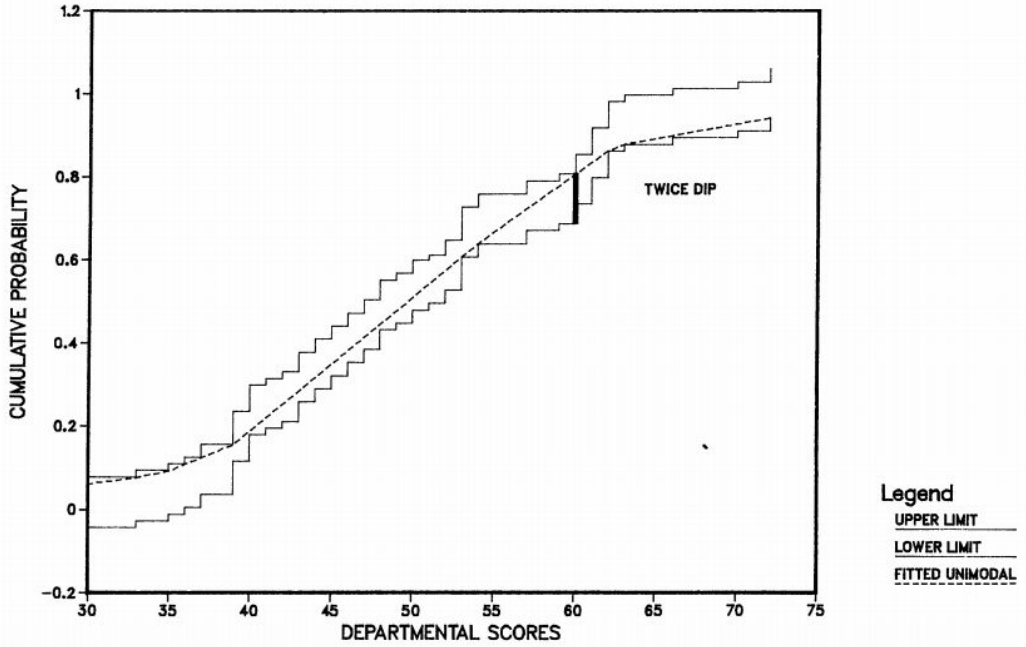
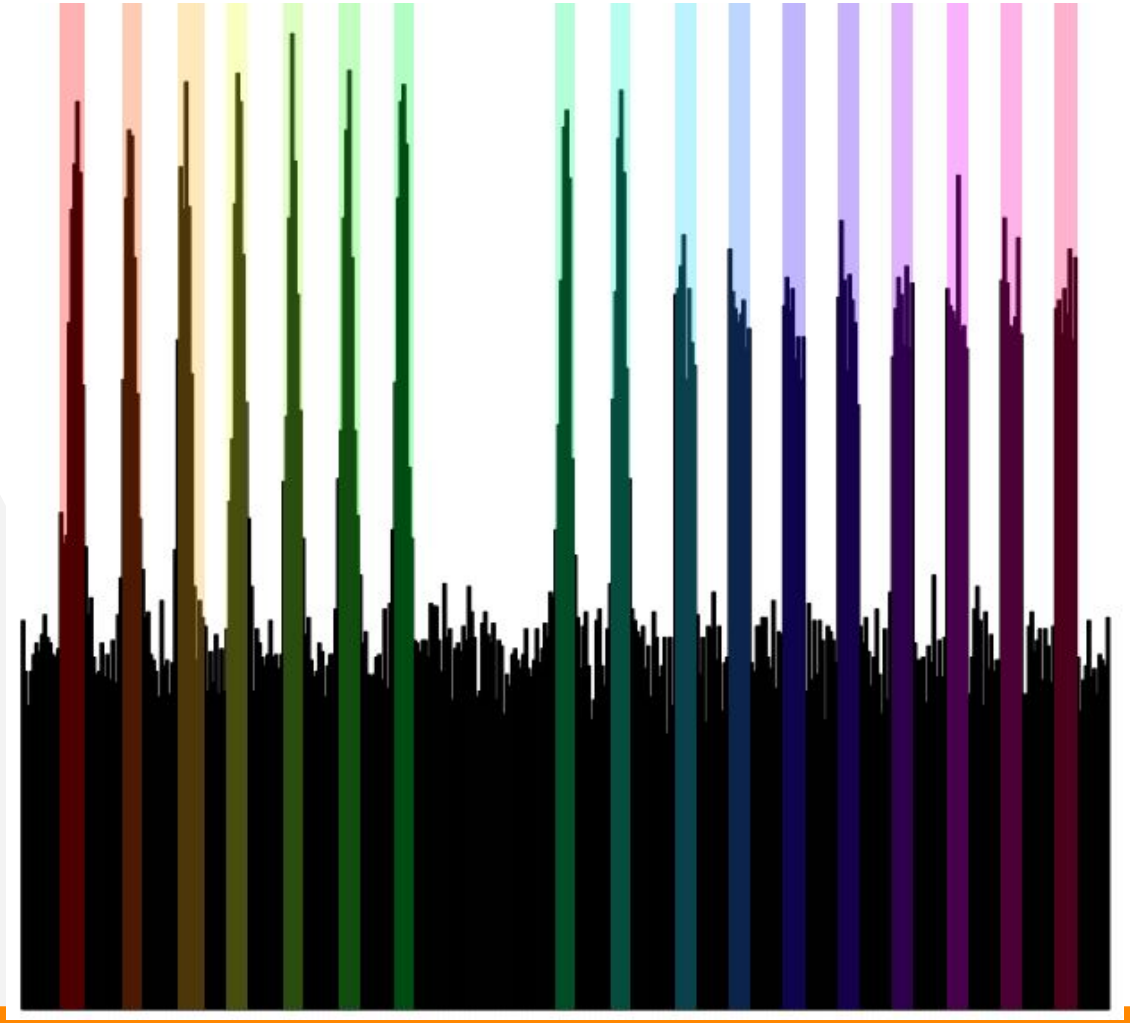
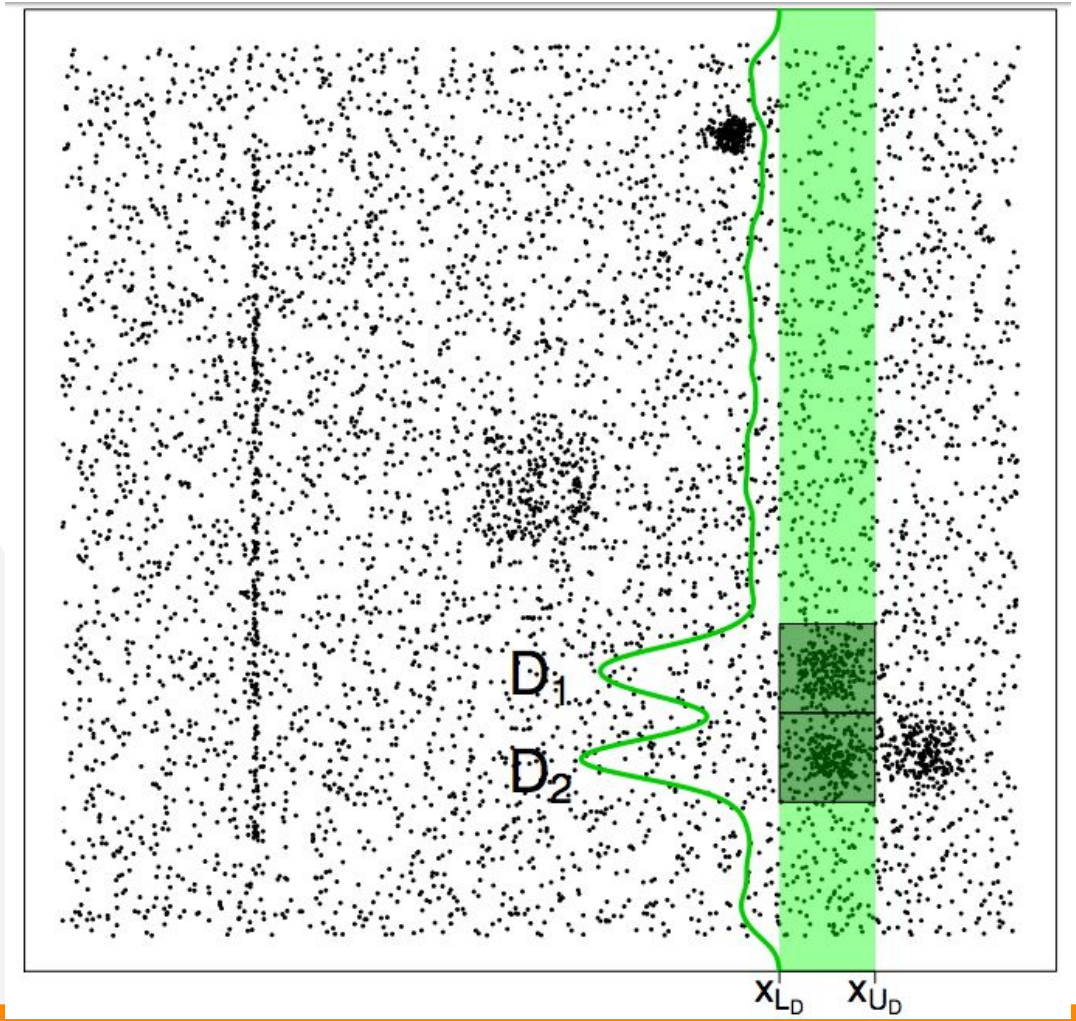


FIG. 1. *Faculty quality in statistics department.*

J. A. Hartigan and P. Hartigan. The dip test of unimodality. The Annals of Statistics, 1985.

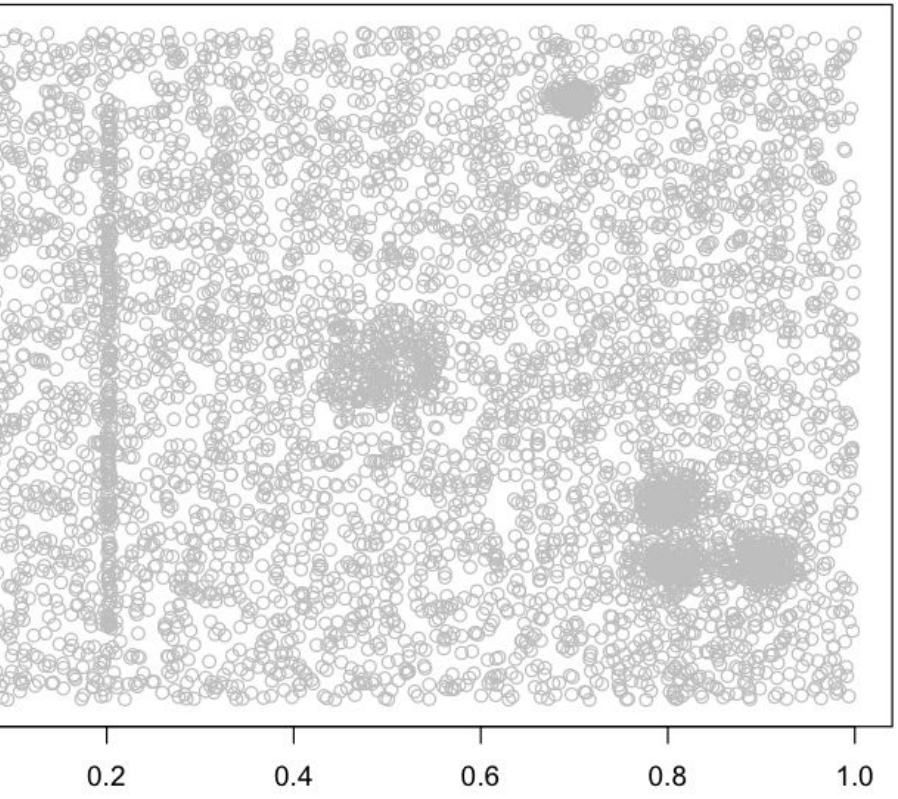



D1
D2
xLD
xUD

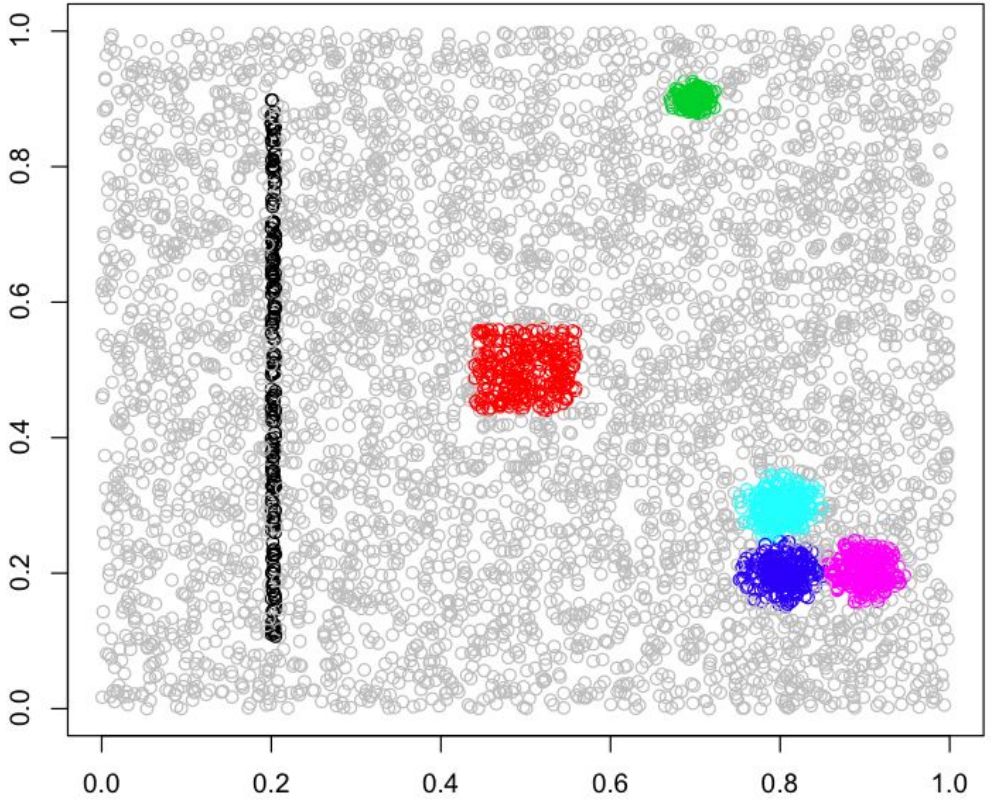
Hypercubic Regions

- SkinnyDip recurses through each dimension of the data and finds approximate unimodal regions
- These intersections of these regions are then combined to form the basis of each cluster
- Accuracy drops exponentially with increasing dimensionality

SkinnyDip Running Example Unlabeled



SkinnyDip Running Example Ground Truth

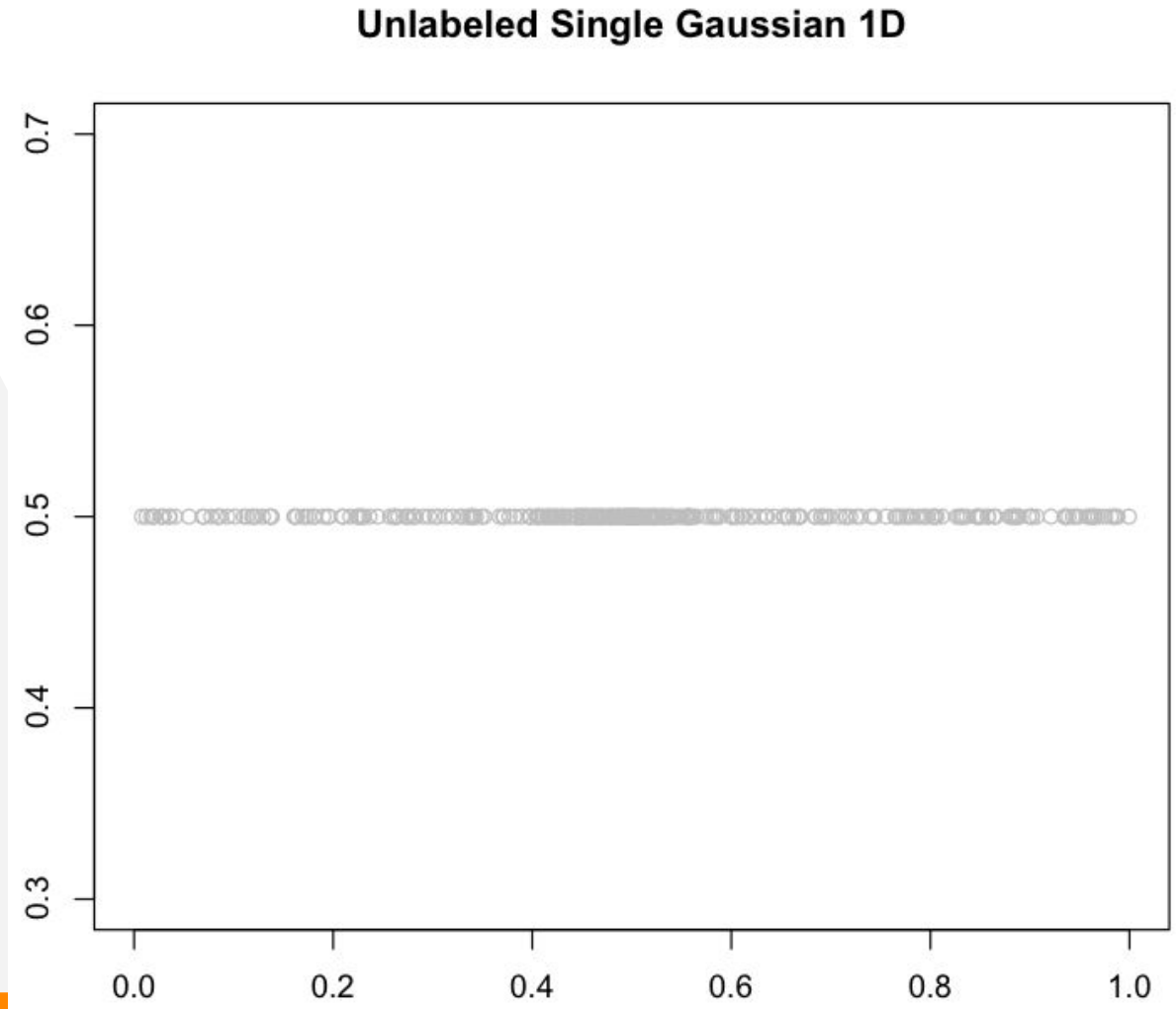


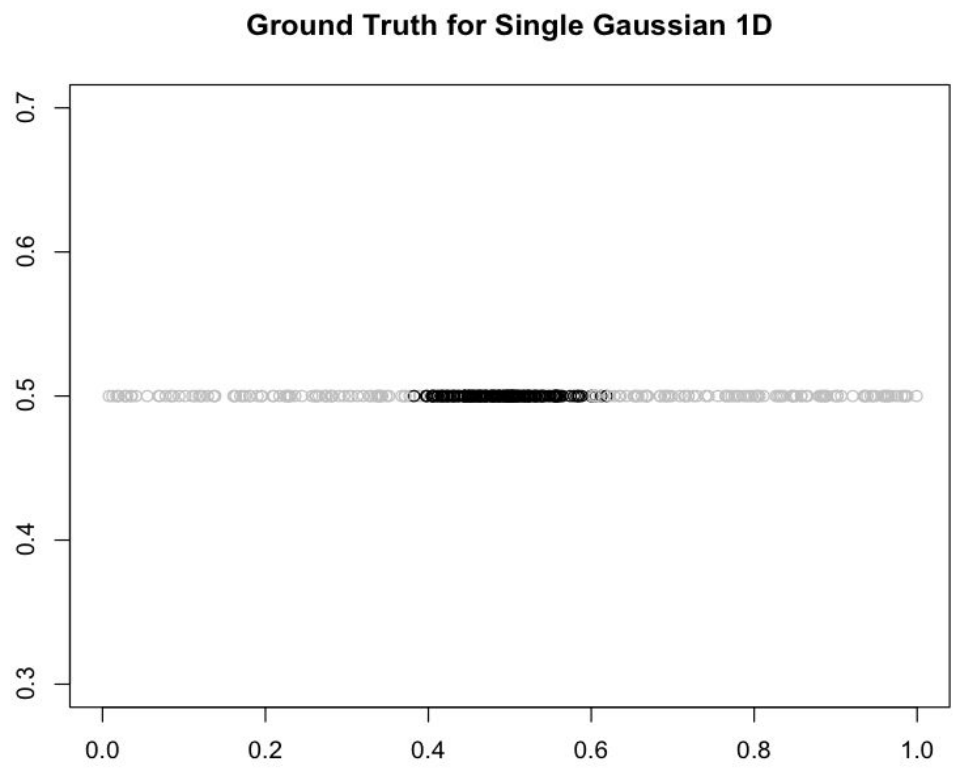
Testing

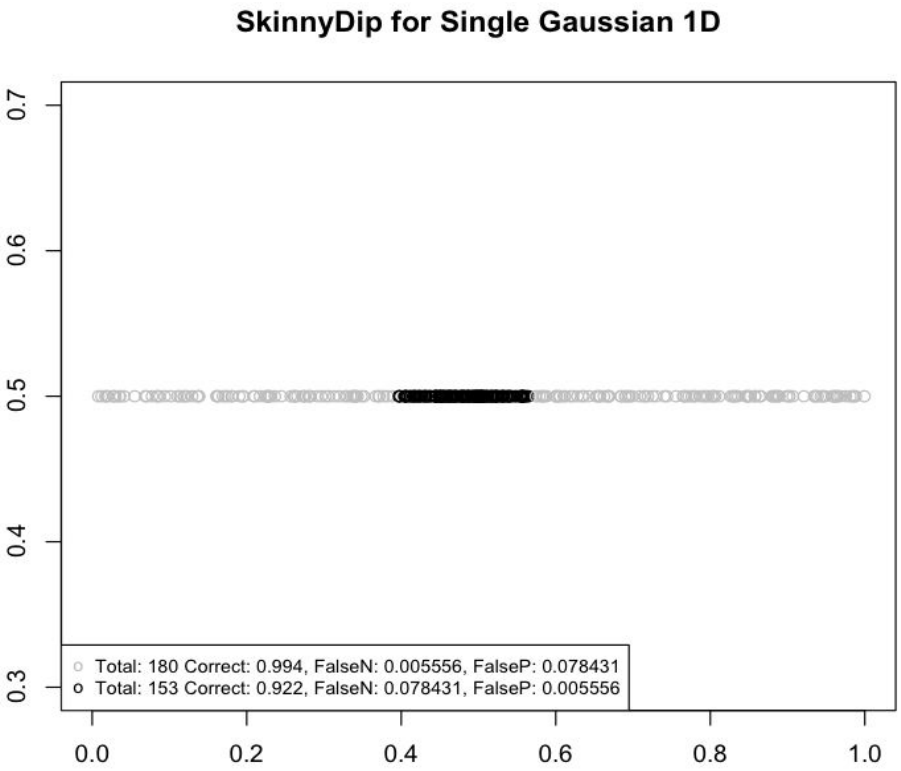
- Map unsupervised clusters to ground truth labels
- Gain a better understanding of inherent error
- Test multiple dimensions and clusters

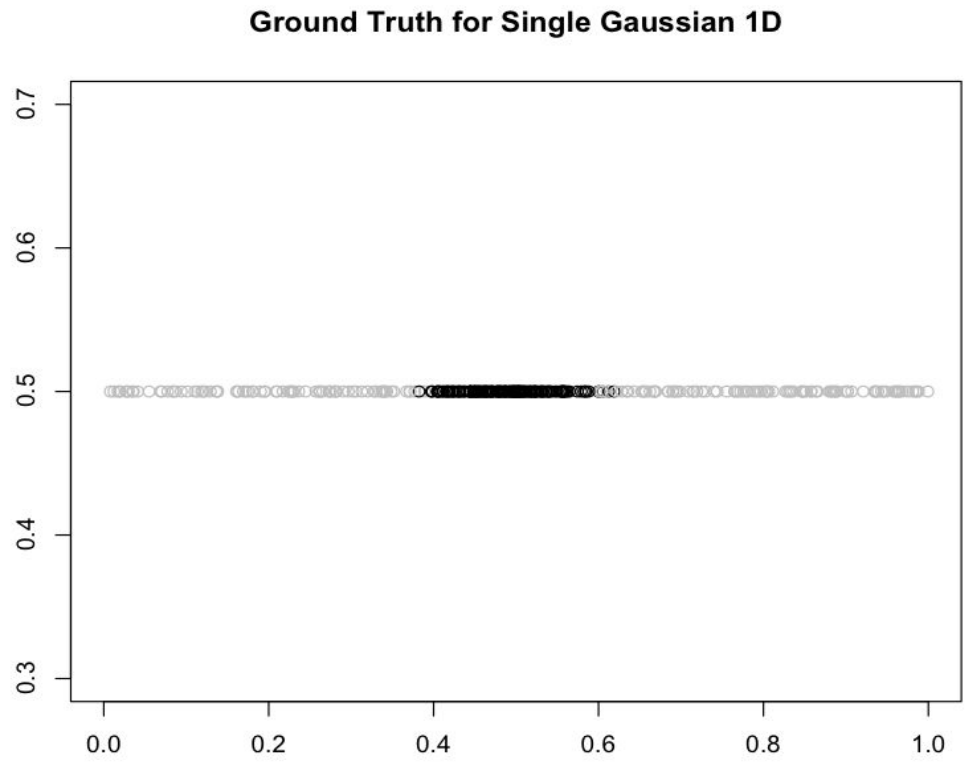
Tuning

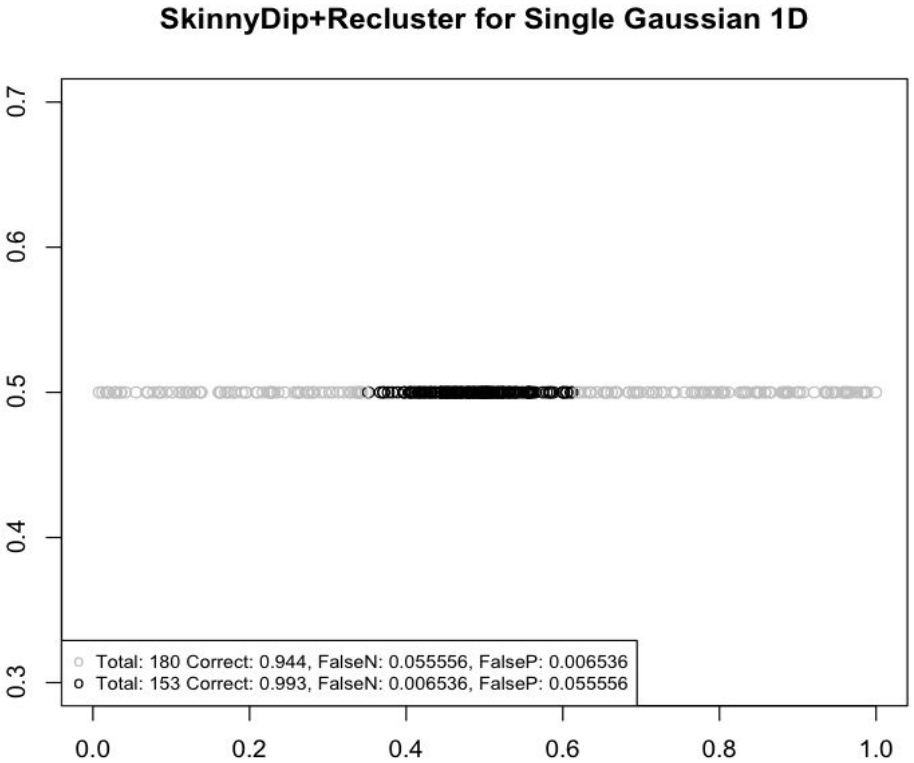
- Use distribution of points in each cluster as evidence for underlying variance
- Remap points in the original data using empirical data
  - e.g. Mahalnobis distance using pseudo-inverse cov
- Retest

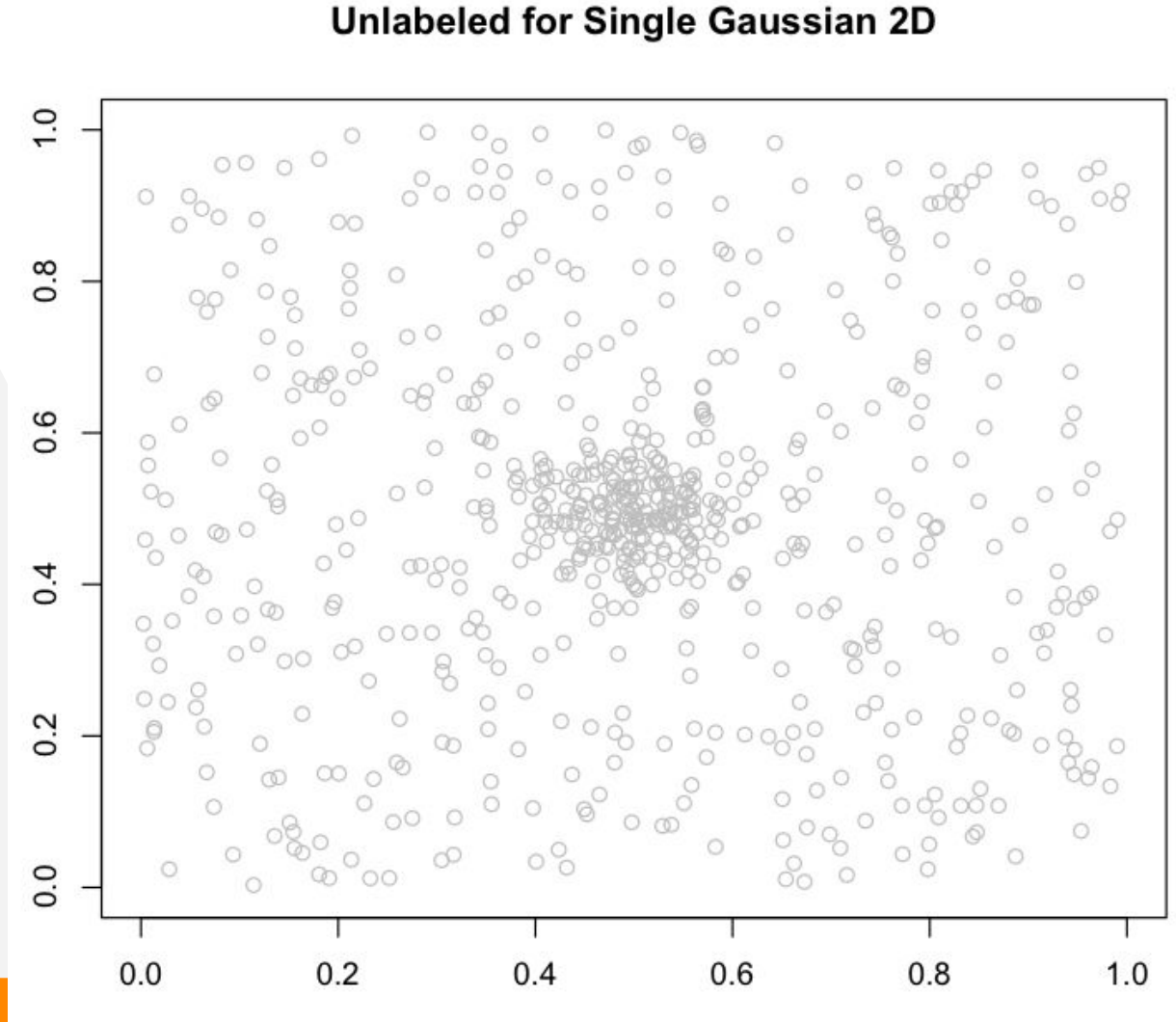

Unlabeled Single Gaussian 1D
0.7
0.6
0.5
0.4
0.3
0.0
0.2
0.4
0.6
0.8
1.0

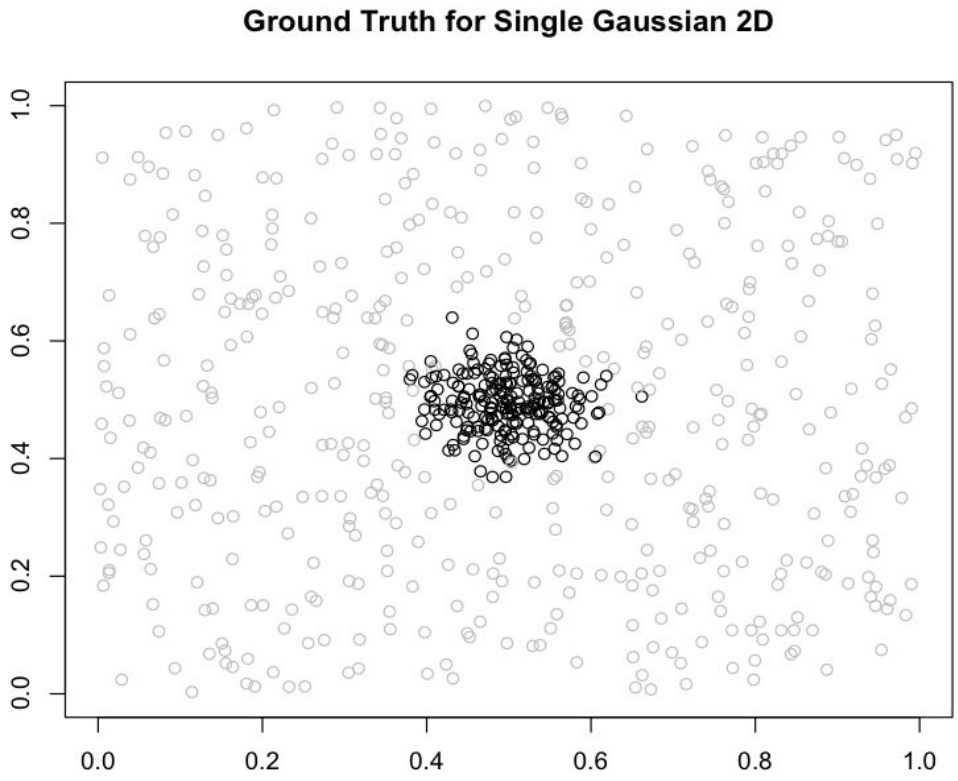

Ground Truth for Single Gaussian 1D
0.7
0.6
0.5
0.4
0.3
0.0
0.2
0.4
0.6
0.8
1.0

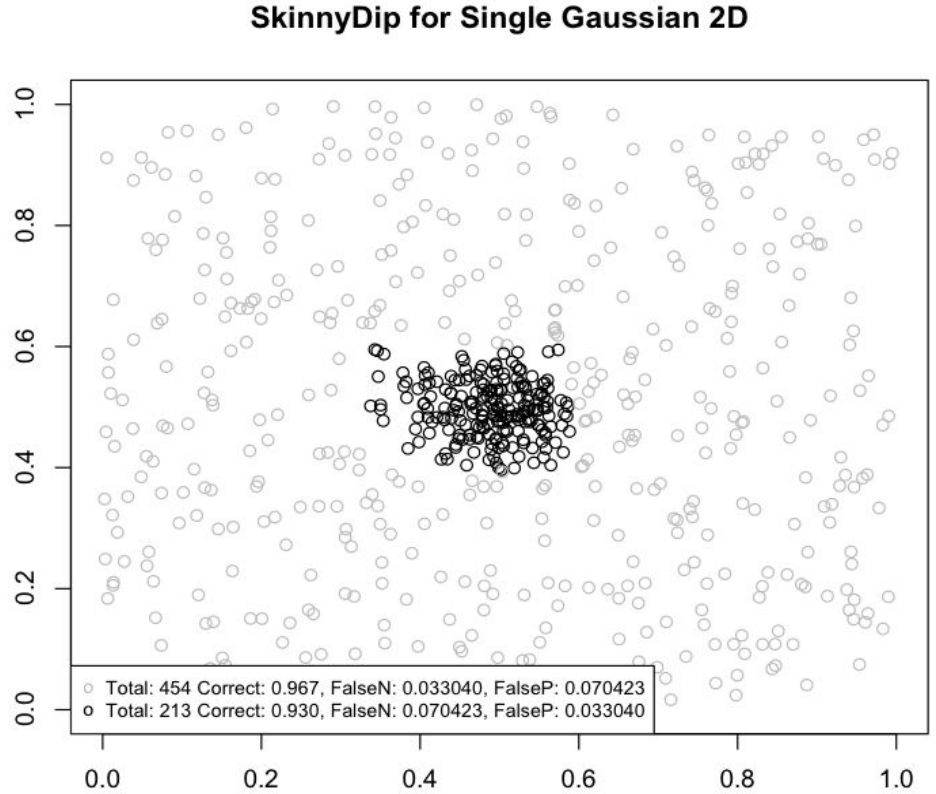

SkinnyDip for Single Gaussian 1D
0.7
0.6
0.5
0.4
0.3
Total: 180 Correct: 0.994, FalseN: 0.005556, FalseP: 0.078431
Total: 153 Correct: 0.922, FalseN: 0.078431, FalseP: 0.005556
0.0
0.2
0.4
0.6
0.8
1.0

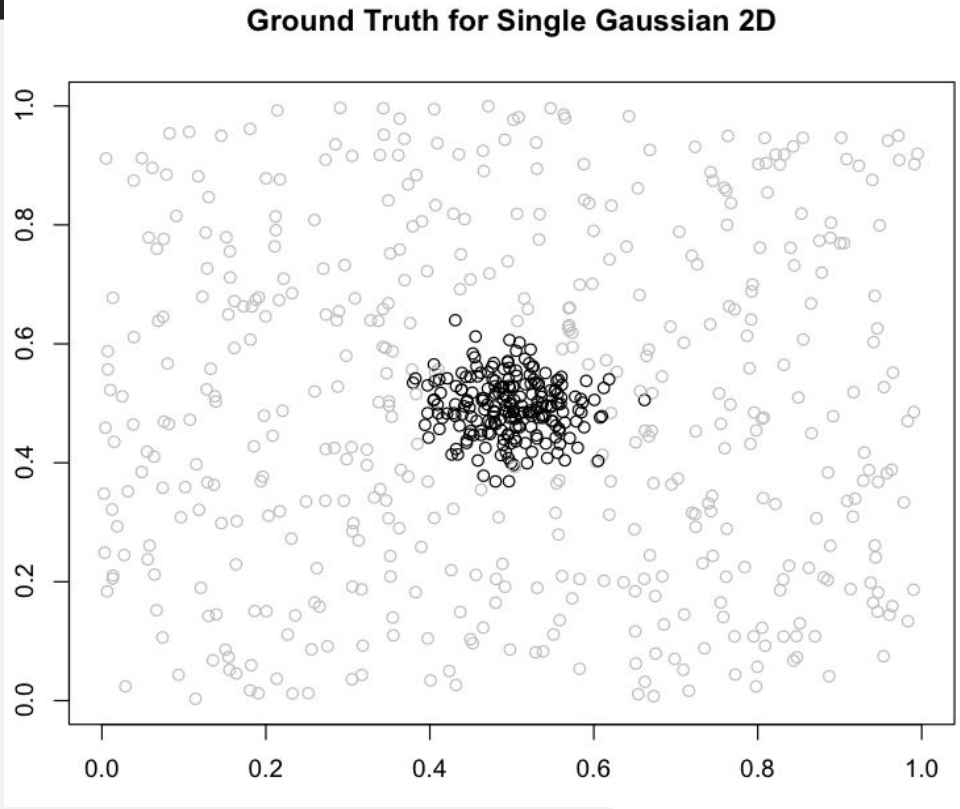

Ground Truth for Single Gaussian 1D
0.7
0.6
0.5
0.4
0.3
0.0
0.2
0.4
0.6
0.8
1.0

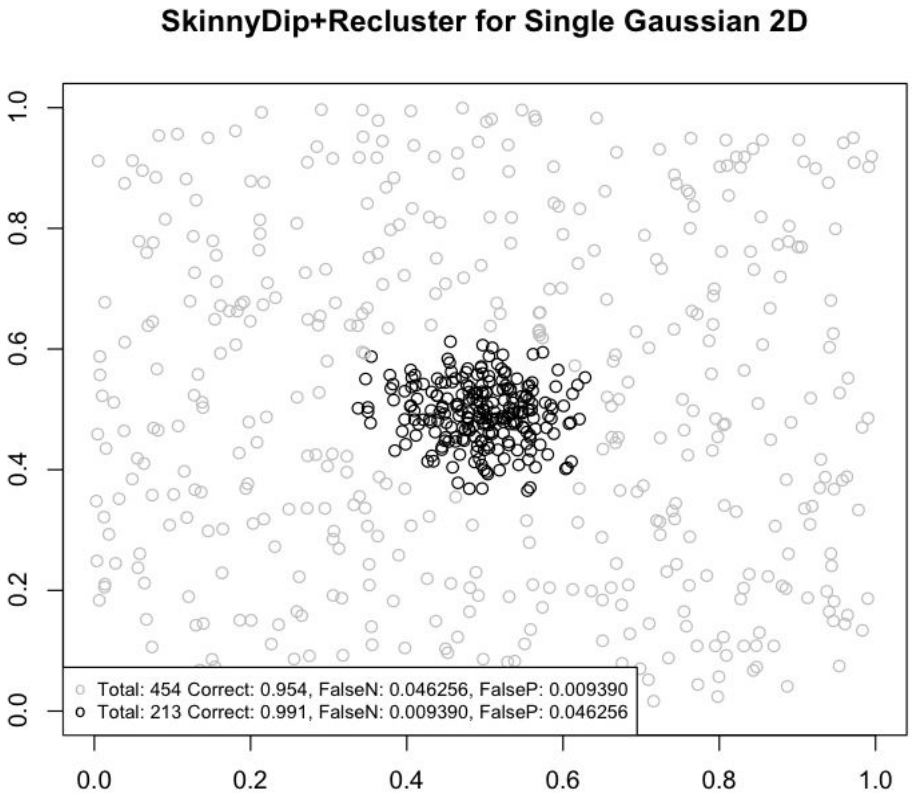

SkinnyDip+Recluster for Single Gaussian 1D
0.7
0.6
0.5
0.4
0.3
Total: 180 Correct: 0.944, FalseN: 0.055556, FalseP: 0.006536
Total: 153 Correct: 0.993, FalseN: 0.006536, FalseP: 0.055556
0.0
0.2
0.4
0.6
0.8
1.0


Unlabeled for Single Gaussian 2D
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0

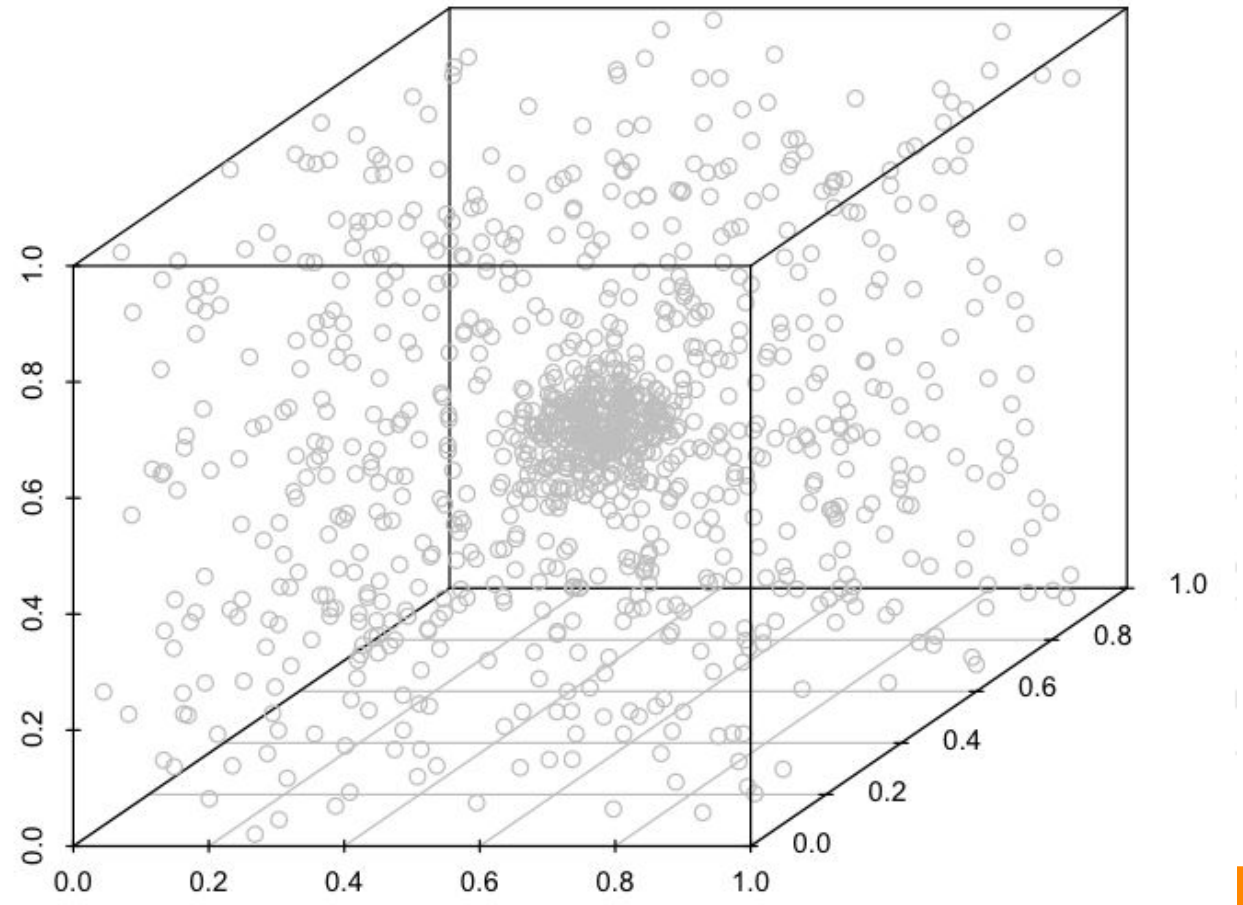

Ground Truth for Single Gaussian 2D
0.0
0.2
0.4
0.6
0.8
1.0


SkinnyDip for Single Gaussian 2D
Total: 454 Correct: 0.967, FalseN: 0.033040, FalseP: 0.070423
Total: 213 Correct: 0.930, FalseN: 0.070423, FalseP: 0.033040
0.0
0.2
0.4
0.6
0.8
1.0

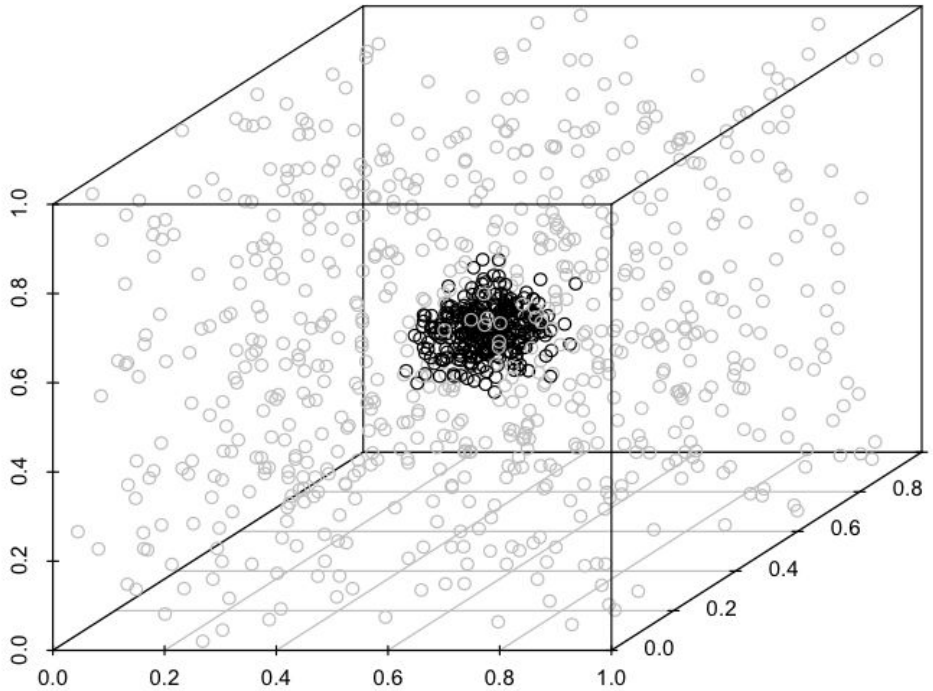

Ground Truth for Single Gaussian 2D
0.0
0.2
0.4
0.6
0.8
1.0


SkinnyDip+Recluster for Single Gaussian 2D
Total: 454 Correct: 0.954, FalseN: 0.046256, FalseP: 0.009390
Total: 213 Correct: 0.991, FalseN: 0.009390, FalseP: 0.046256
0.0
0.2
0.4
0.6
0.8
1.0

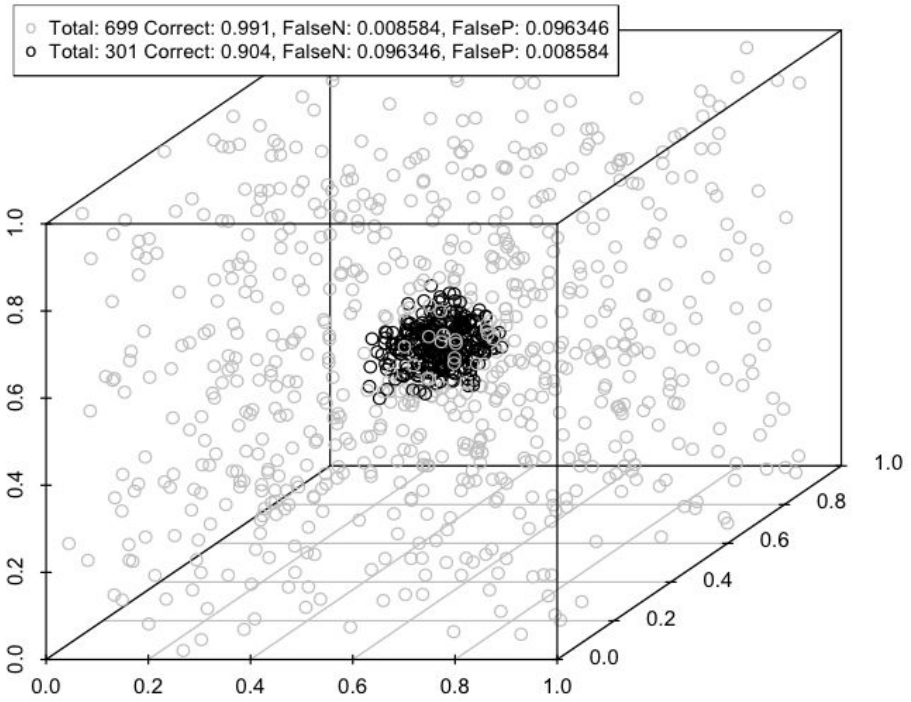
Unlabeled Single Gaussian 3D

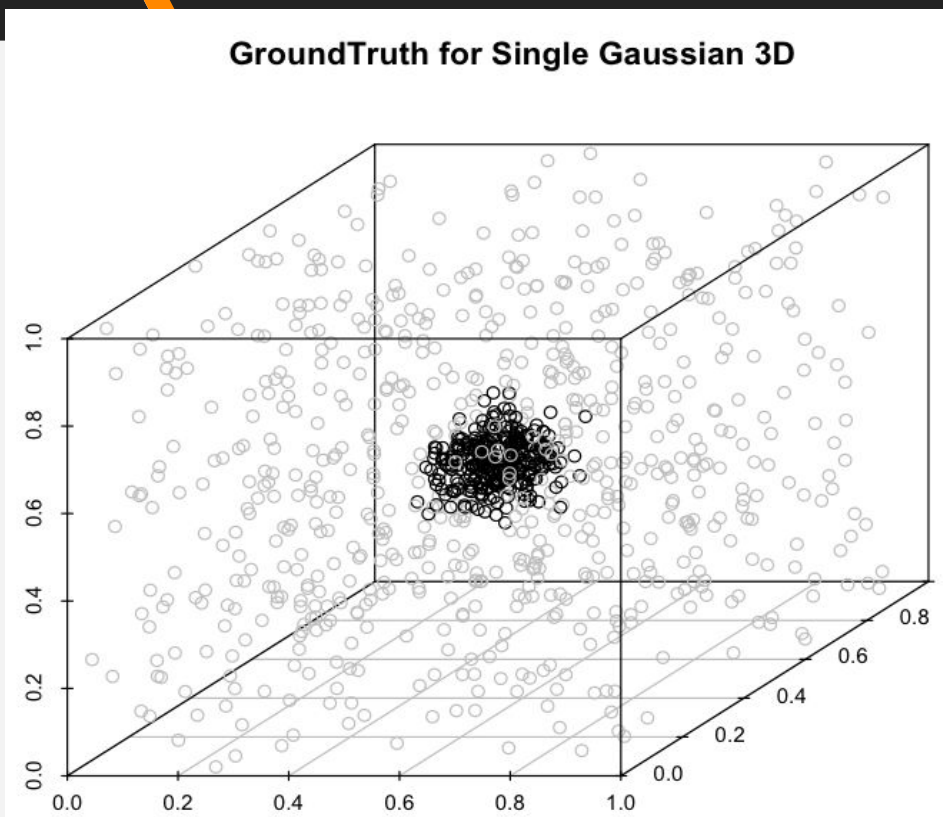


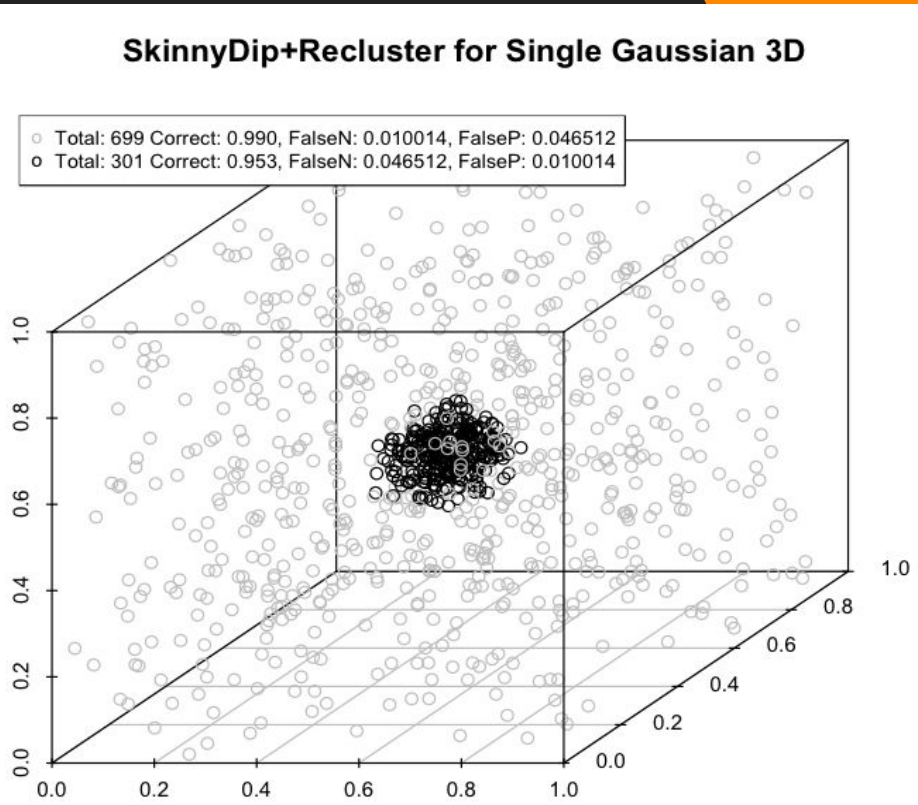
GroundTruth for Single Gaussian 3D



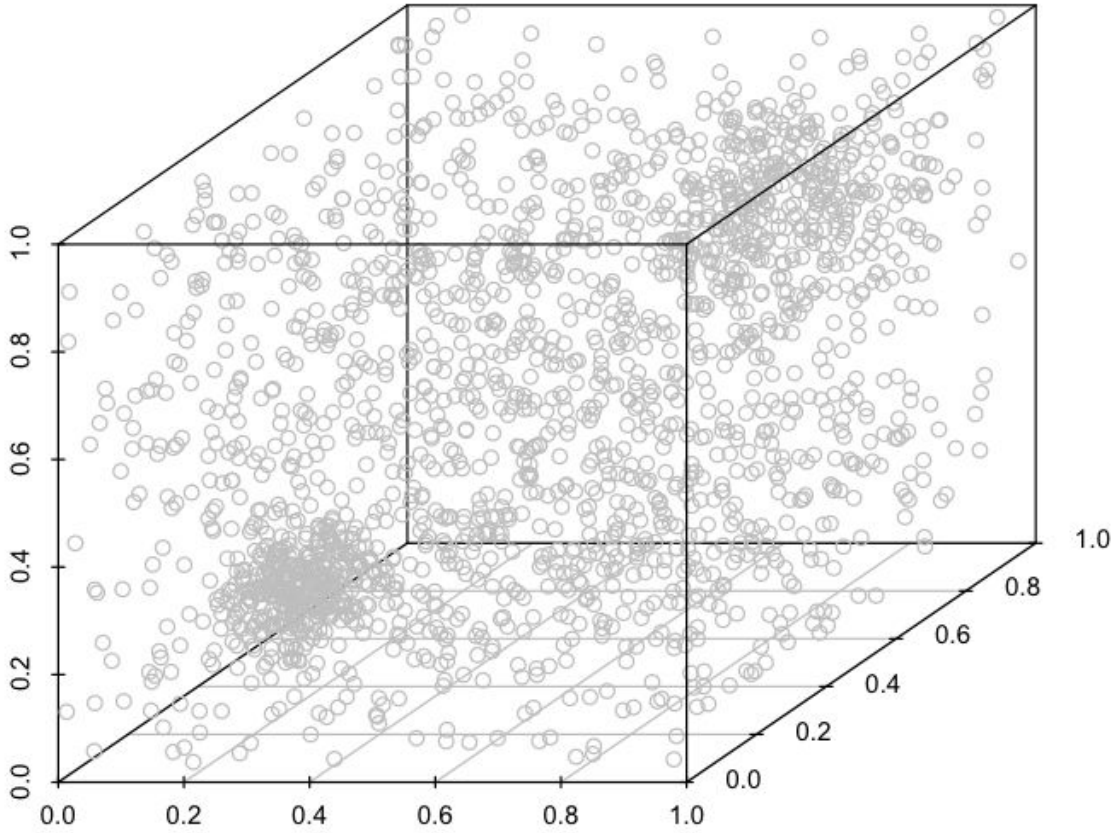
SkinnyDip for Single Gaussian 3D



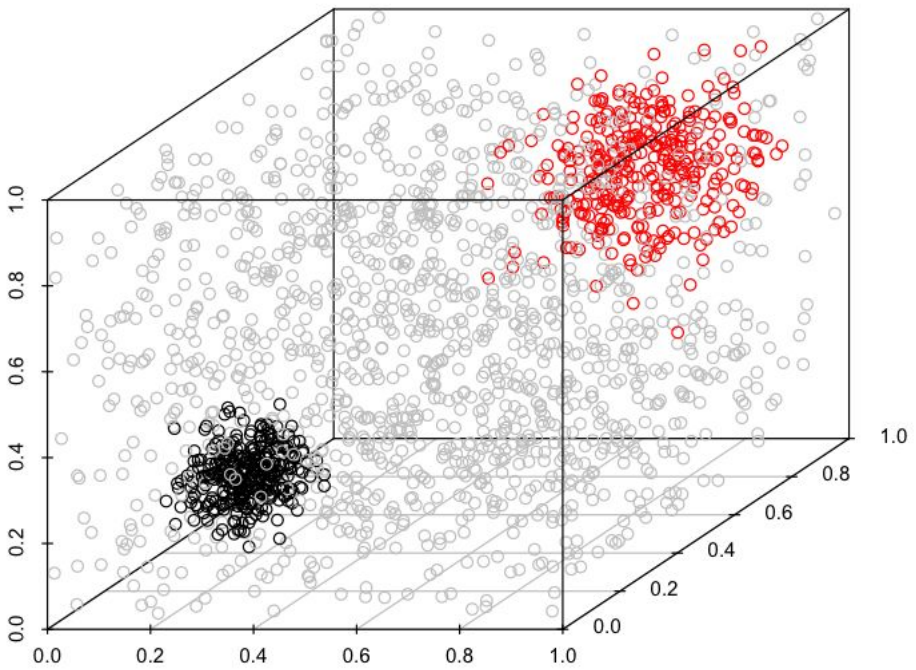

GroundTruth for Single Gaussian 3D
0.0
0.2
0.4
0.6
0.8
1.0


SkinnyDip+Recluster for Single Gaussian 3D
Total: 699 Correct: 0.990, FalseN: 0.010014, FalseP: 0.046512
Total: 301 Correct: 0.953, FalseN: 0.046512, FalseP: 0.010014
0.0
0.2
0.4
0.6
0.8
1.0

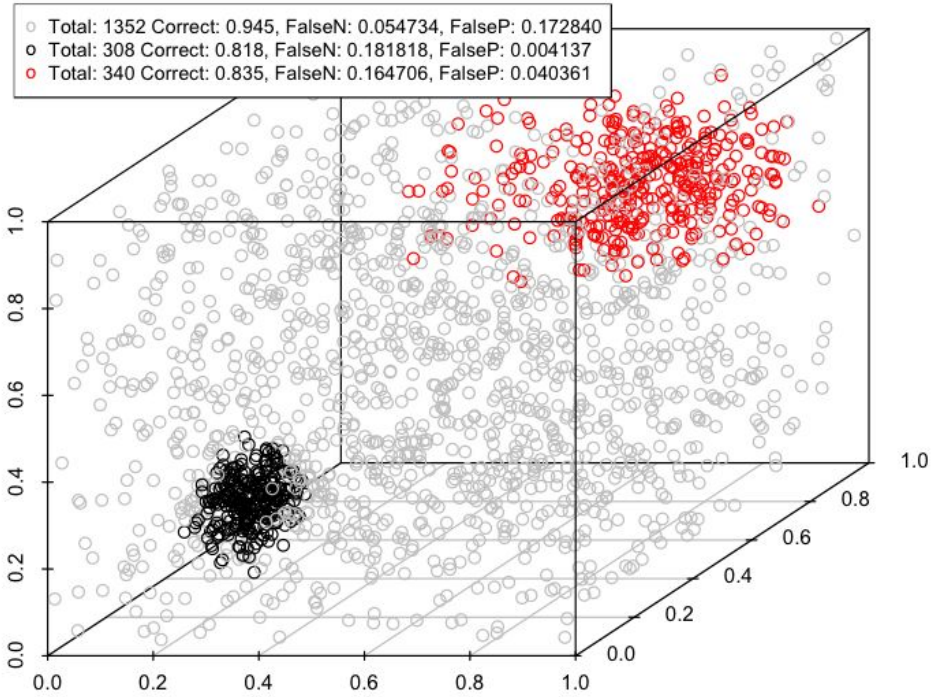
Random Data Unlabeled

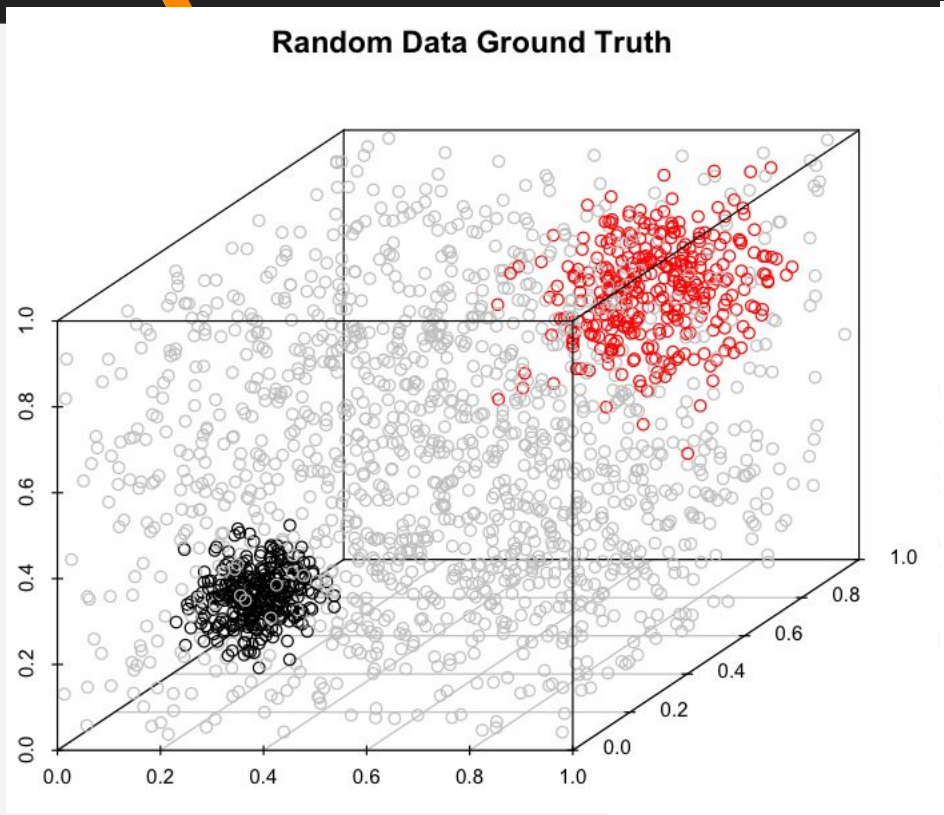


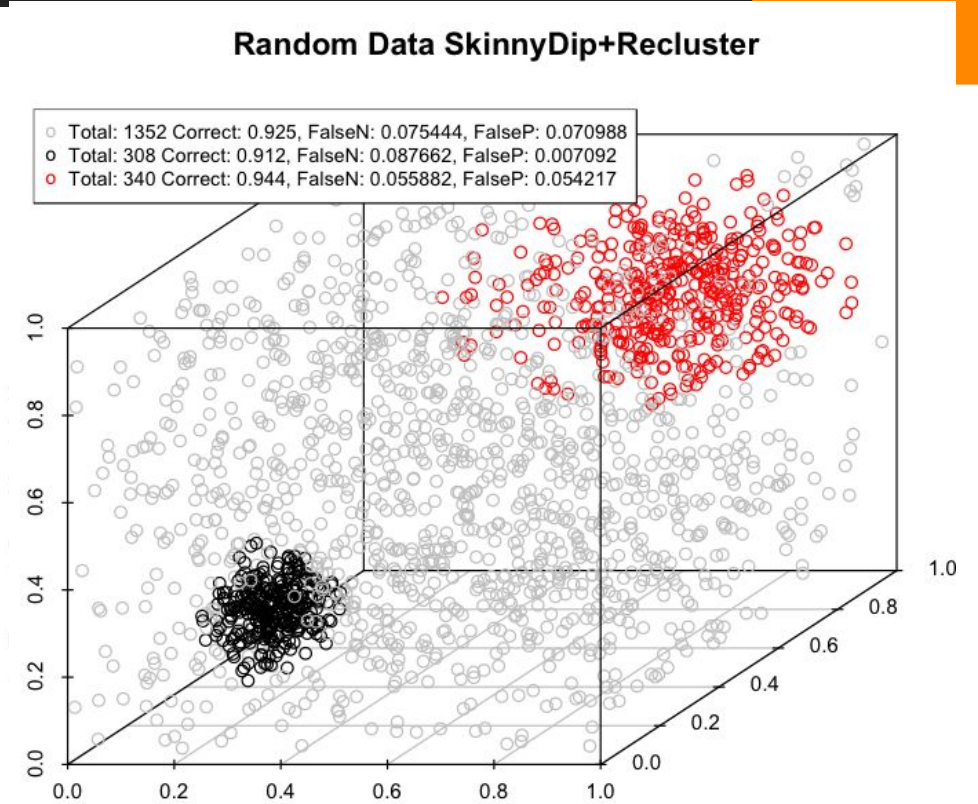
Random Data Ground Truth



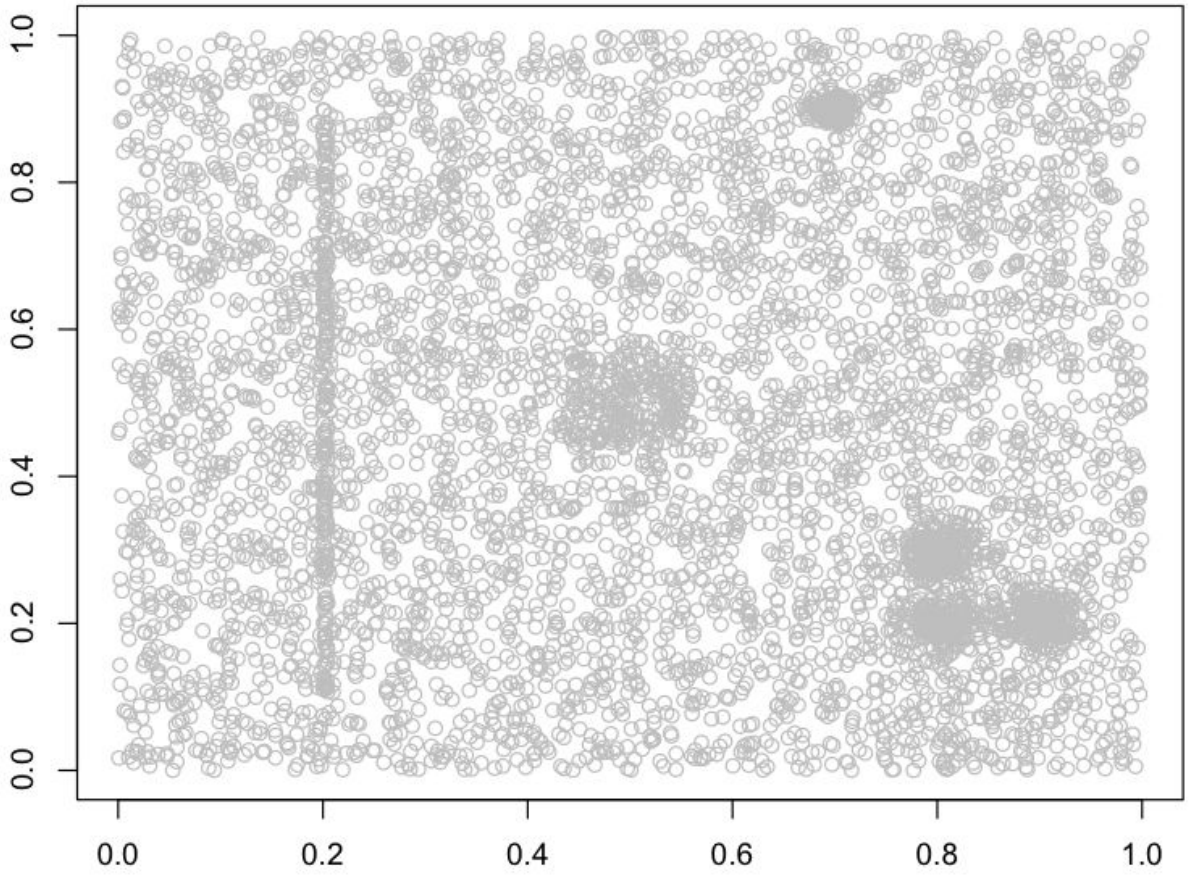
Random Data SkinnyDip

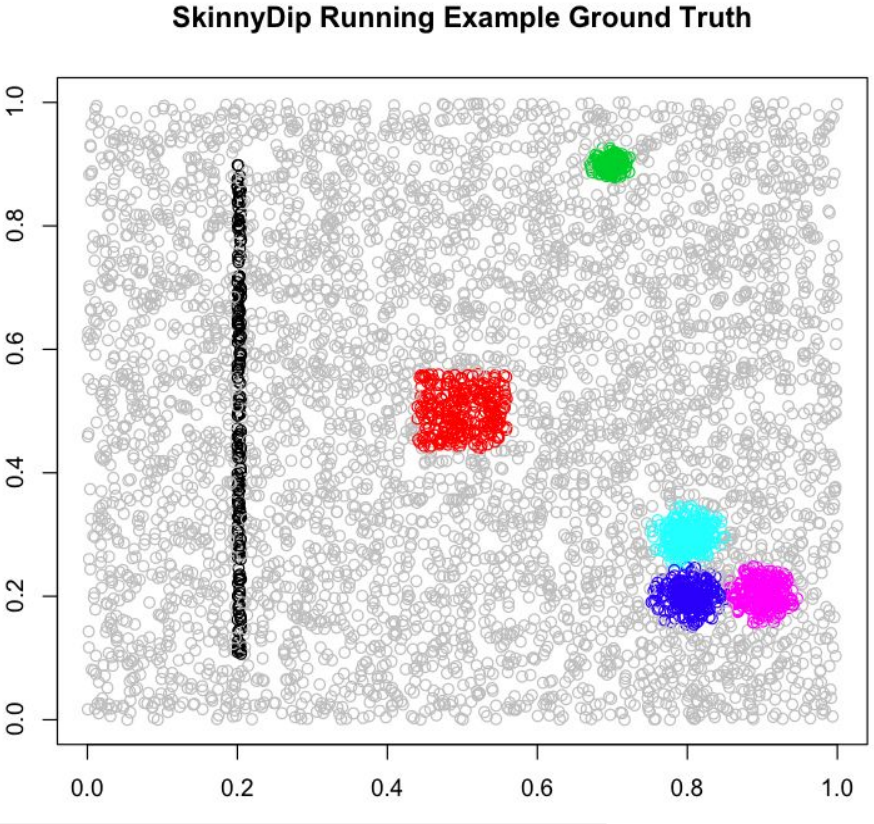


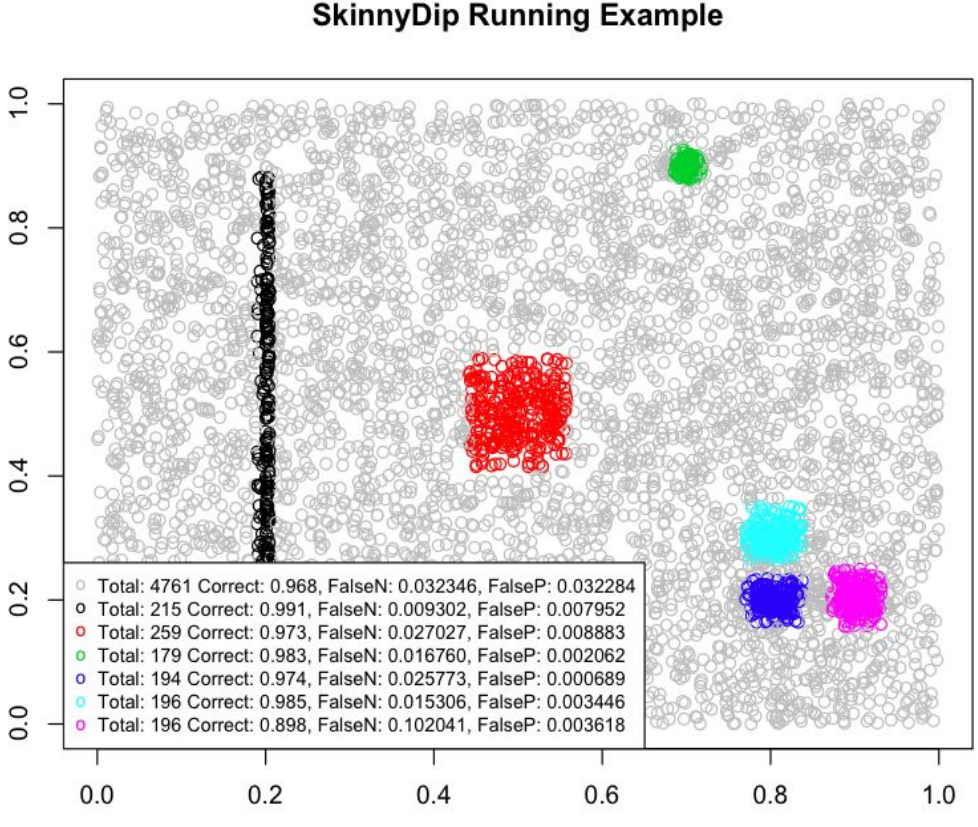

Random Data Ground Truth
0.0
0.2
0.4
0.6
0.8
1.0

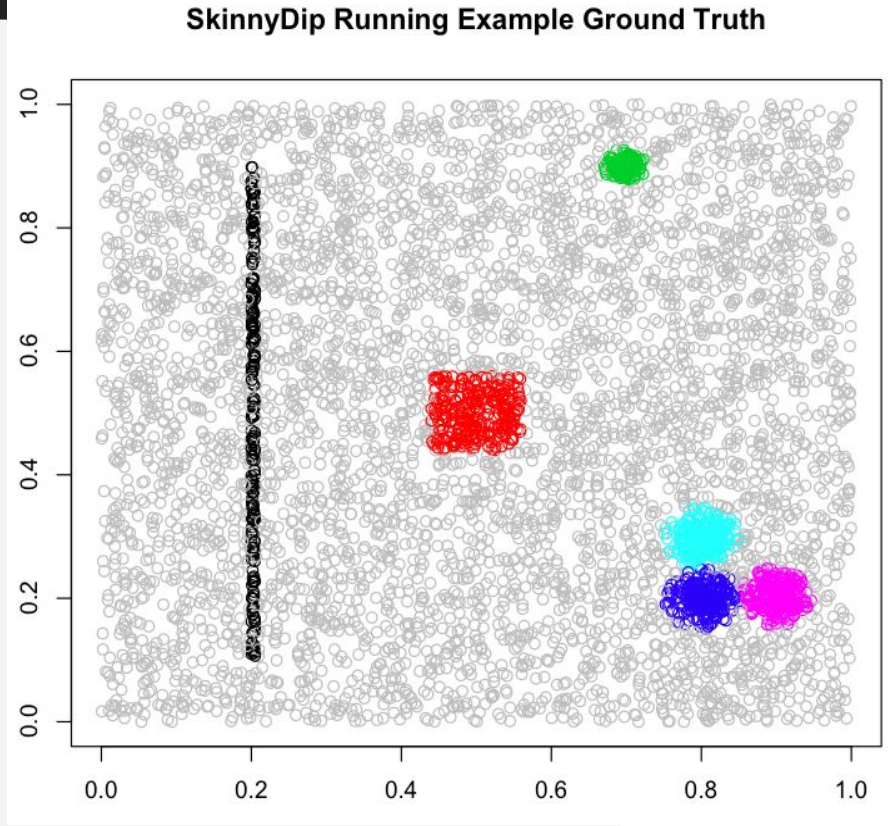

Random Data SkinnyDip+Recluster
Total: 1352 Correct: 0.925, FalseN: 0.075444, FalseP: 0.070988
Total: 308 Correct: 0.912, FalseN: 0.087662, FalseP: 0.007092
Total: 340 Correct: 0.944, FalseN: 0.055882, FalseP: 0.054217
0.0
0.2
0.4
0.6
0.8
1.0

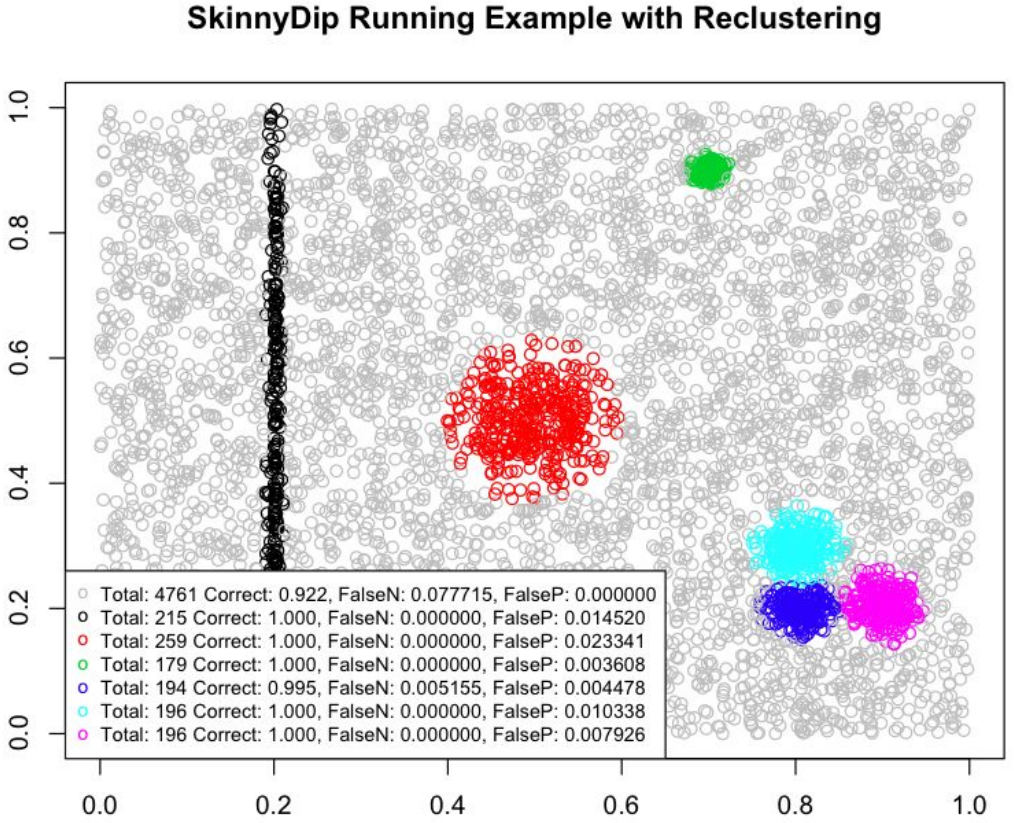
SkinnyDip Running Example Unlabeled



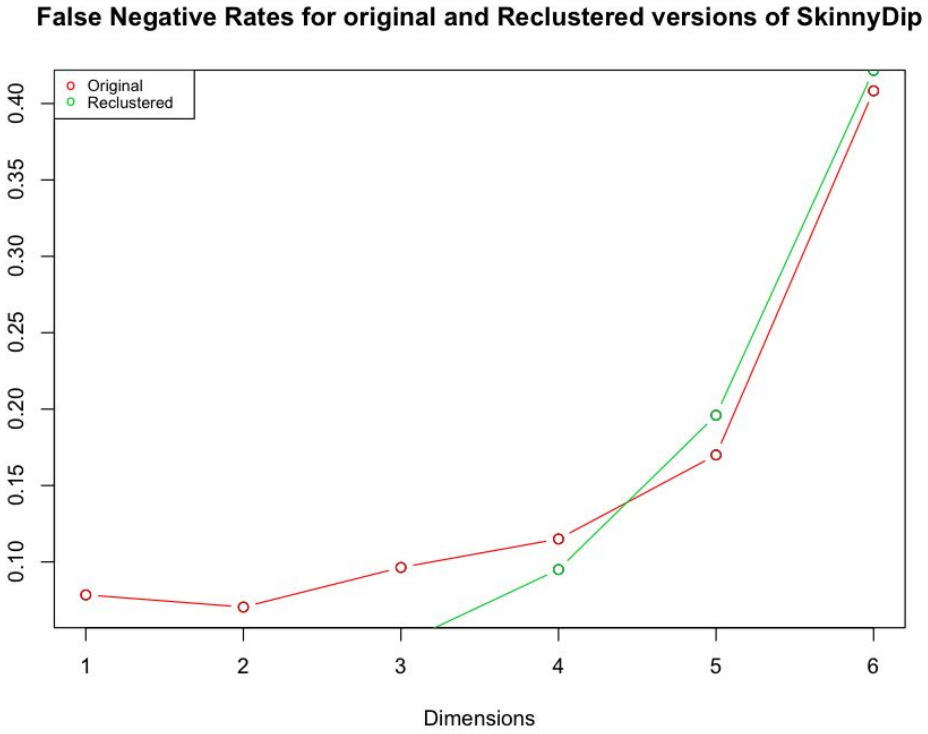

SkinnyDip Running Example Ground Truth
0.0
0.2
0.4
0.6
0.8
1.0


SkinnyDip Running Example
Total: 4761 Correct: 0.968, FalseN: 0.032346, FalseP: 0.032284
Total: 215 Correct: 0.991, FalseN: 0.009302, FalseP: 0.007952
Total: 259 Correct: 0.973, FalseN: 0.027027, FalseP: 0.008883
Total: 179 Correct: 0.983, FalseN: 0.016760, FalseP: 0.002062
Total: 194 Correct: 0.974, FalseN: 0.025773, FalseP: 0.000689
Total: 196 Correct: 0.985, FalseN: 0.015306, FalseP: 0.003446
Total: 196 Correct: 0.898, FalseN: 0.102041, FalseP: 0.003618
0.0
0.2
0.4
0.6
0.8
1.0


SkinnyDip Running Example Ground Truth
0.0
0.2
0.4
0.6
0.8
1.0


SkinnyDip Running Example with Reclustering
Total: 4761 Correct: 0.922, FalseN: 0.077715, FalseP: 0.000000
Total: 215 Correct: 1.000, FalseN: 0.000000, FalseP: 0.014520
Total: 259 Correct: 1.000, FalseN: 0.000000, FalseP: 0.023341
Total: 179 Correct: 1.000, FalseN: 0.000000, FalseP: 0.003608
Total: 194 Correct: 0.995, FalseN: 0.005155, FalseP: 0.004478
Total: 196 Correct: 1.000, FalseN: 0.000000, FalseP: 0.010338
Total: 196 Correct: 1.000, FalseN: 0.000000, FalseP: 0.007926
0.0
0.2
0.4
0.6
0.8
1.0

Conclusions



- Reclustering to balance out error rates is only practical in lower dimensions
- Small neglected regions still exponentially increase error in higher dimensions
- Less uniform the background noise means more prone to extraneous clusters
- Authors discuss the uses of SkinnyDip on real-world noisy data. Locational classification in 2 or 3 dimensions is still practical

References

- Samuel Maurus and Claudia Plant. 2016. Skinny-dip: Clustering in a Sea of Noise. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '16). ACM, New York, NY, USA, 1055-1064. DOI: https://doi.org/10.1145/2939672.2939740

- J. A. Hartigan and P. Hartigan. The dip test of unimodality. The Annals of Statistics, 1985.

Questions?