

# Supplemental Material

Empirical Guidance on Scatterplot and  
Dimension Reduction Technique Choices

*Michael Sedlmair, Tamara Munzner, Melanie Tory*

# Content

<b>A. List of Datasets .....</b>	<b>Slide 4</b>
<b>B. Heatmaps</b>	
B.1. Labeled Heatmaps .....	Slides 6-13
B.2. Separate Heatmaps for Coders ....	Slides 15-20
B.3. Difference between Coders .....	Slides 22-24
<b>C. Selected Examples</b>	
C.1. Within-DR: SPLOM .....	Slides 26-31
C.2. Within-DR: i3D .....	Video
C.3. Between-DR .....	Slides 34-36
<b>D. SPLOM Evaluation .....</b>	<b>Slide 38</b>
<b>E. Coding Guidelines .....</b>	<b>Slide 40</b>

# Pointers from the paper

## Section 4.4 / 4.5

- all scatterplot samples ..... Slide 4
- labeled heatmaps ..... Slides 6-13
- raw data / heatmaps for both coders .. Slides 15-20
- further examples ..... Slides 26-36 / Video

## Figure 5 / Section 5.3.1

- video of entangled dataset ..... Video
- PCA and robPCA 2D Scatterplots ..... Slide 34

## Section 5.5

- details about SPLOM analysis ..... Slide 38

# A. List of Datasets

List sorting = heatmap sorting

<b>id</b>	<b>Dataset name</b>	<b>Category</b>	<b>Points</b>	<b>Dimensions</b>	<b>Classes</b>	<b>Provenance</b>
1	abalone	real	4154	7	28	uci
2	cars-2	real	7404	22	53	colleagues
3	industryIndices	real	102	6	13	uci
4	worldMap	real	192	3	13	visumap
5	cars-3	real	7404	22	12	colleagues
6	fisheriesEscapementTarget	real	121	12	11	colleagues
7	fisheriesHarvestRule	real	121	12	11	colleagues
8	yeast	real	1452	8	10	uci
9	ecoliProteins	real	332	7	8	visumap
10	eFashion	real	3272	4	8	sap
11	tse300	real	244	49	8	visumap
12	cereal	real	77	12	7	xmdv
13	shuttle-big	real	43500	9	7	uci
14	shuttle-small	real	14500	9	7	uci
15	musicNetGroups	real	171	9	6	visumap
16	hiv	real	78	159	6	colleagues
17	bbdm13	real	200	13	5	umass
18	pageBlocks	real	5473	10	5	uci
19	world-12d	real	151	12	5	visumap
20	world-10d	real	151	10	5	visumap
21	bostonHousing	real	155	13	3	uci
22	iris	real	147	4	3	uci
23	olive	real	572	8	3	colleagues
24	swanson	real	1875	6	3	xmdv
25	wine	real	178	13	3	uci
26	breastCancer-original	real	454	9	2	uci
27	cars-1	real	7404	22	2	colleagues
28	ionosphere	real	351	34	2	visumap
29	parkinson	real	195	11	2	uci
30	spamBase	real	4601	57	2	uci
31	breastCancer-diagnostic	real	569	30	2	uci
32	gauss-n100-10d-5smallCI	gaussian	100	10	5	
33	gauss-n100-10d-5largeCI	gaussian	100	10	5	
34	gauss-n100-5d-5smallCI	gaussian	100	5	5	
35	gauss-n100-5d-5largeCI	gaussian	100	5	5	
36	gauss-n500-10d-5smallCI	gaussian	500	10	5	
37	gauss-n500-10d-5largeCI	gaussian	500	10	5	
38	gauss-n500-5d-5smallCI	gaussian	500	5	5	
39	gauss-n500-5d-5largeCI	gaussian	500	5	5	
40	gauss-n100-10d-3smallCI	gaussian	100	10	3	
41	gauss-n100-10d-3largeCI	gaussian	100	10	3	
42	gauss-n100-5d-3smallCI	gaussian	100	5	3	
43	gauss-n100-5d-3largeCI	gaussian	100	5	3	
44	gauss-n500-10d-3smallCI	gaussian	500	10	3	
45	gauss-n500-10d-3largeCI	gaussian	500	10	3	
46	gauss-n500-5d-3smallCI	gaussian	500	5	3	
47	gauss-n500-5d-3largeCI	gaussian	500	5	3	

<b>id</b>	<b>Dataset name</b>	<b>Category</b>	<b>Points</b>	<b>Dimensions</b>	<b>Classes</b>
48	entangled2-15d-adjacent	interleaved	2049	15	15
49	entangled2-15d-overlap	interleaved	2318	15	15
50	entangled2-10d-adjacent	interleaved	1490	10	10
51	entangled2-10d-overlap	interleaved	1479	10	10
52	entangled2-6d-adjacent	interleaved	837	6	6
53	entangled2-6d-overlap	interleaved	1034	6	6
54	entangled2-5d-adjacent	interleaved	741	5	5
55	entangled2-5d-overlap	interleaved	696	5	5
56	entangled1-3d-5cl-separate	interleaved	500	3	5
57	entangled2-4d-adjacent	interleaved	1254	4	4
58	entangled2-4d-overlap	interleaved	538	4	4
59	entangled1-3d-4cl-separate	interleaved	400	3	4
60	entangled2-3d-overlap	interleaved	857	3	3
61	entangled2-3d-adjacent	interleaved	1098	3	3
62	entangled1-3d-3cl-separate	interleaved	600	3	3
63	entangled3-l-3d-smallOver	interleaved	496	3	3
64	entangled3-l-3d-bigOverlap	interleaved	571	3	3
65	entangled3-s-3d-adjacent	interleaved	185	3	3
66	entangled3-s-3d-bigOverlap	interleaved	205	3	3
67	entangled3-xl-3d-adjacent	interleaved	1821	3	3
68	entangled3-xl-3d-bigOverlap	interleaved	1892	3	3
69	entangled3-m-3d-adjacent	interleaved	309	3	3
70	entangled3-m-3d-smallOverlap	interleaved	292	3	3
71	entangled3-m-3d-bigOverlap	interleaved	325	3	3
72	grid-4d	grid	1296	4	16
73	grid-3d	grid	1000	3	8
74	twoSquare	grid	968	3	4
75	unevenDensity	grid	905	3	2

A video of all resulting 816 scatterplot samples can be found at:

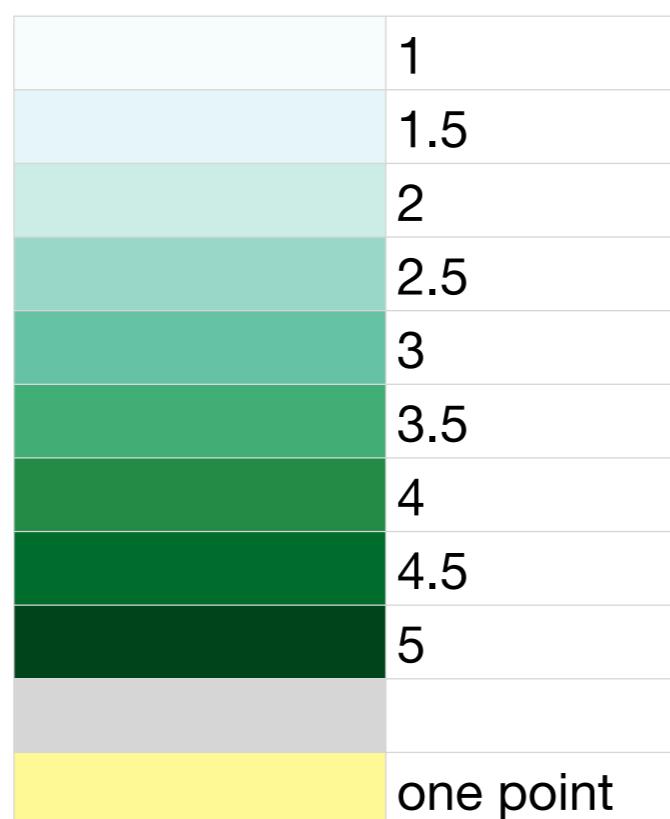
[http://www.cs.ubc.ca/labs/imager/video/2012/VisClusterSep/VisClusterSep\\_video1.avi](http://www.cs.ubc.ca/labs/imager/video/2012/VisClusterSep/VisClusterSep_video1.avi)

(60.8 MB, no audio, tested on VLC 1.1.12)

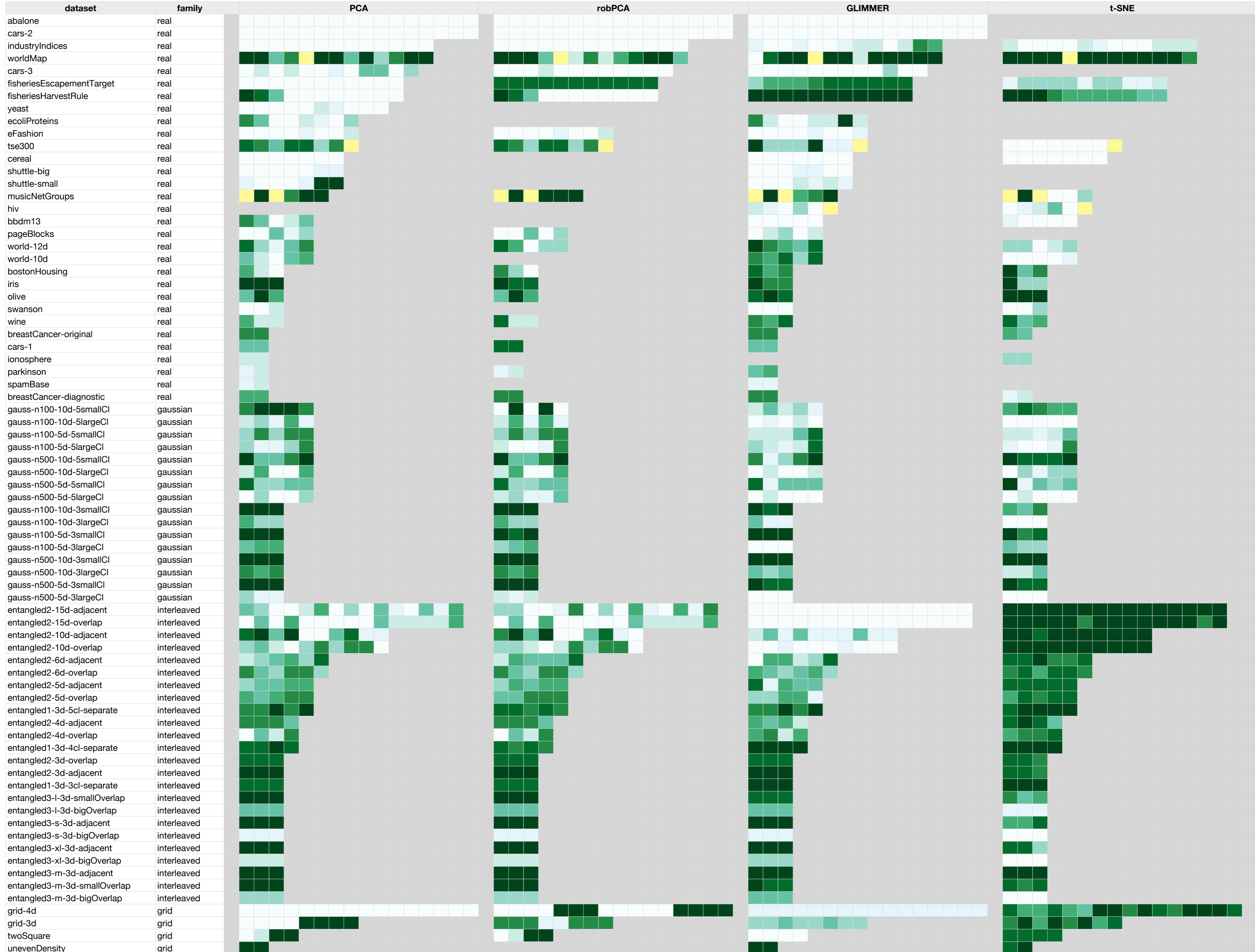
# B. Heatmaps

## B I. Labeled heatmaps

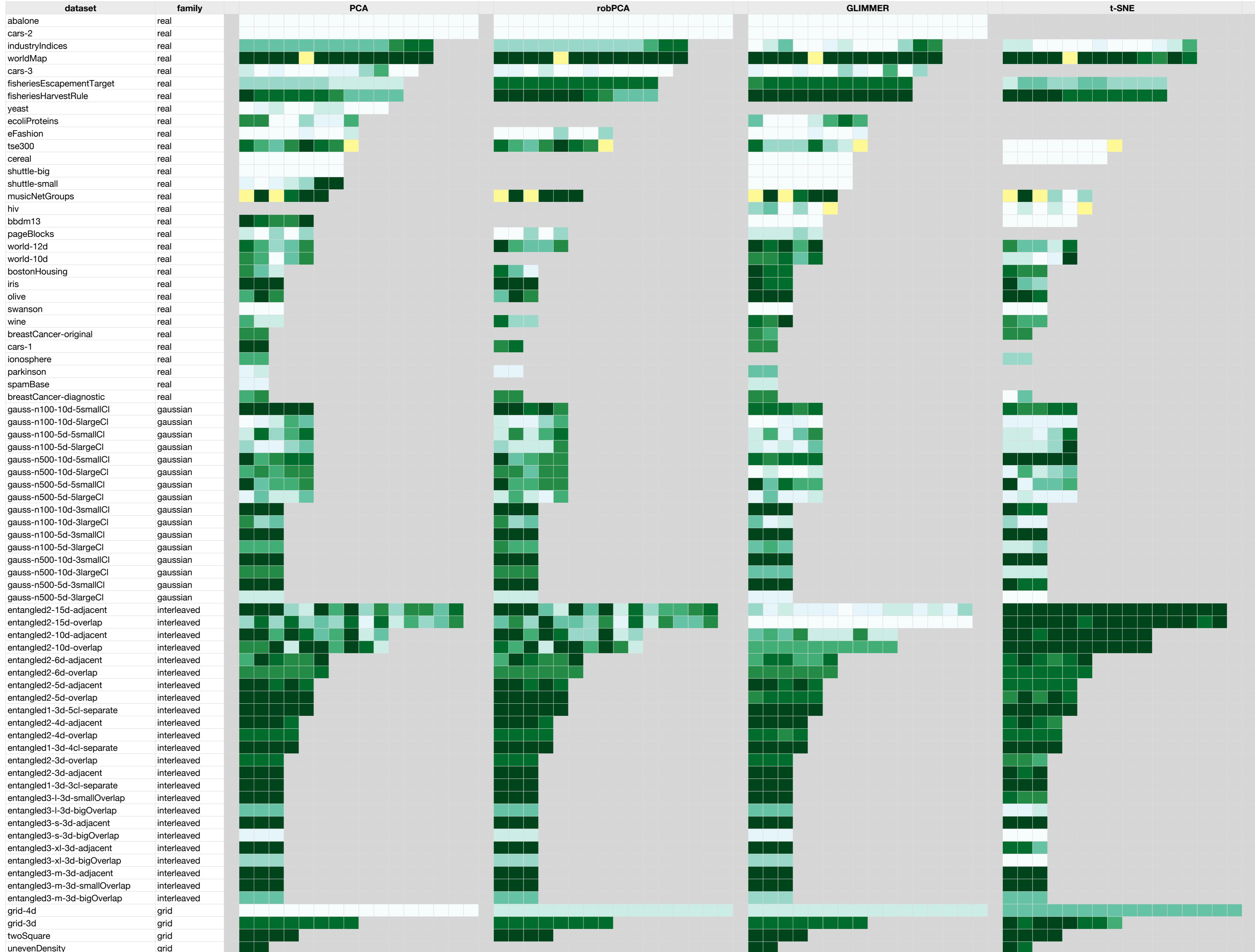
as extensions to Figure 2, 3, and 6 in  
the paper



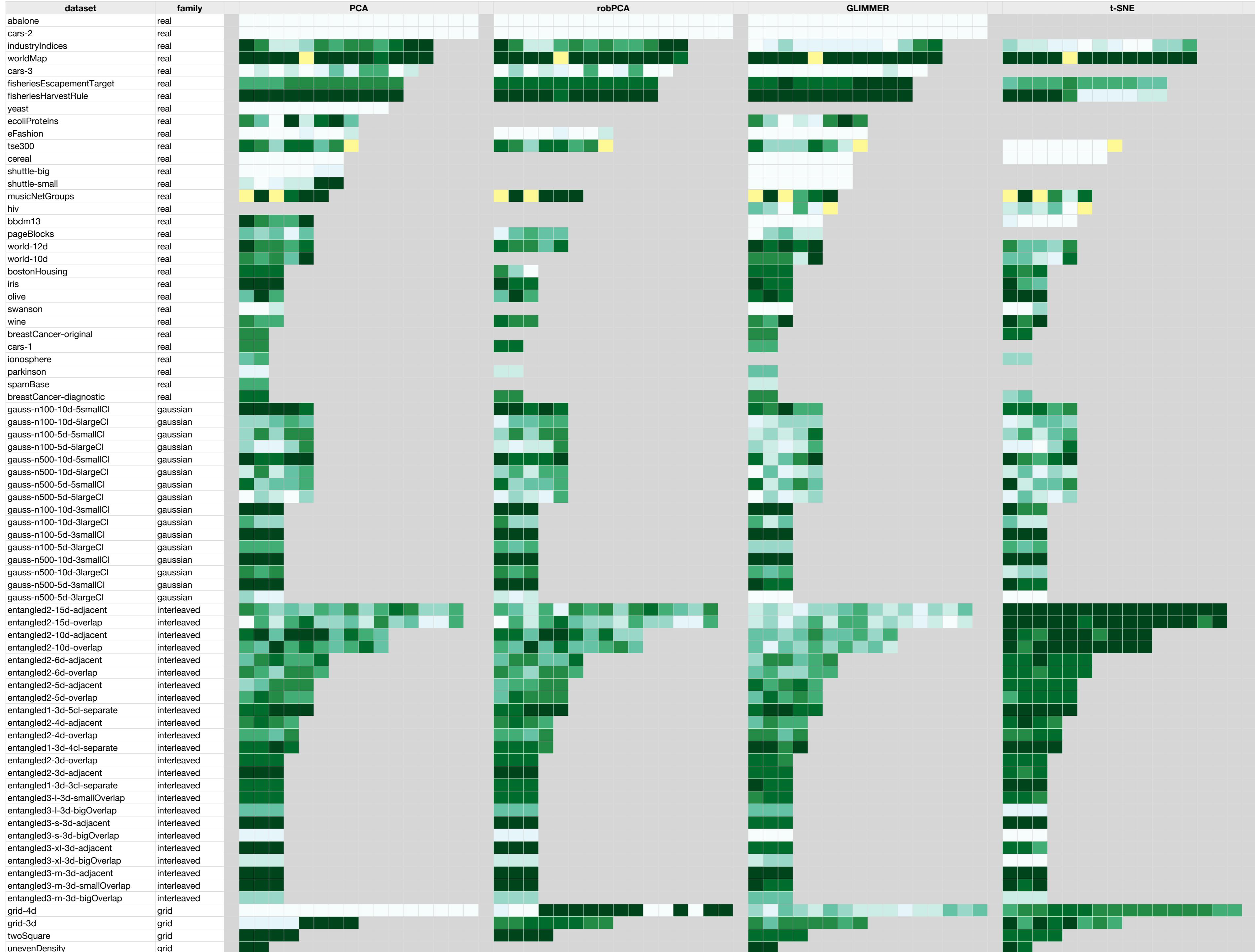
# B I. Labeled Heatmaps: Averaged base data 2D.



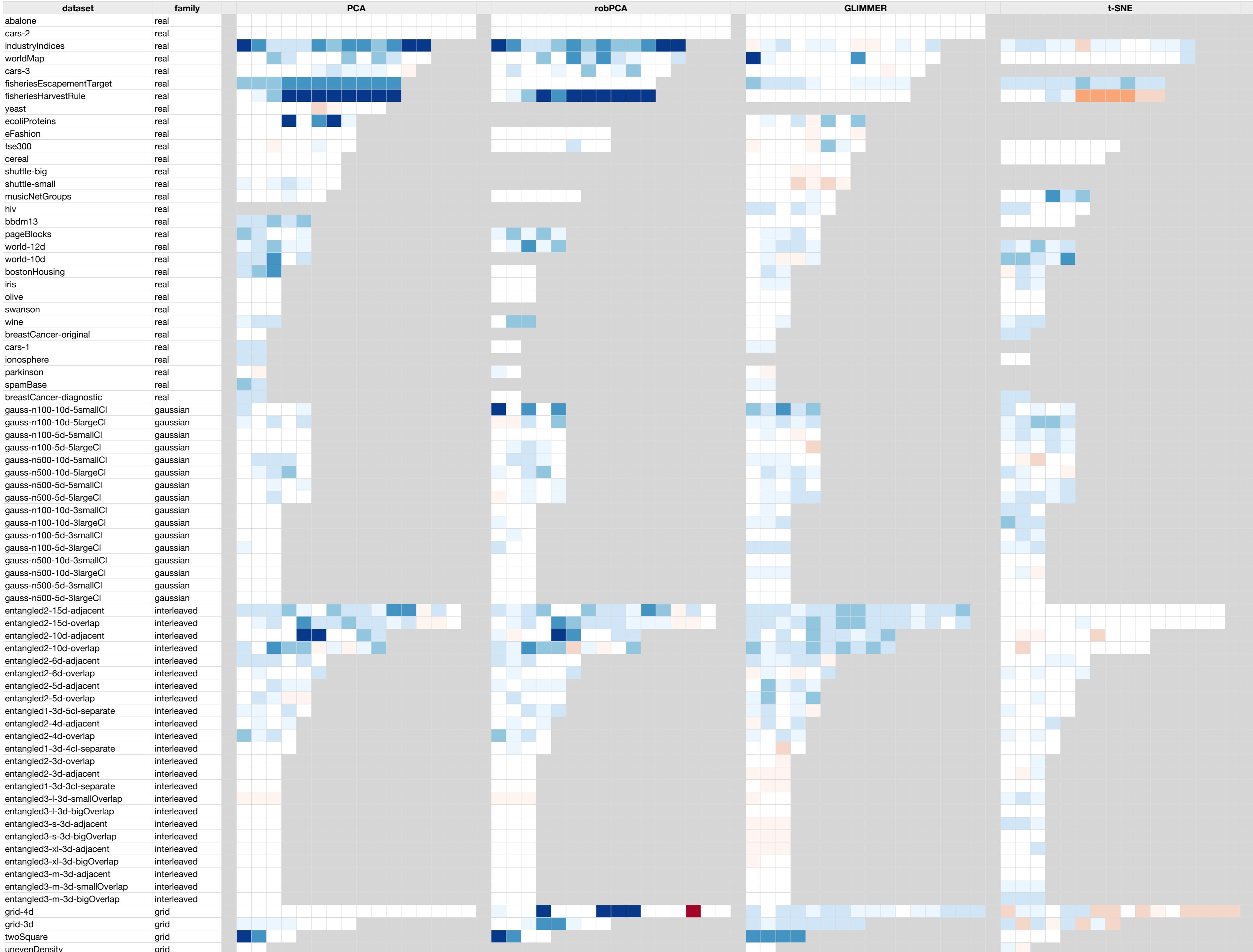
# B I. Labeled Heatmaps: Averaged base data i3D.



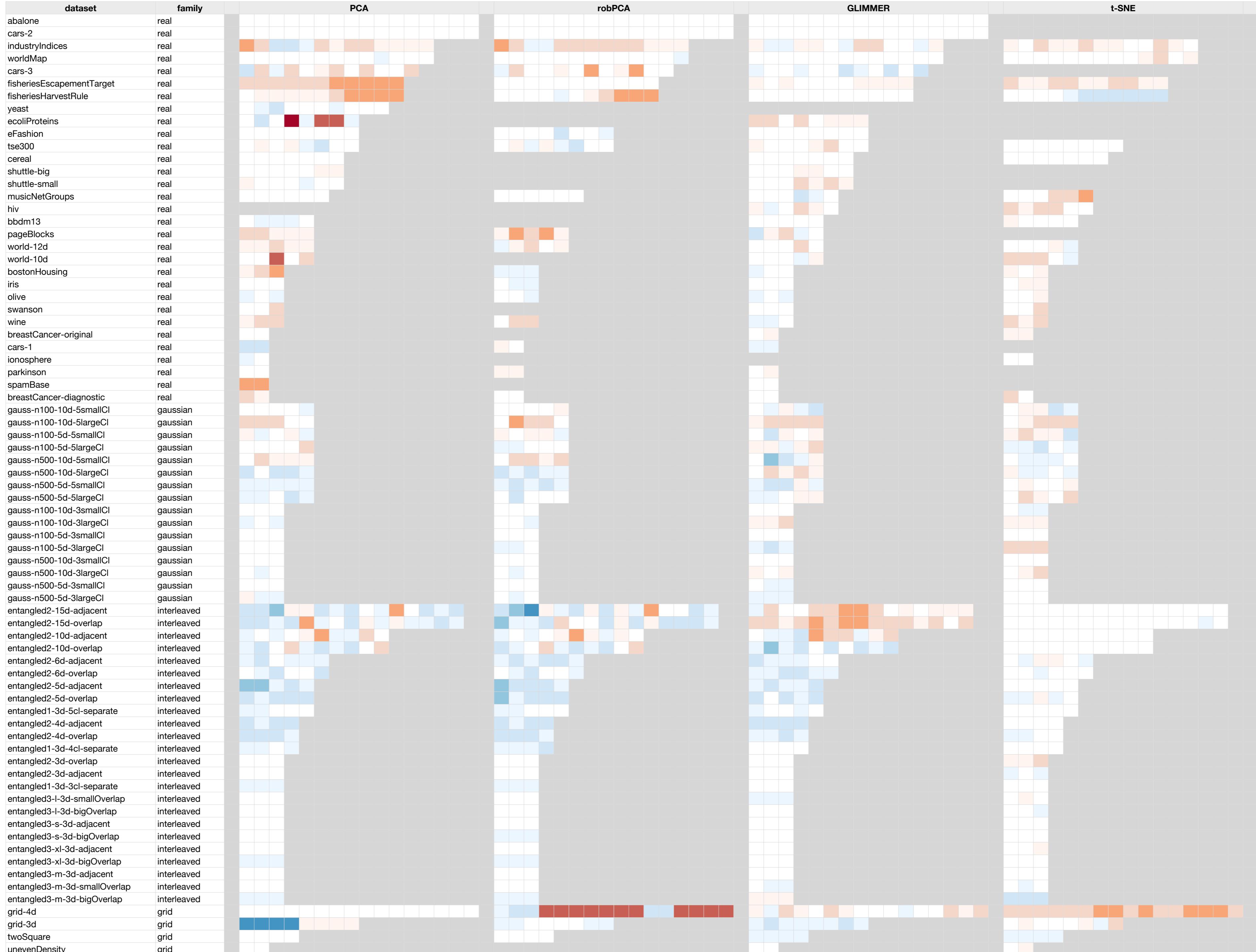
# B I. Labeled Heatmaps: Averaged base data SPLOM.



# B I. Labeled Heatmaps: SPLOM - 2D.



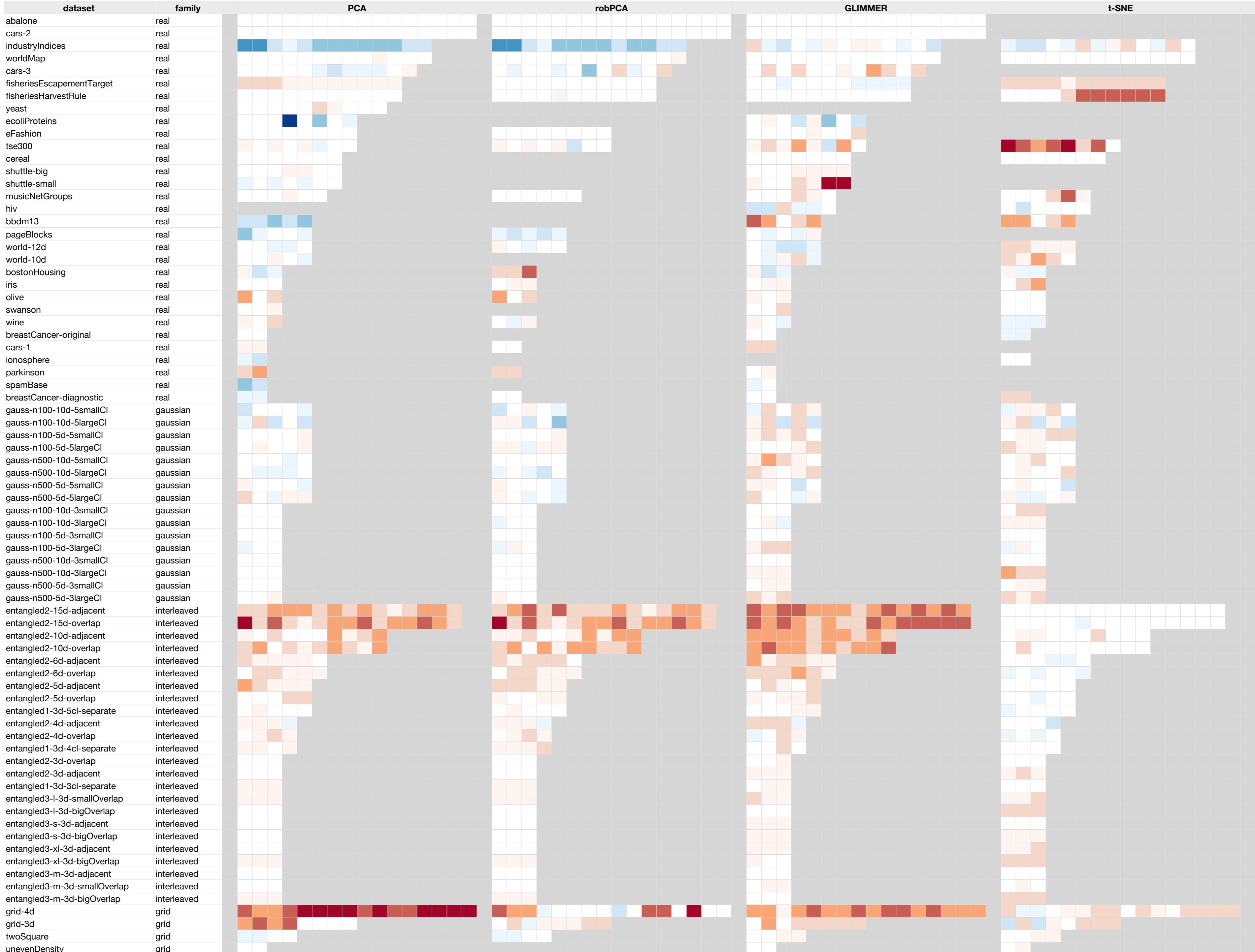
# B I. Labeled Heatmaps: i3D - max(2D, SPLOM).



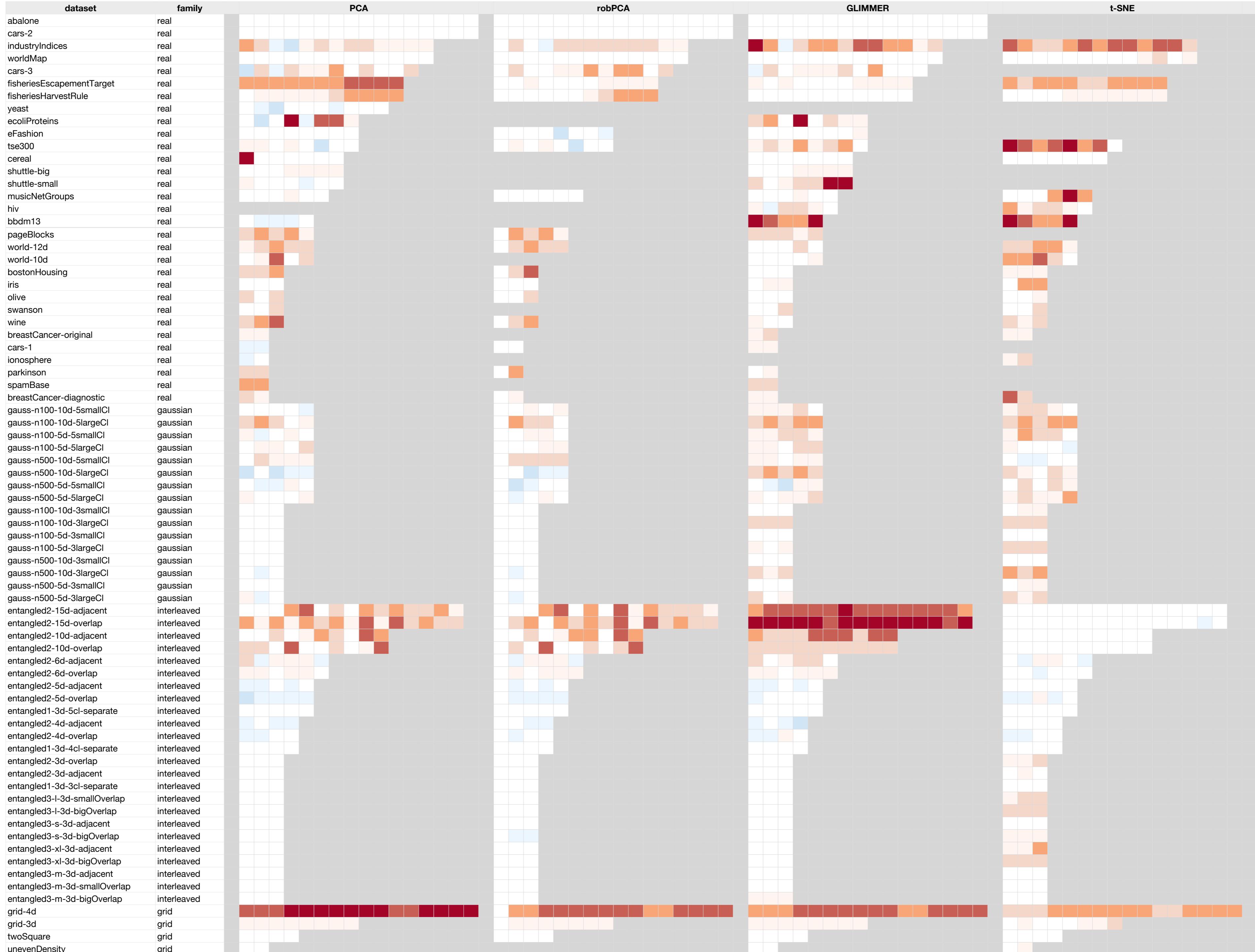
# B I. Labeled Heatmaps: 2D - max(2D from other DRs).



# B1. Labeled Heatmaps: SPLOM - max(2D from all DRs).

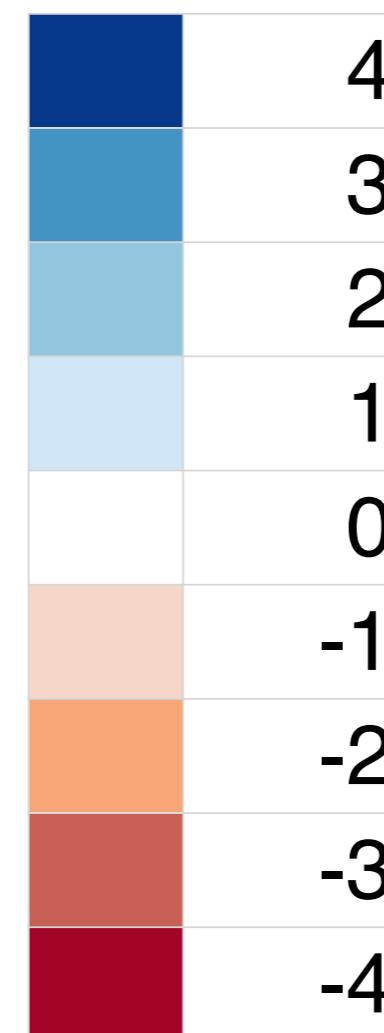
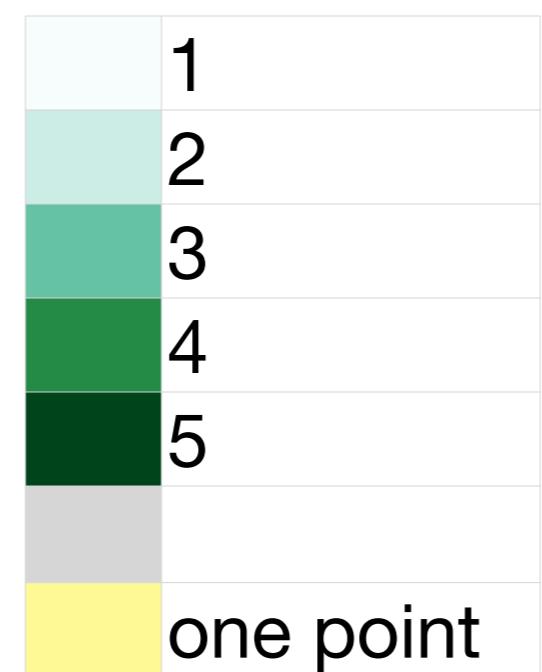


# B1. Labeled Heatmaps: i3D - max(SPLOM, 2D from all DRs).

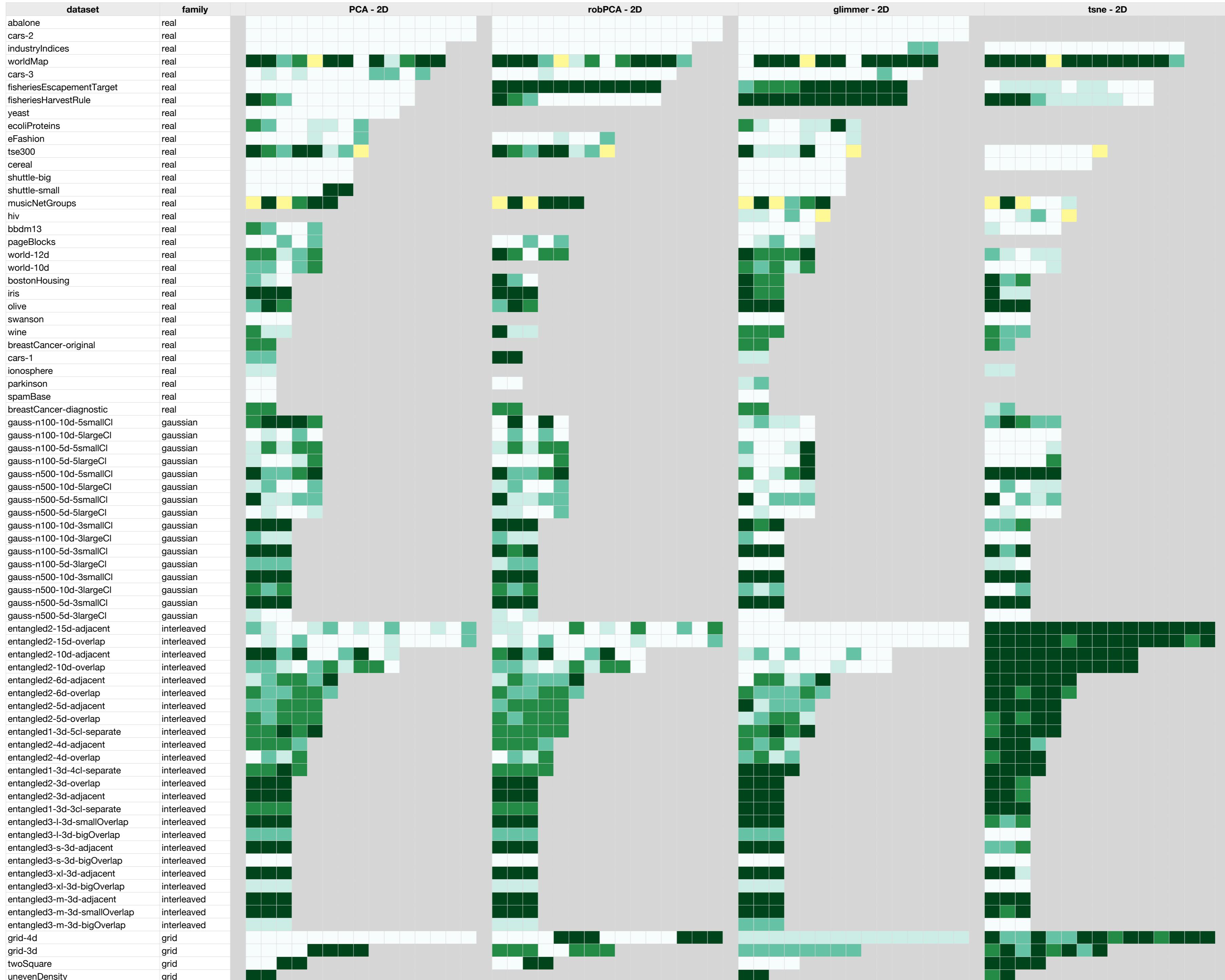


## B2. Separate Heatmaps for Coders

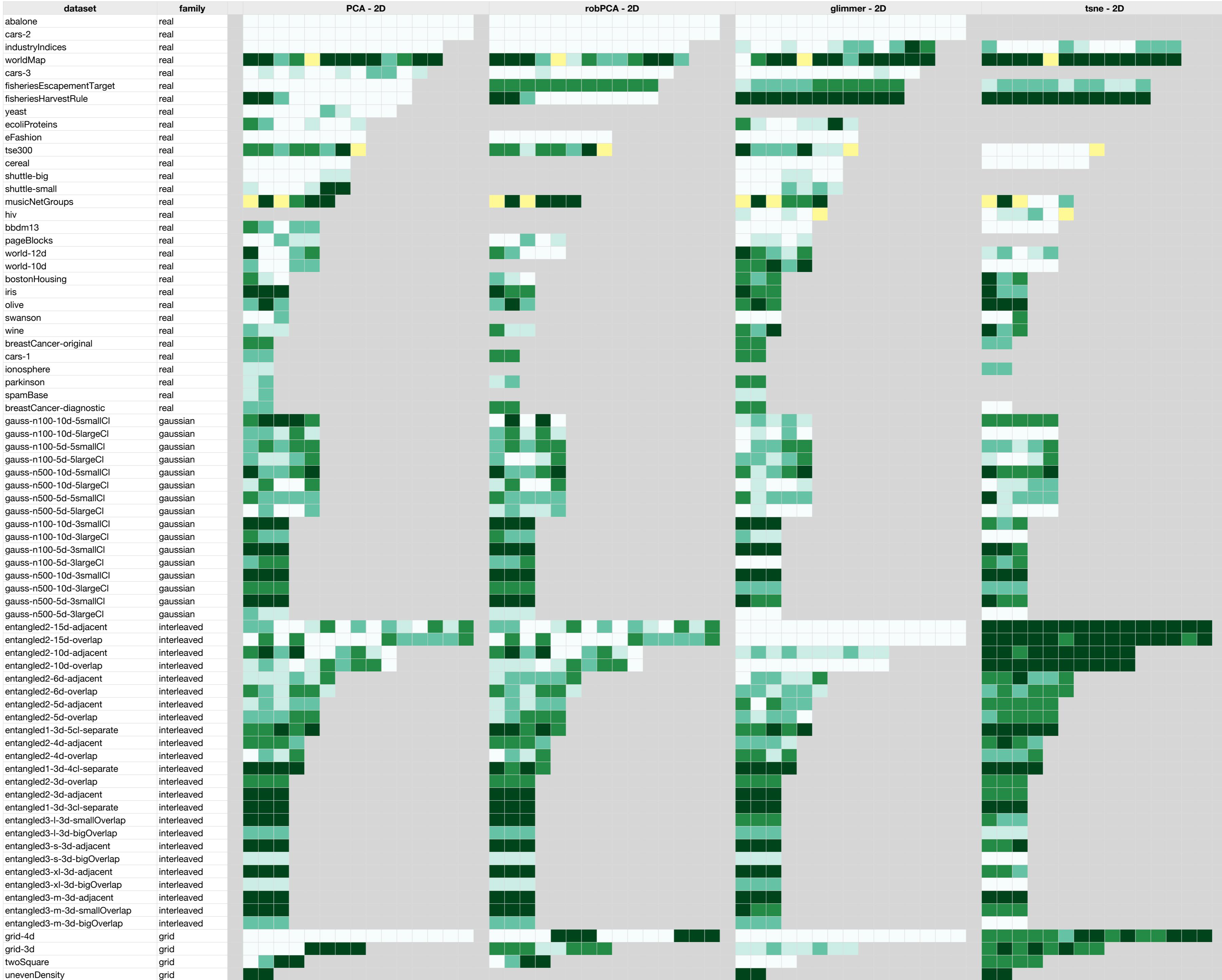
not in the paper



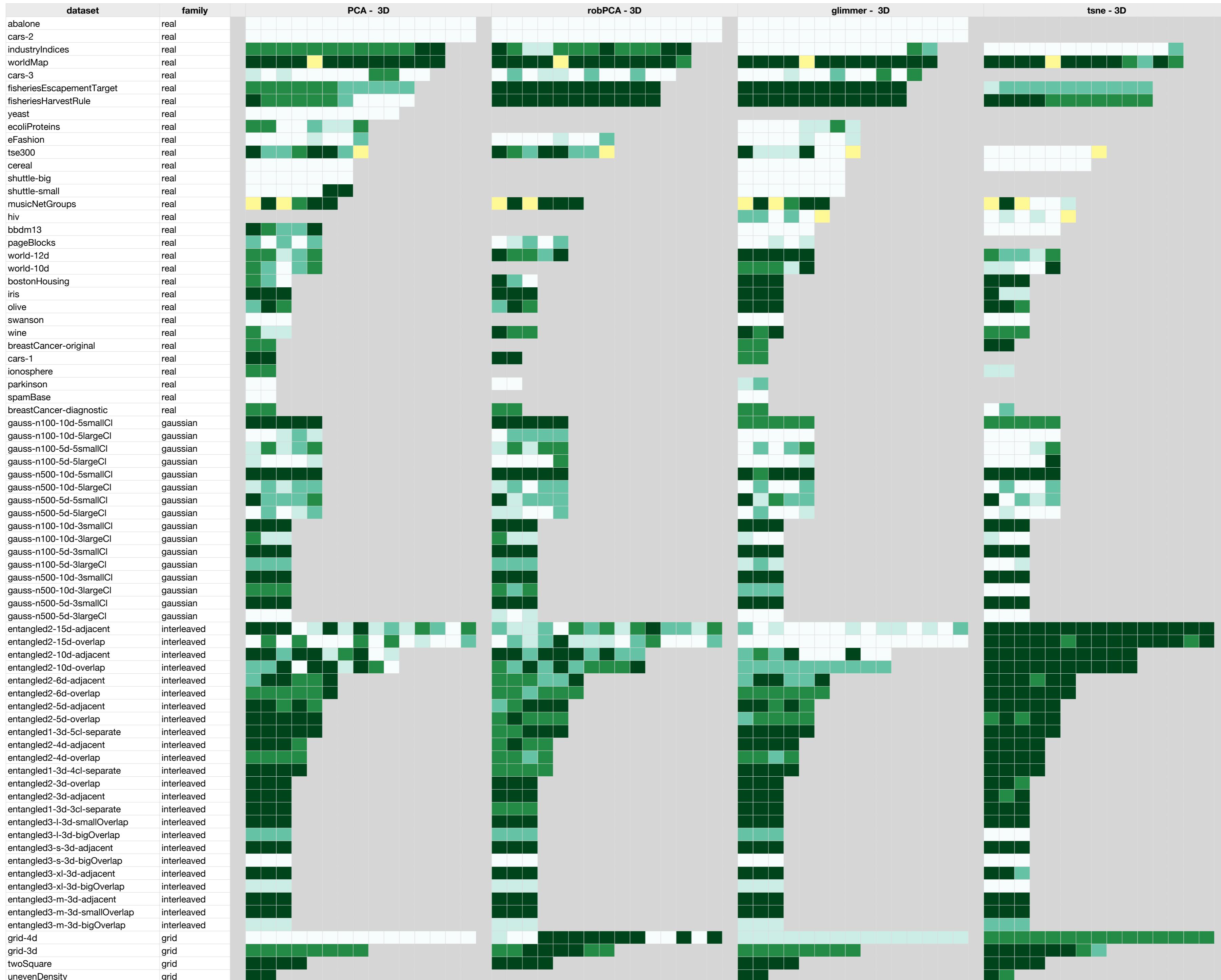
## B2. Heatmaps Coder A: Base data 2D.



## B2. Heatmaps Coder B: Base data 2D.



## B2. Heatmaps Coder A: Base data i3D.



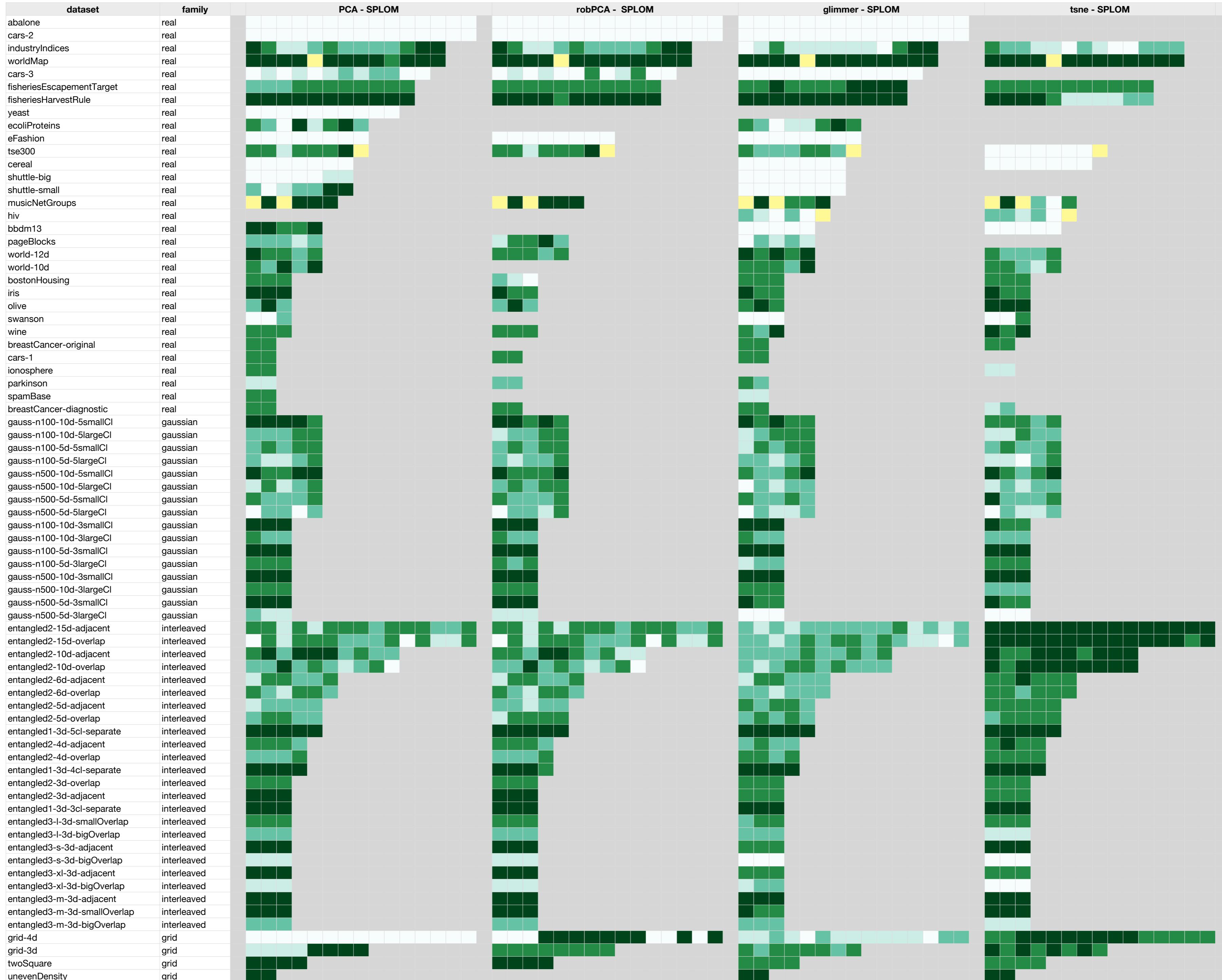
# B2. Heatmaps Coder B: Base data i3D.



## B2. Heatmaps Coder A: Base data SPLOM.



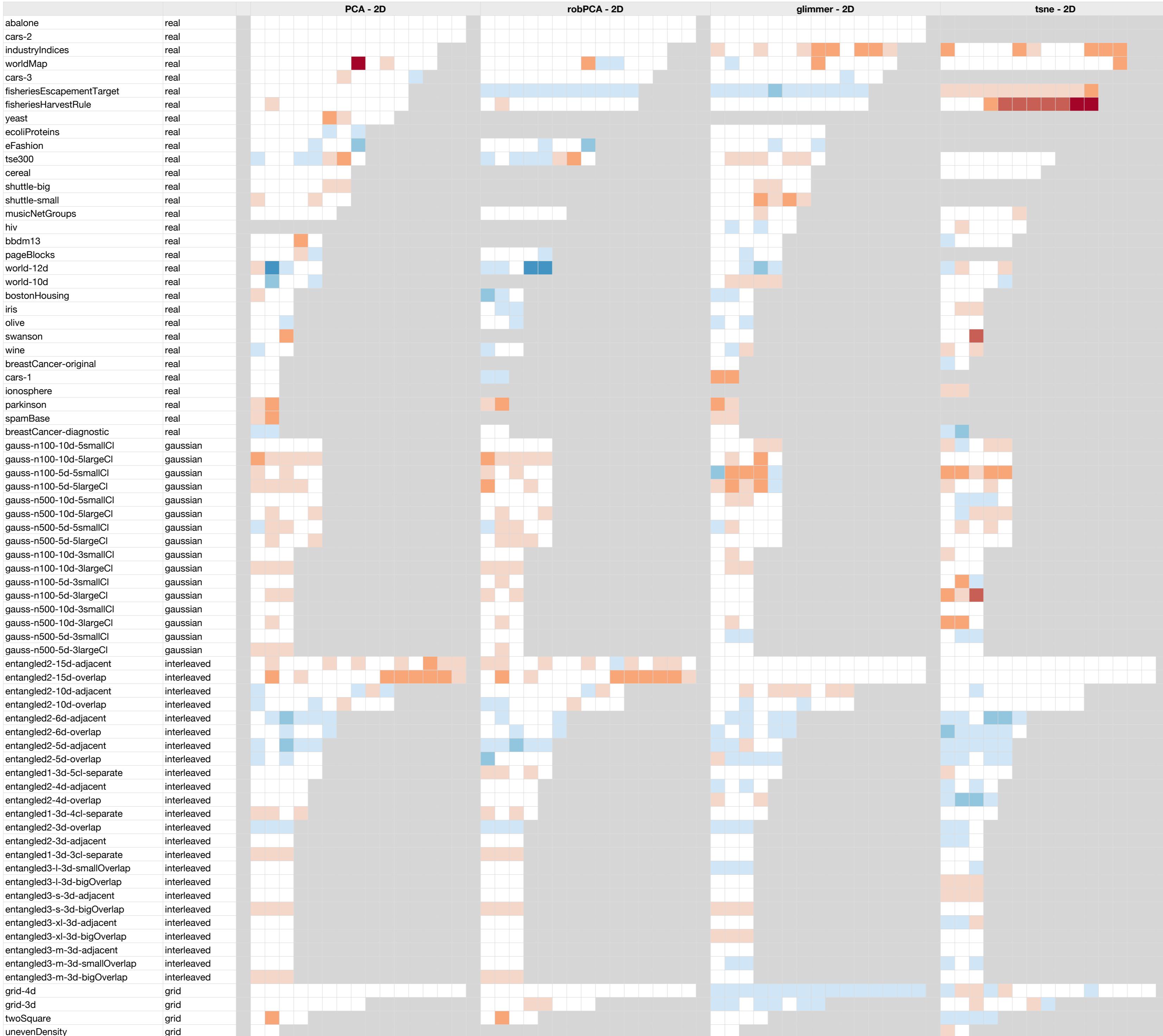
# B2. Heatmaps Coder B: Base data SPLOM.



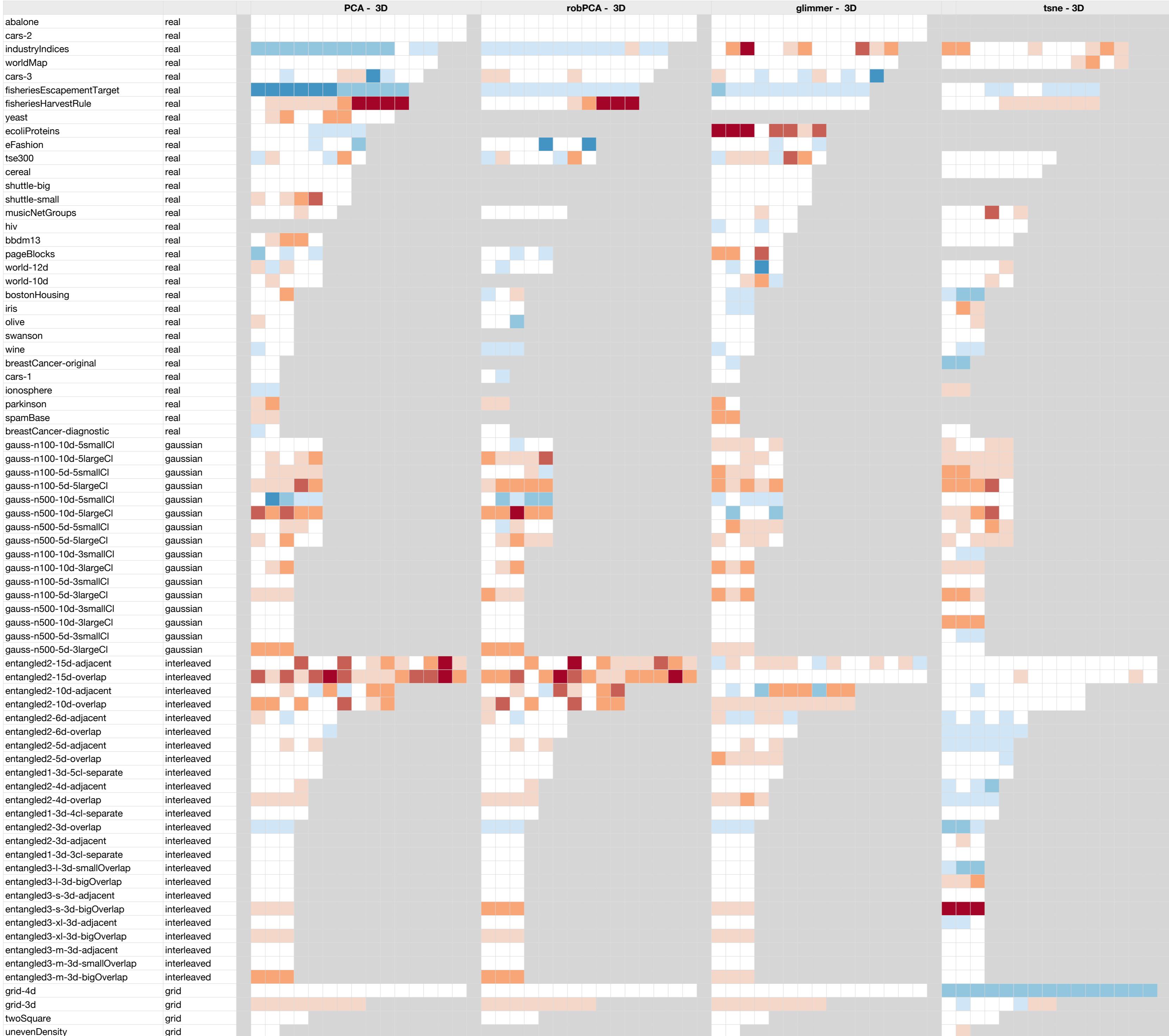
## **B3. Difference Heatmap between Coders**

not in the paper

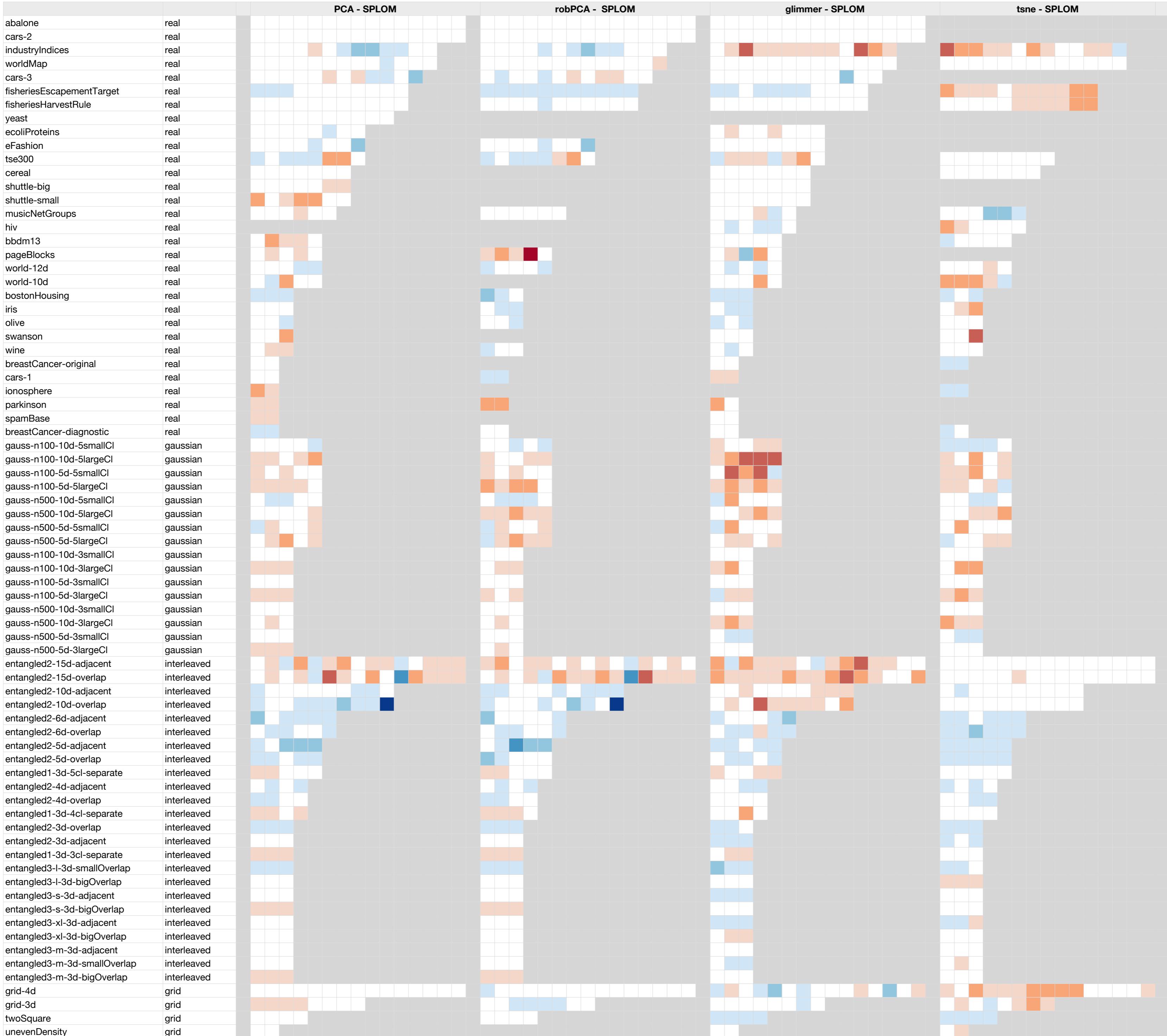
# B3. Difference Heatmap: Coder A - Coder B for 2D.



# B3. Difference Heatmap: Coder A - Coder B for i3D.



# B3. Difference Heatmap: Coder A - Coder B for SPLOM.



# C. Selected Examples

## C1. Within-DR: SPLOM > 2D

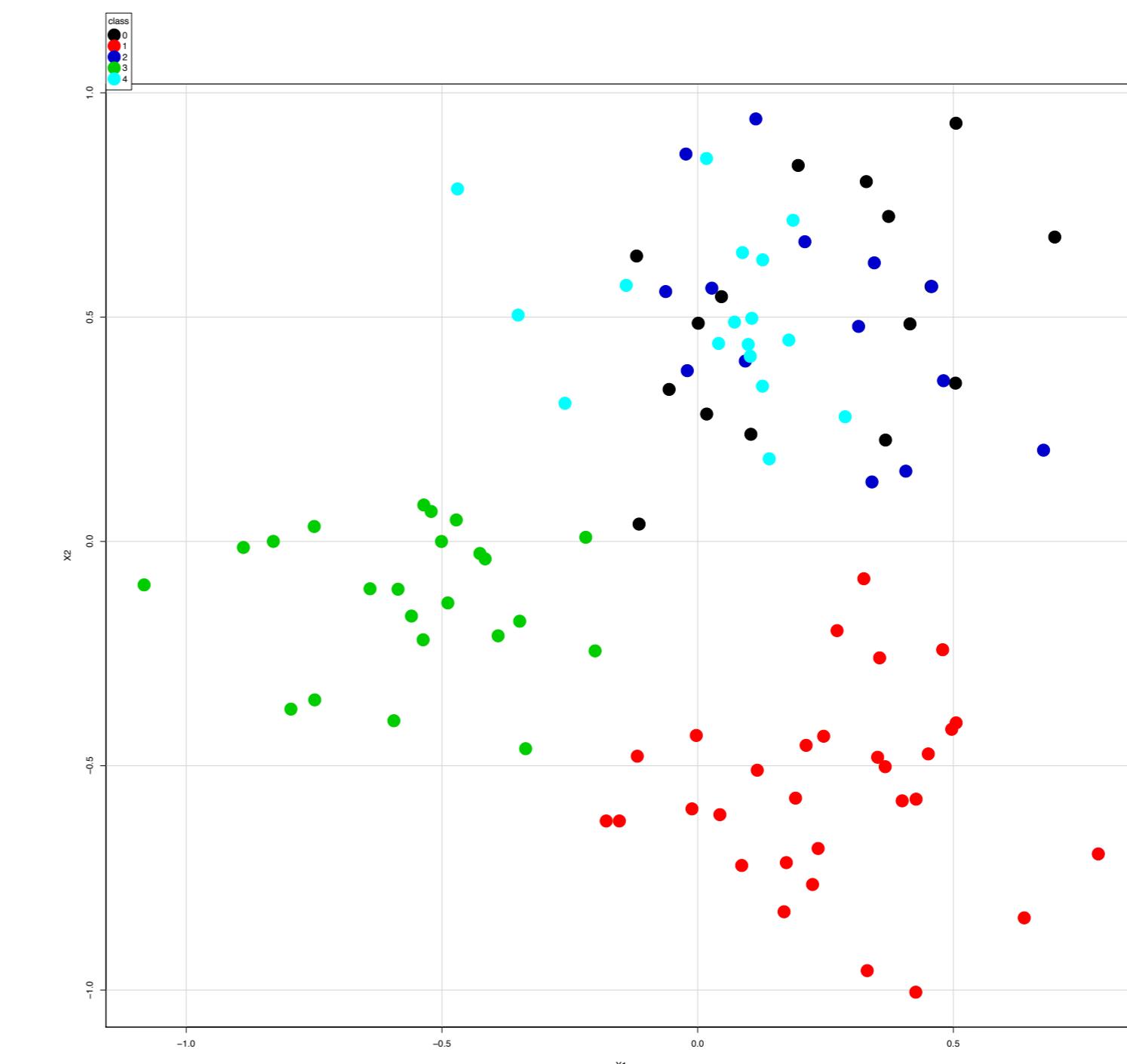
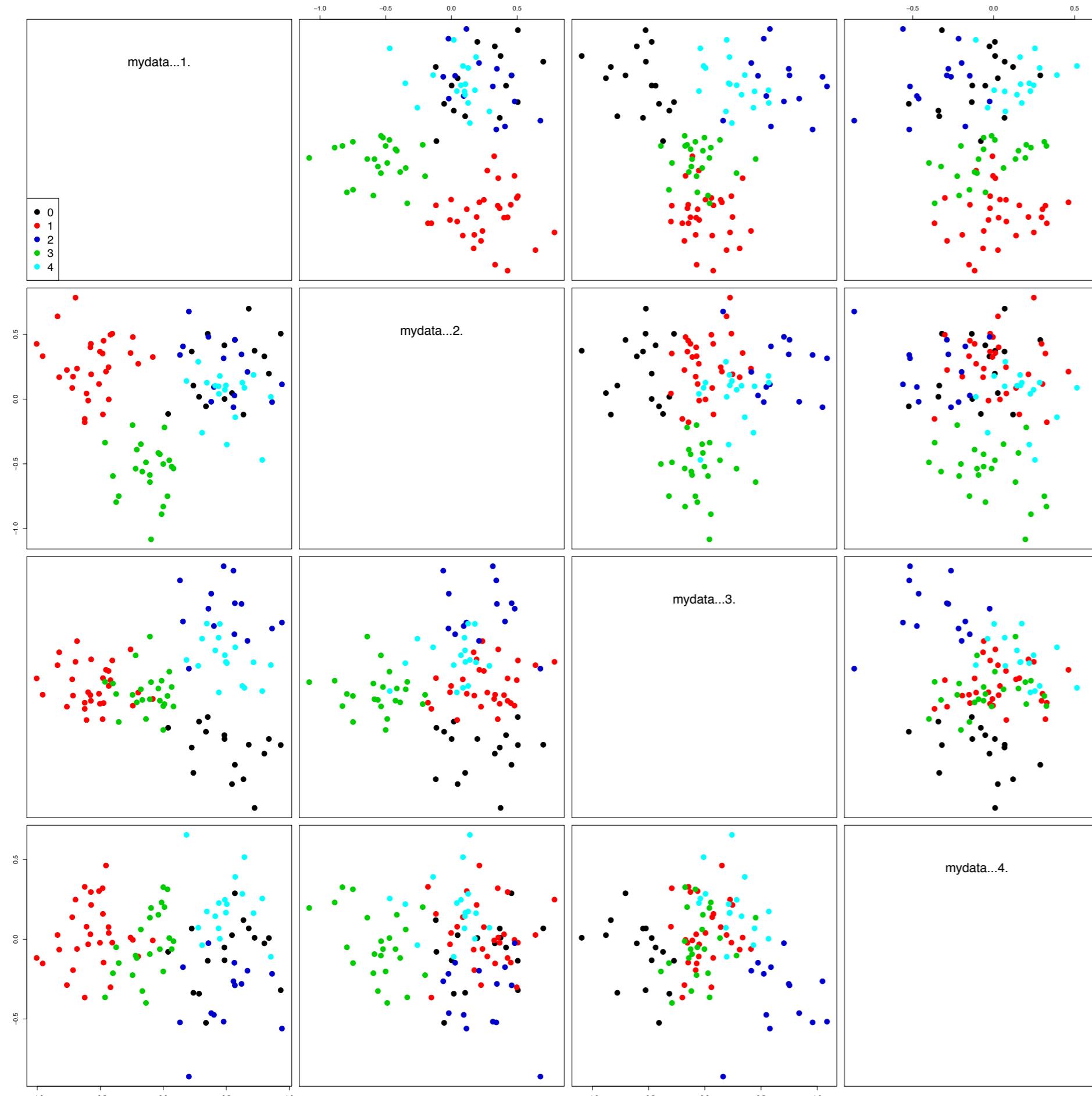
Examples where SPLOM added class separability within the same DR technique.

Note: Ratings are provided from Coder A

# CI. Within-DR - SPLOM (Figure 4a from the paper):

dataset: gauss-n|100-10d-5smallCI

DR: robust PCA



Classwise Ratings:

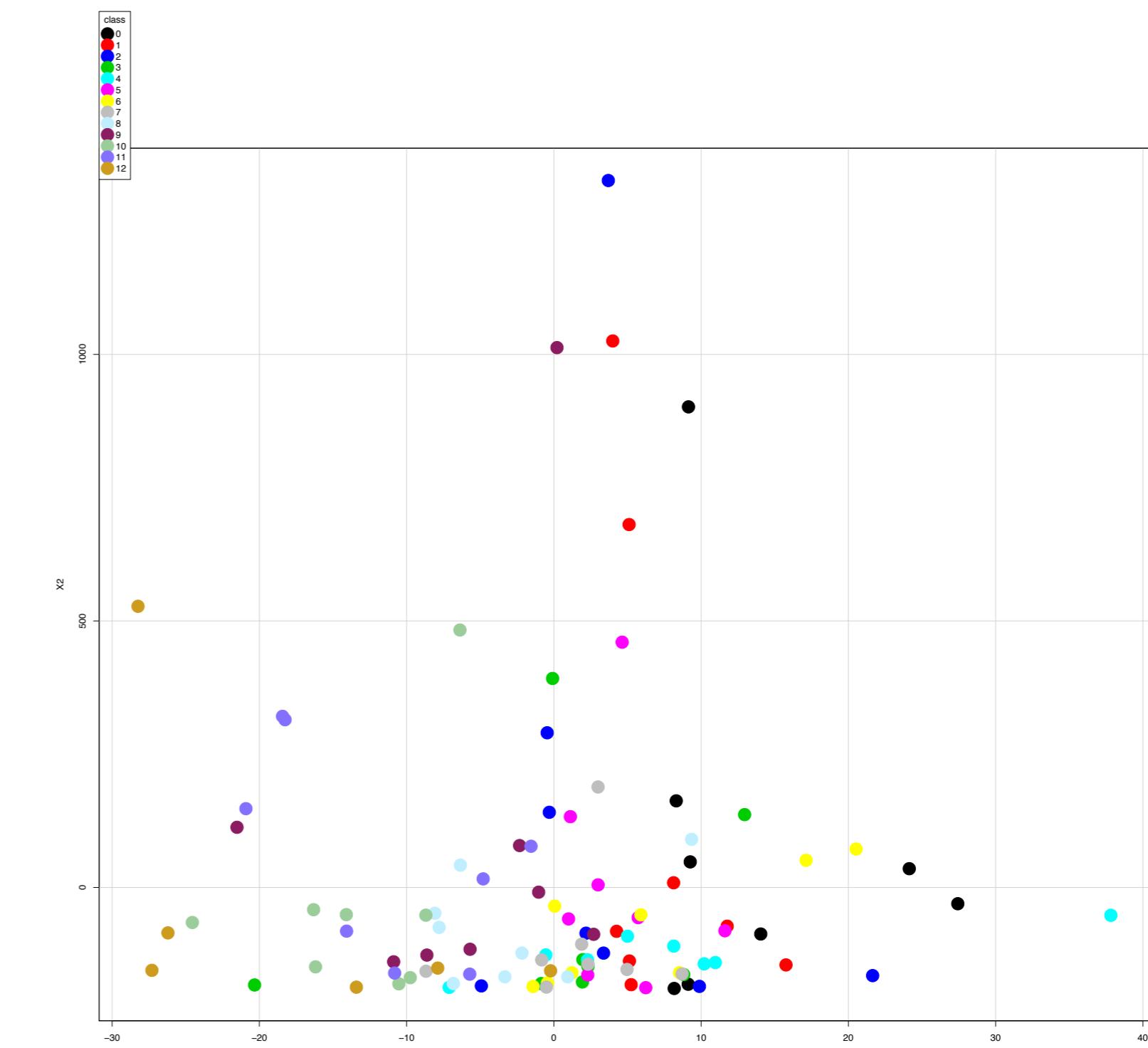
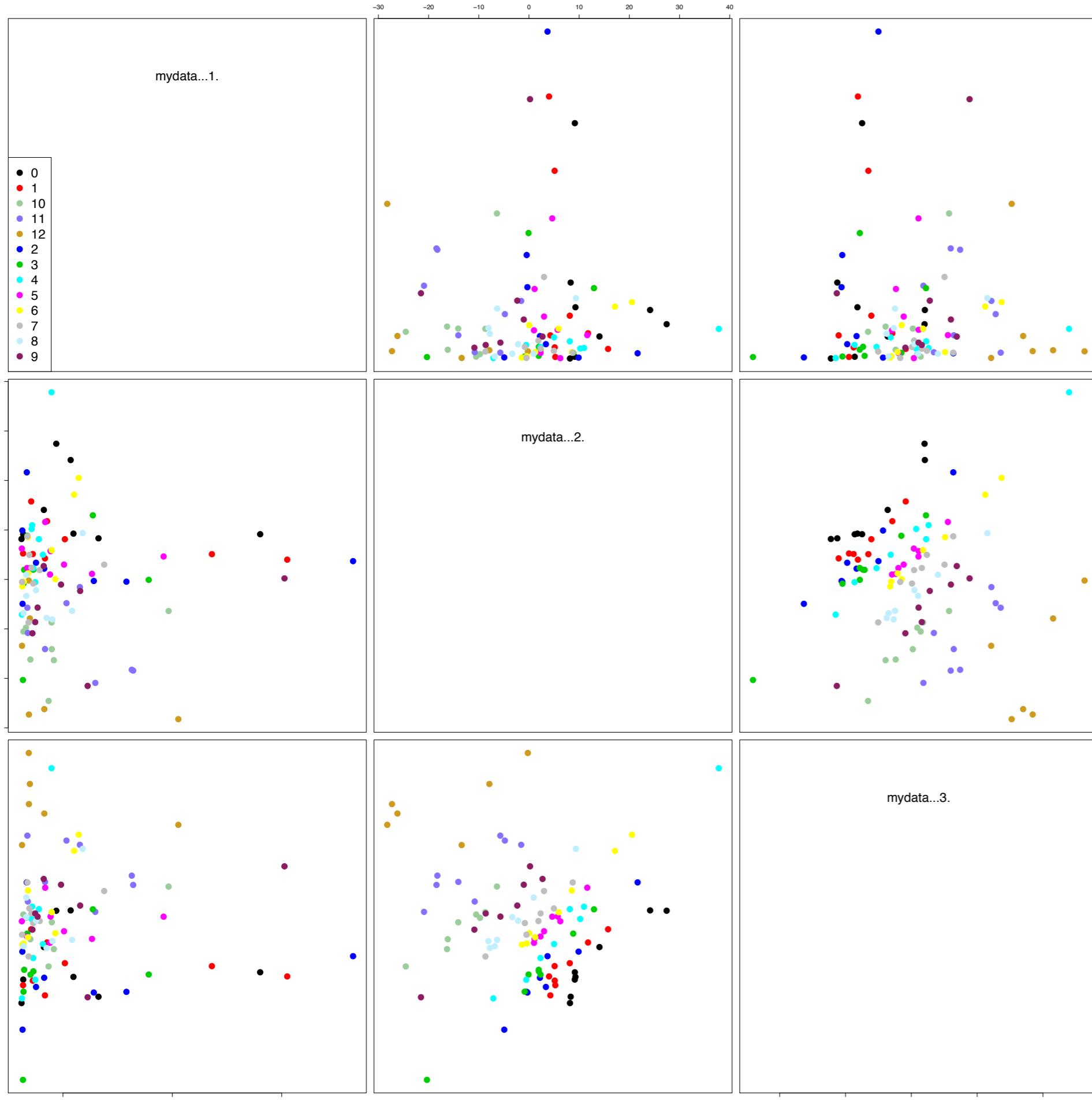
1 5 1 5 1

Classwise ratings and selected views:

5-2x3 5-1x2 5-3x4 5-1x2 5-1x3

# CI. Within-DR - SPLOM (Figure 4b from the paper):

dataset: industryIndices  
 DR: PCA



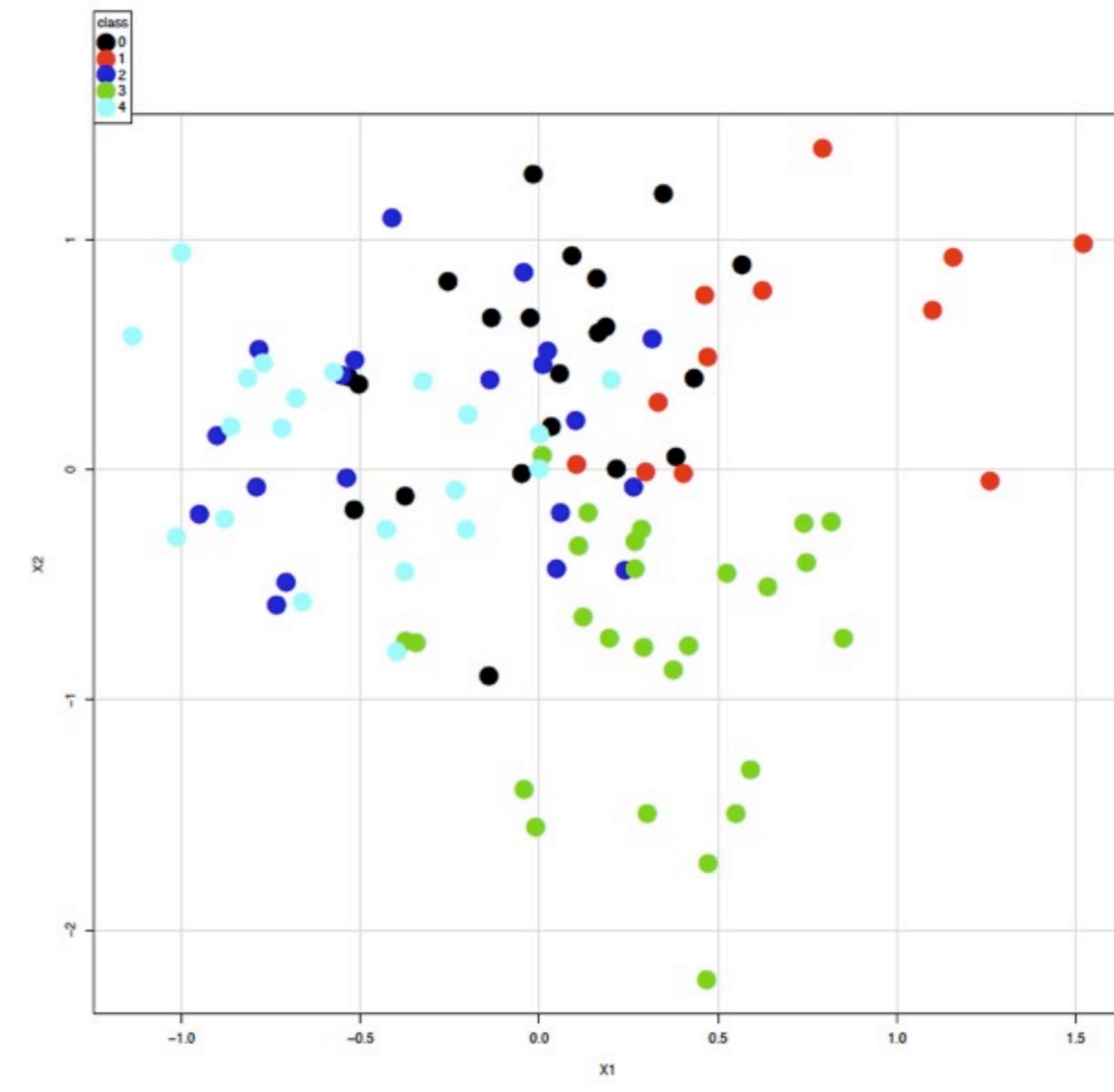
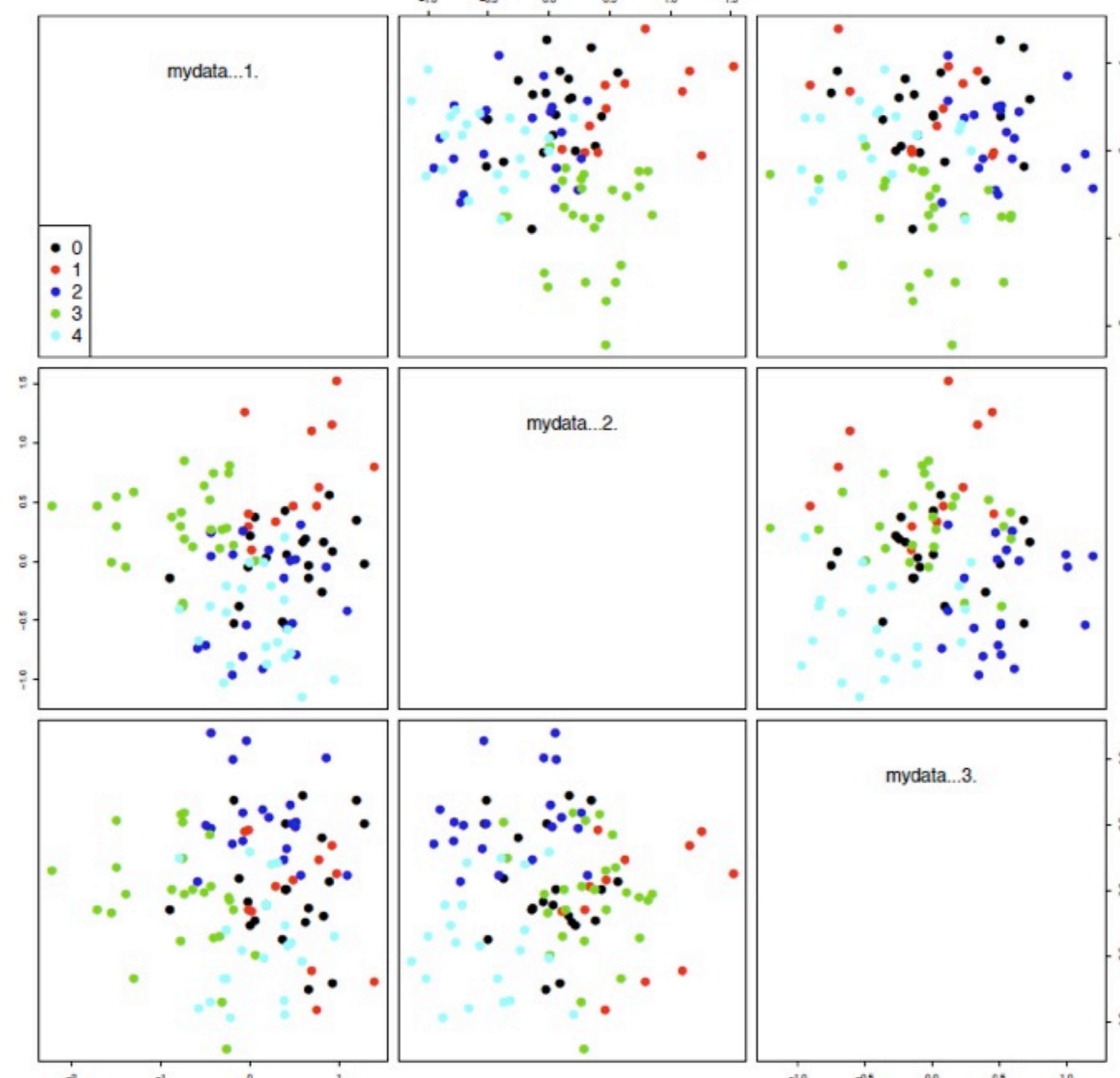
**1 1 1 1 1 1 1 1 1 1 1 1 1**

5-2x3	4-2x3	2-2x3	2-2x3	2-2x3	4-2x3	4-2x3
5-2x3	5-2x3	4-2x3	5-2x3	5-2x3	5-2x3	

# C1. Within-DR - SPLOM (not in the paper):

dataset: gauss-n100-10d-5largeCI

DR: robust PCA



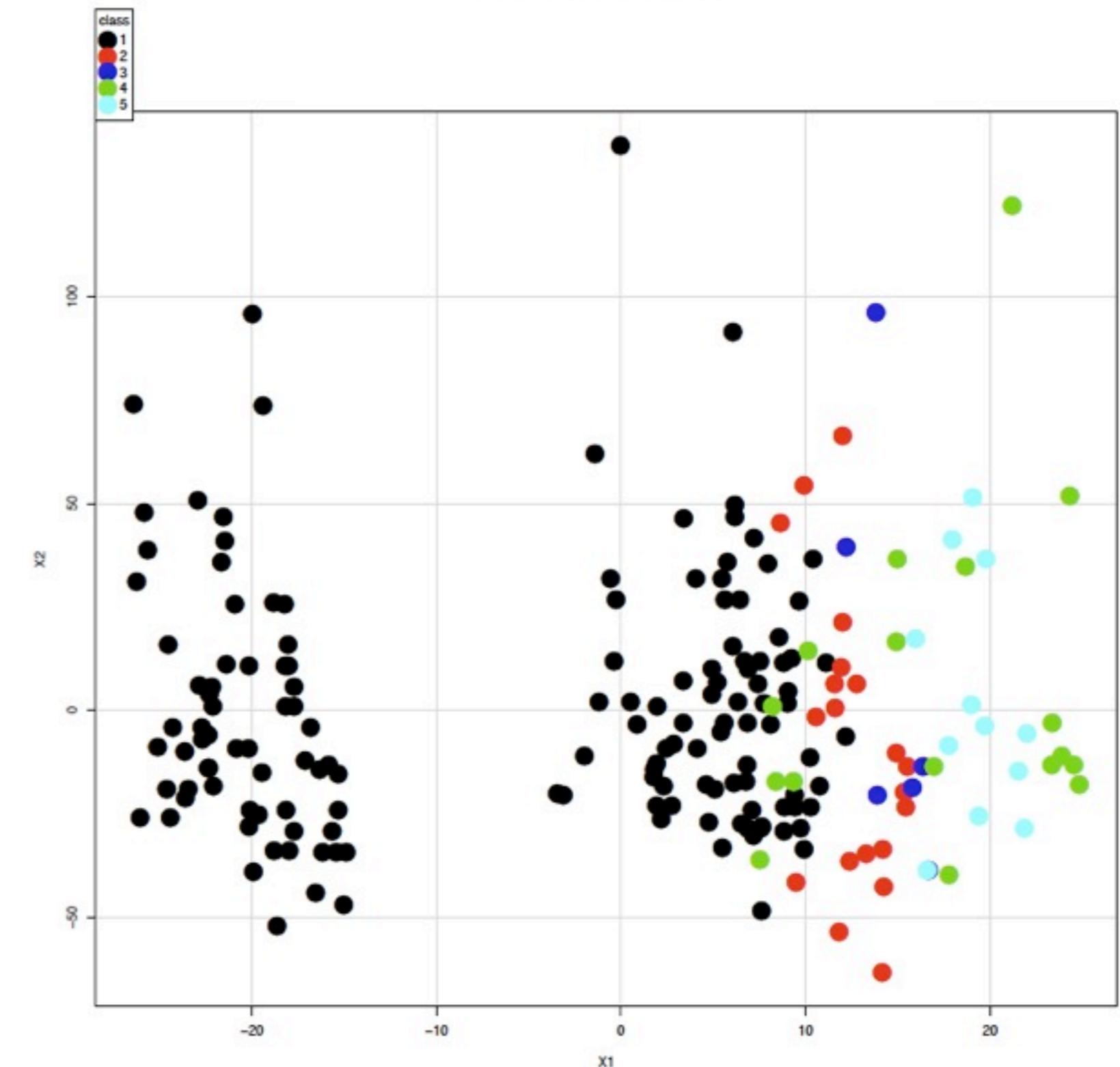
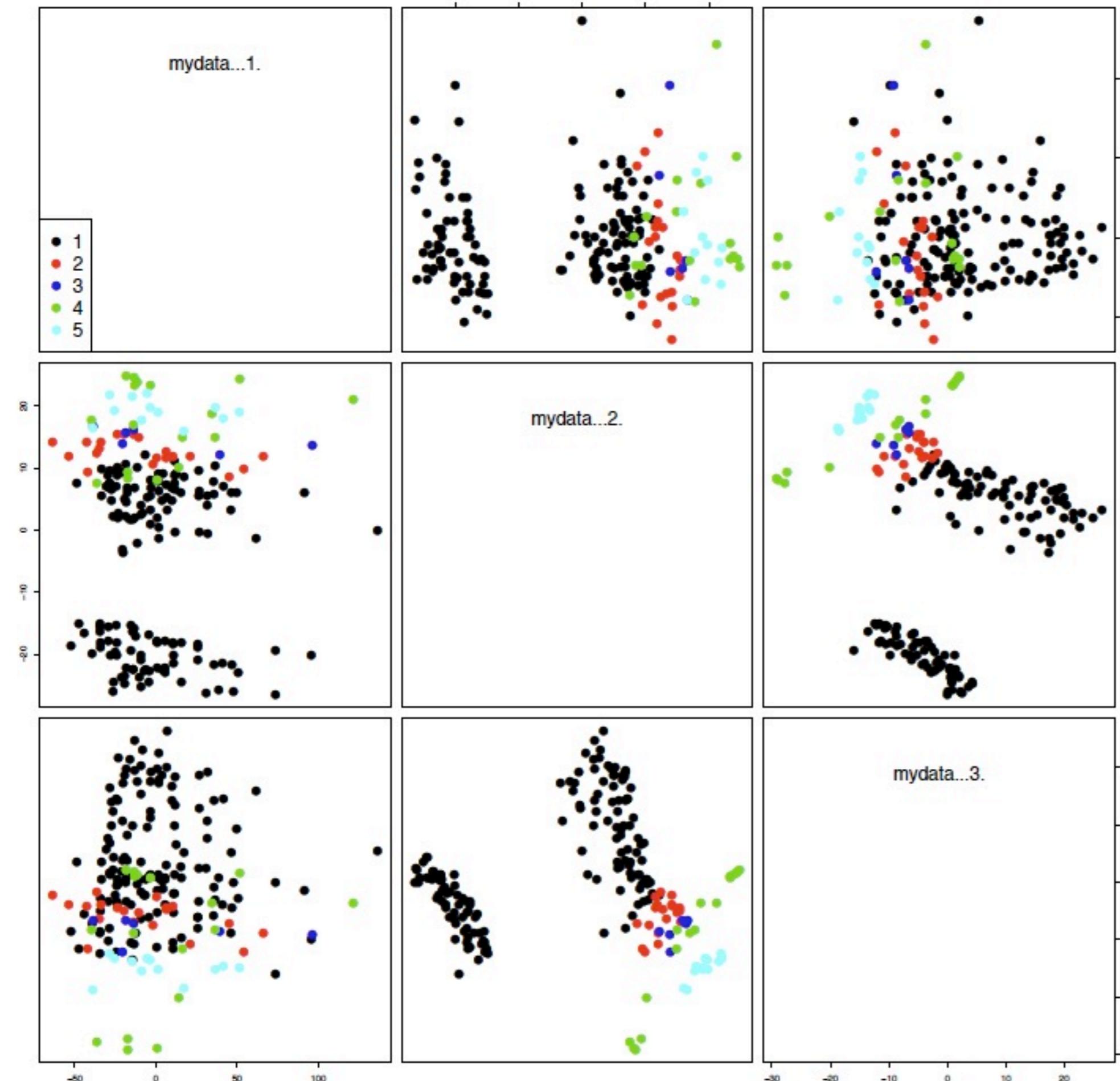
1 3 1 3 1

1 3-1x2 3-2x3 3-1x2 3-2x3

# C1. Within-DR - SPLOM (not in the paper):

dataset: bbdm13

DR: PCA



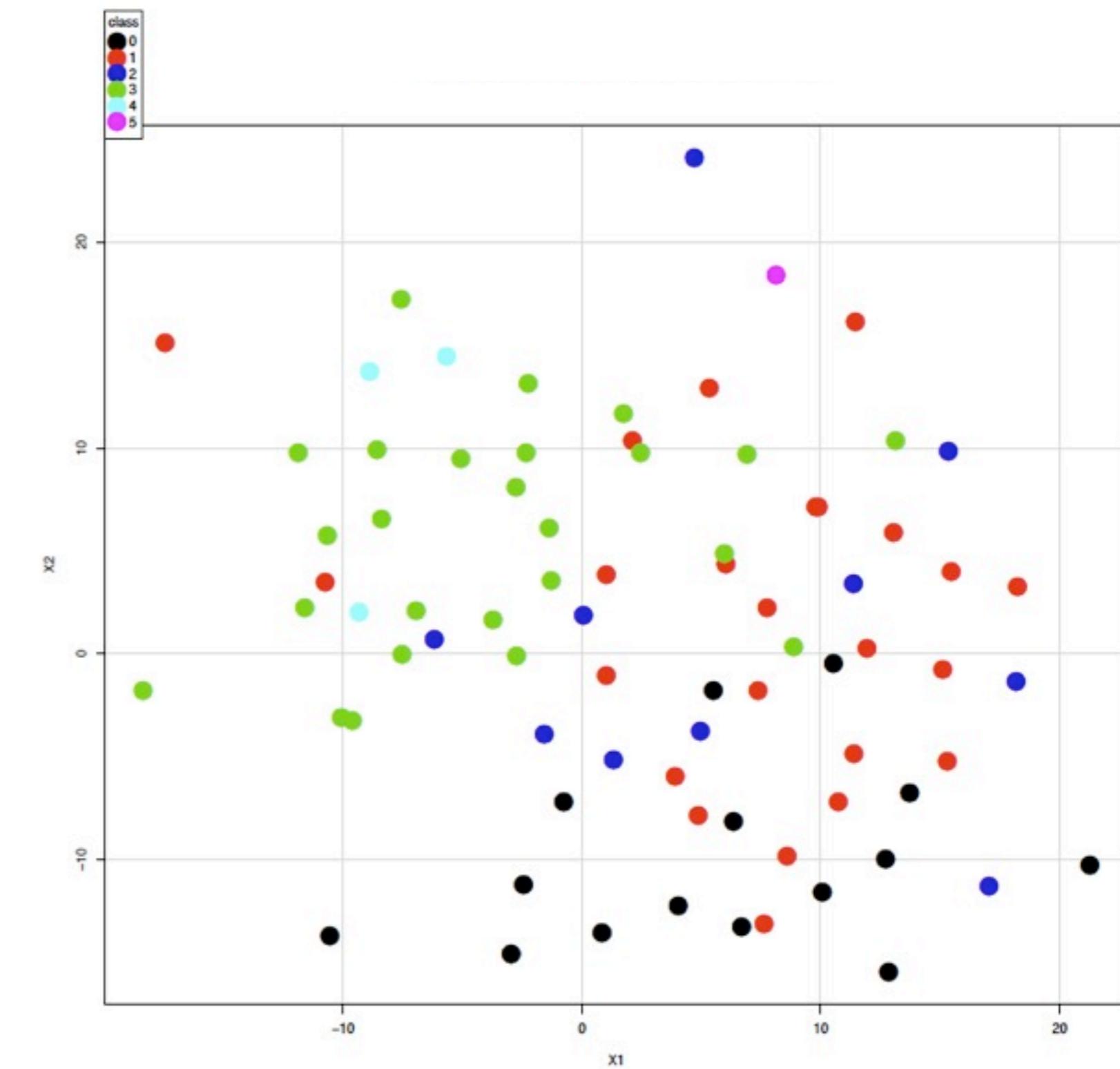
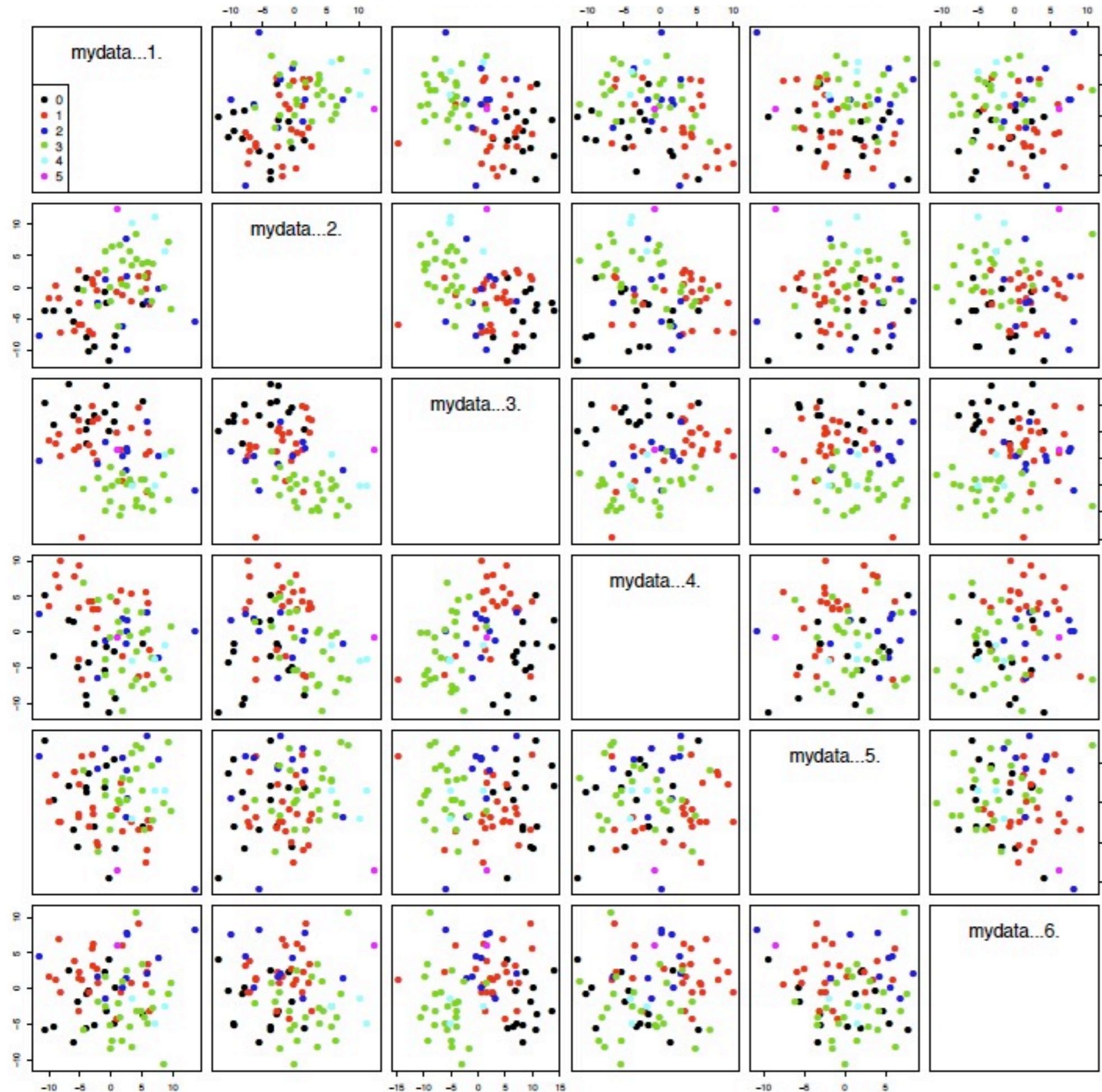
4 3 1 1 3

5-2x3 | 3-2x3 | 3-2x3 | 3-2x3 | 5-2x3

# C1. Within-DR - SPLOM (not in the paper):

dataset: hiv

DR: glimmer MDS



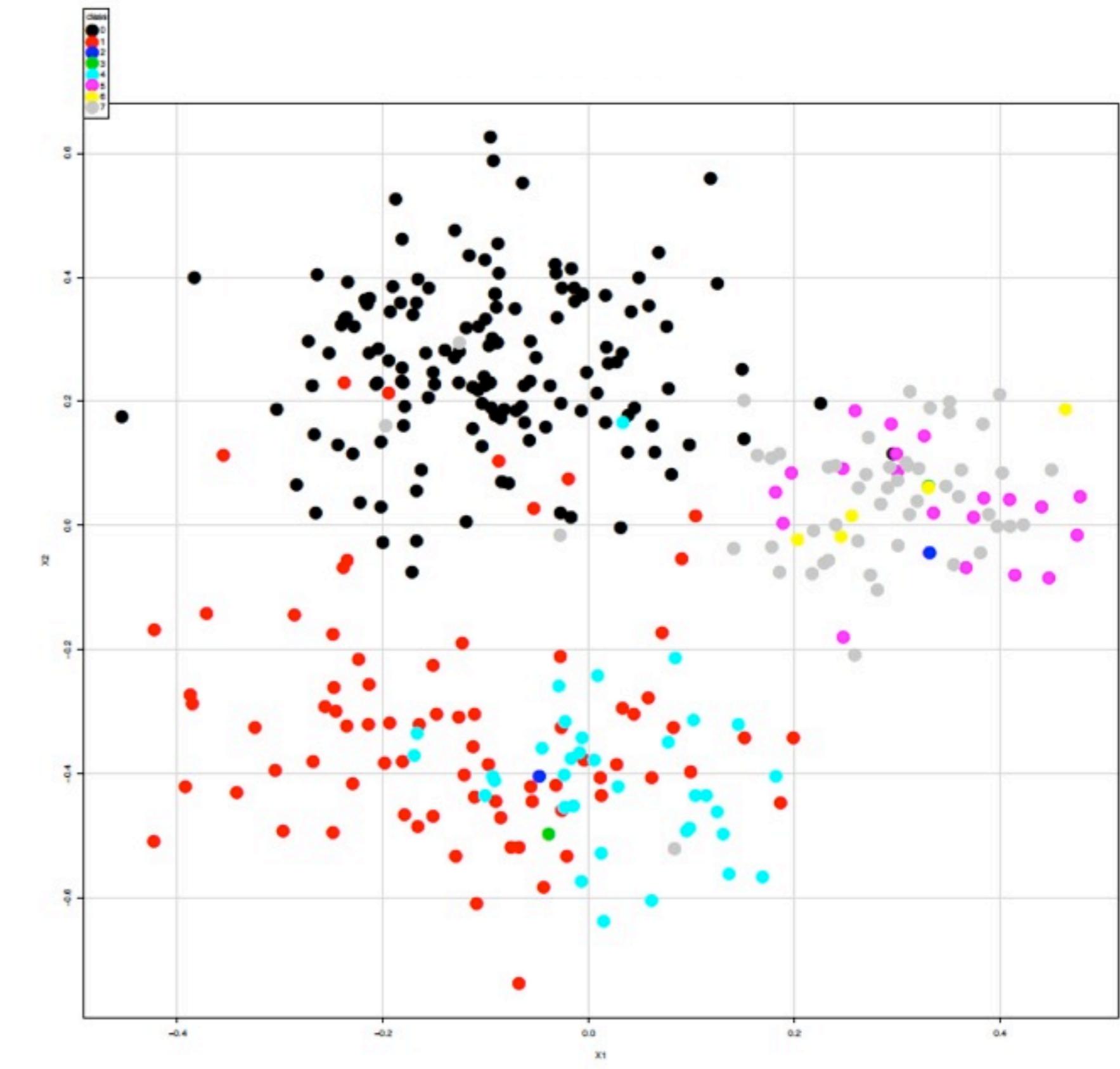
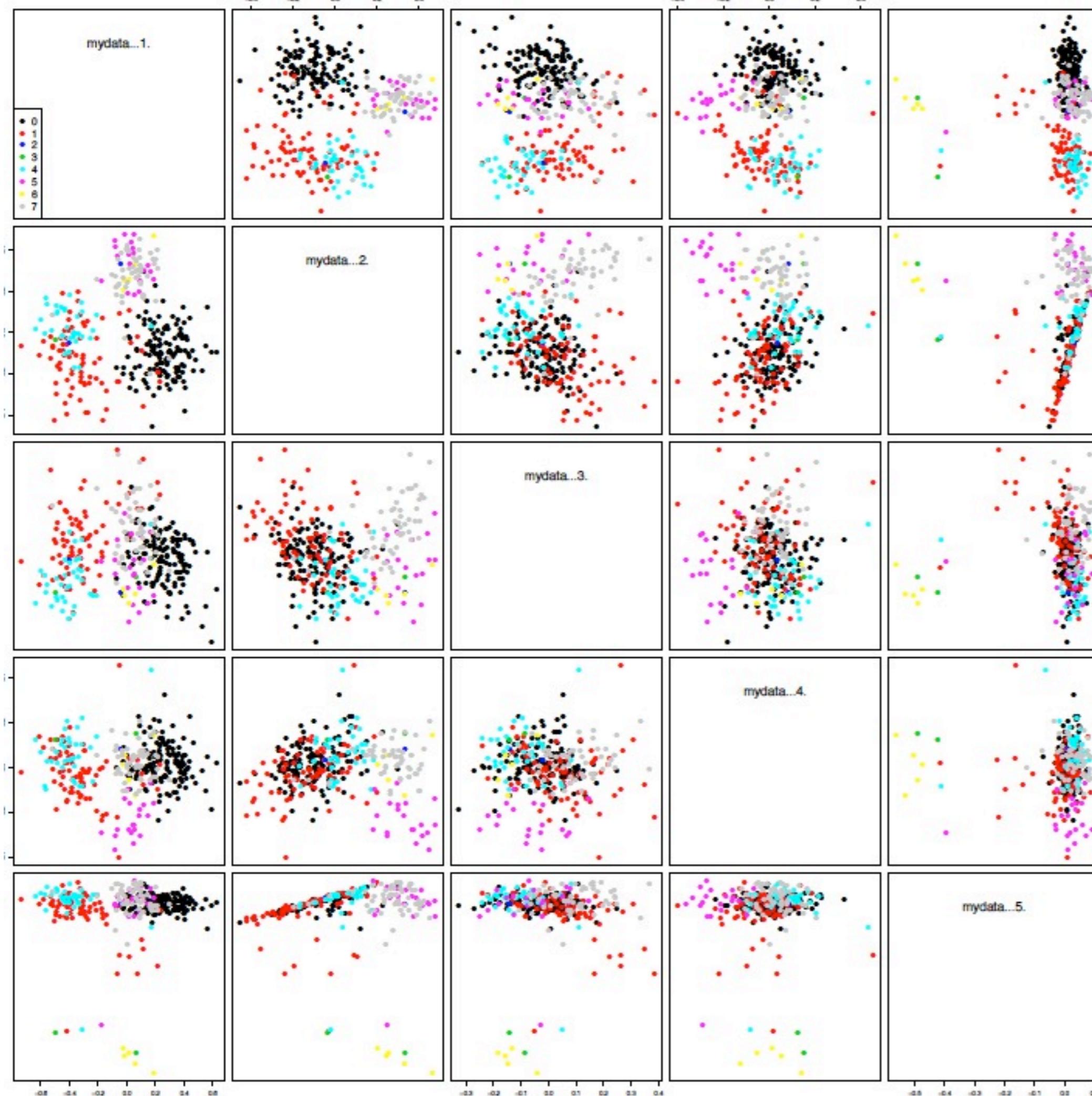
2 2 1 3 1 99

3-2x3 3-4x6 1 4-3x6 2-2x5 99

# C1. Within-DR - SPLOM (not in paper):

dataset: ecoliProteins

DR: PCA



4-1x2 3-1x2 1 5-4x5 2-1x3 5-2x4 5-4x5 3-2x3

## **C2. Within-DR: i3D > max(2D,SPLOM)**

see video

(17.7 MB, no audio, tested on VLC 1.1.12)

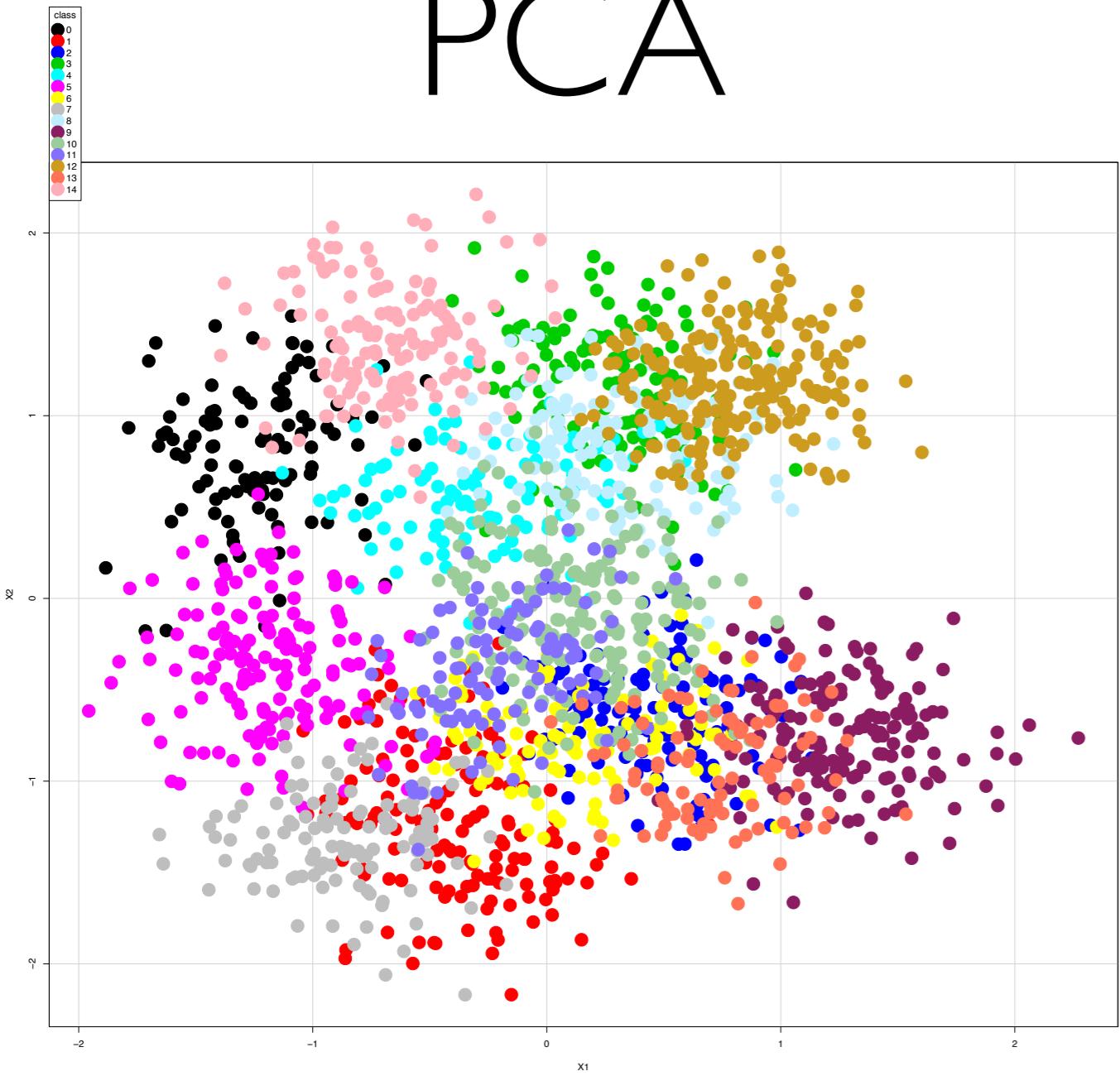
## **C3. Between-DR**

Examples where one DR provided notably or substantially better class separability than other DR techniques.

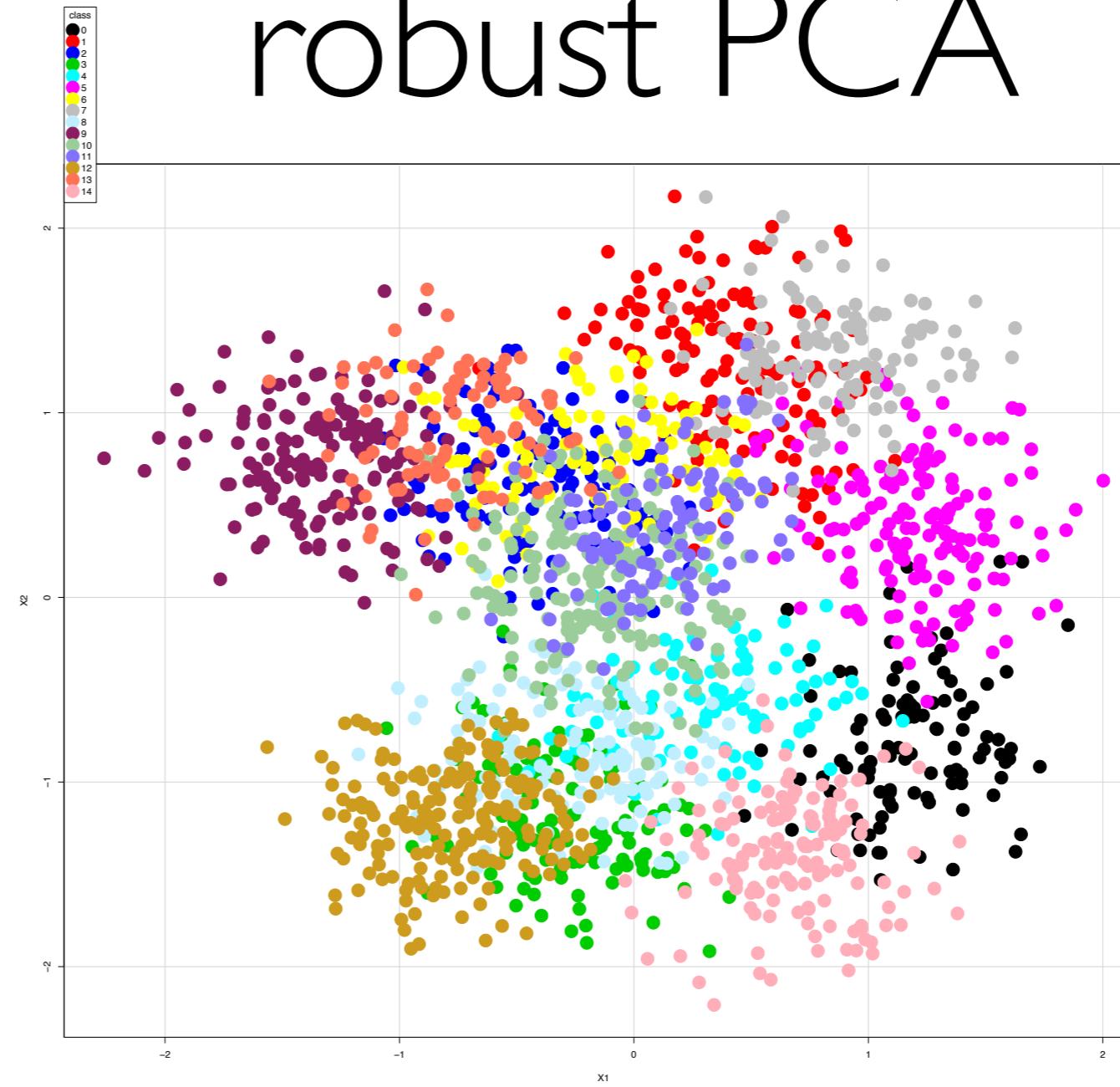
# C3. Between-DR (Figure 5 e/f from the paper):

dataset: entangled2-15d-adjacent

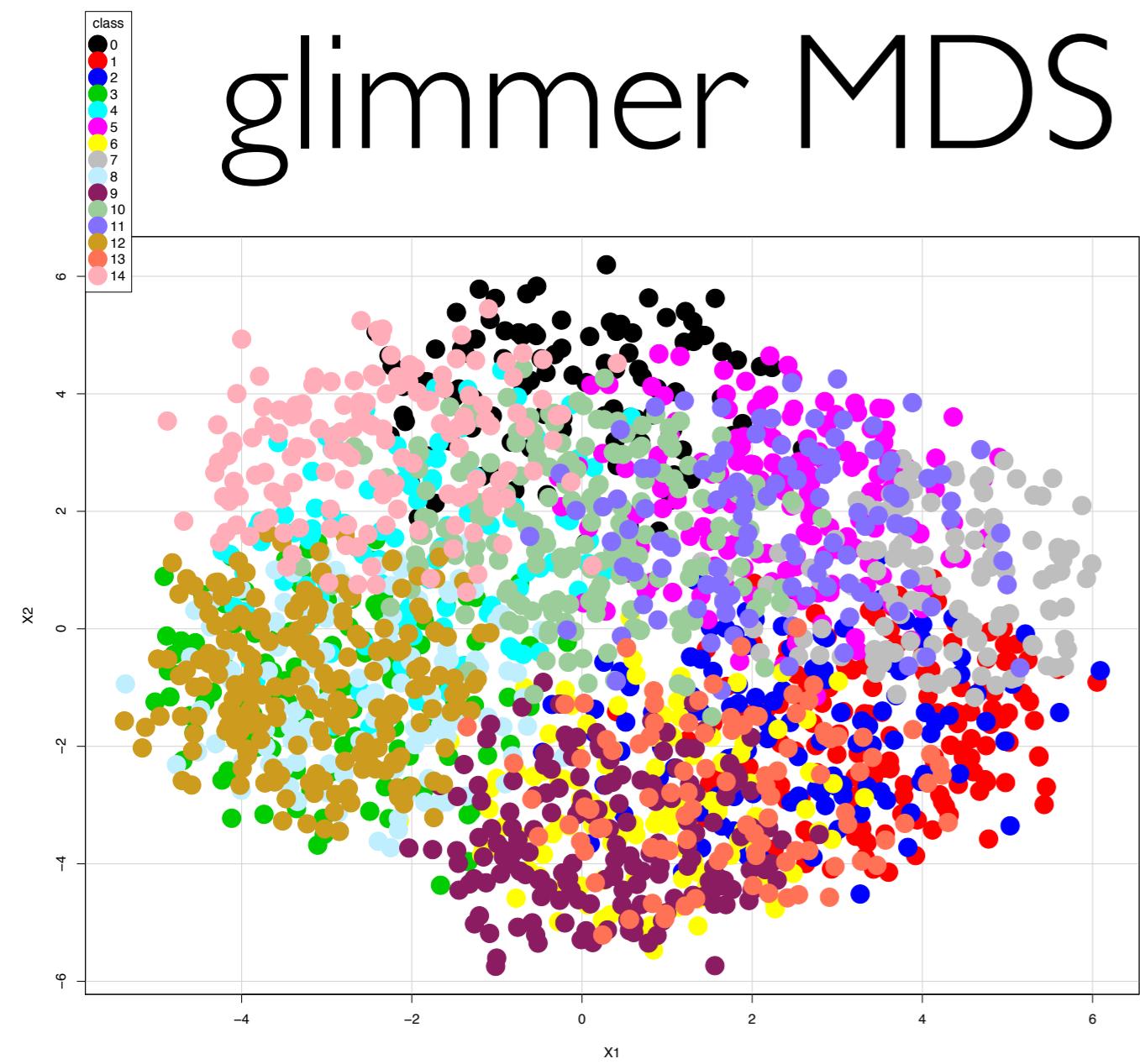
PCA



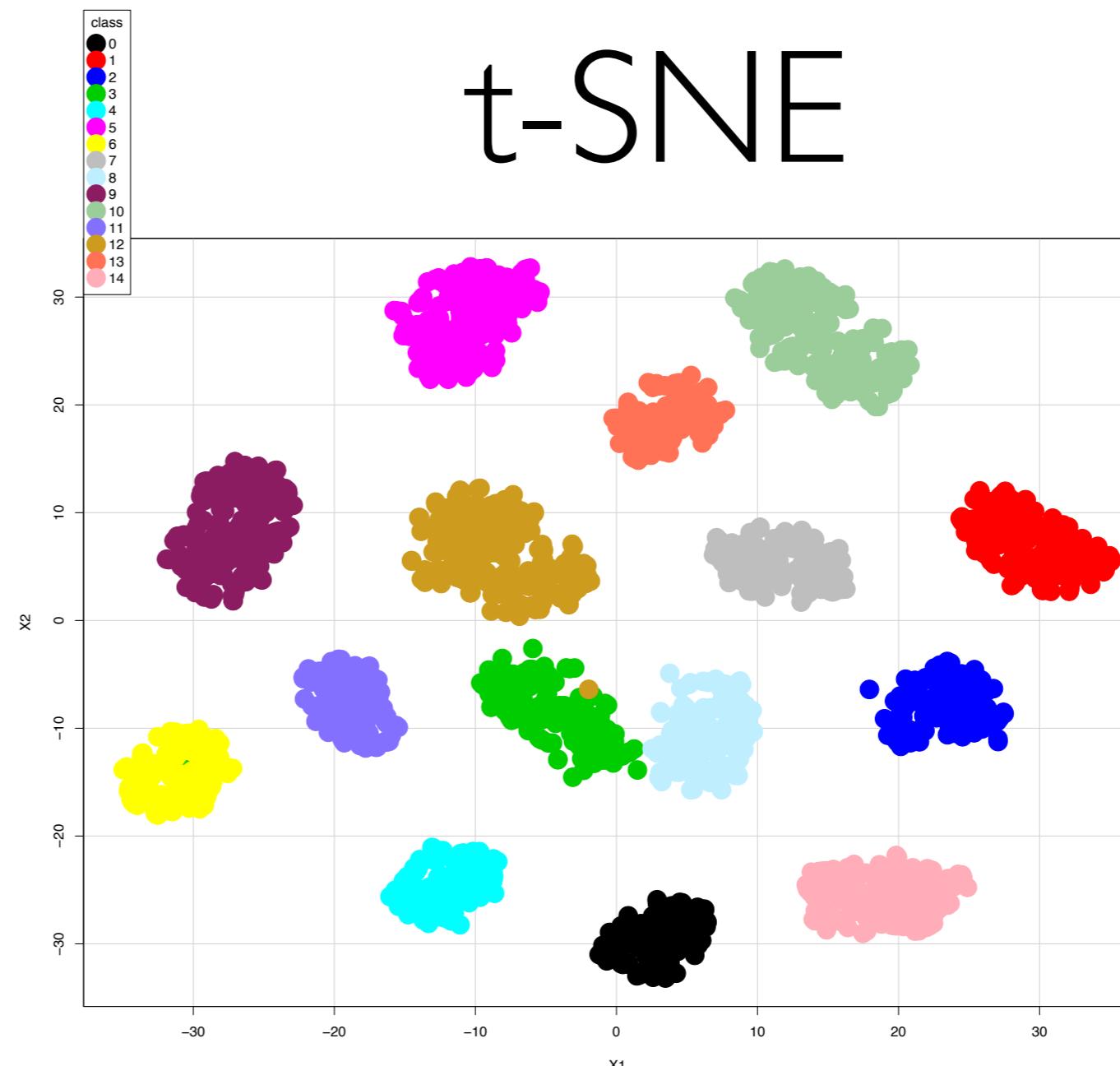
robust PCA



glimmer MDS



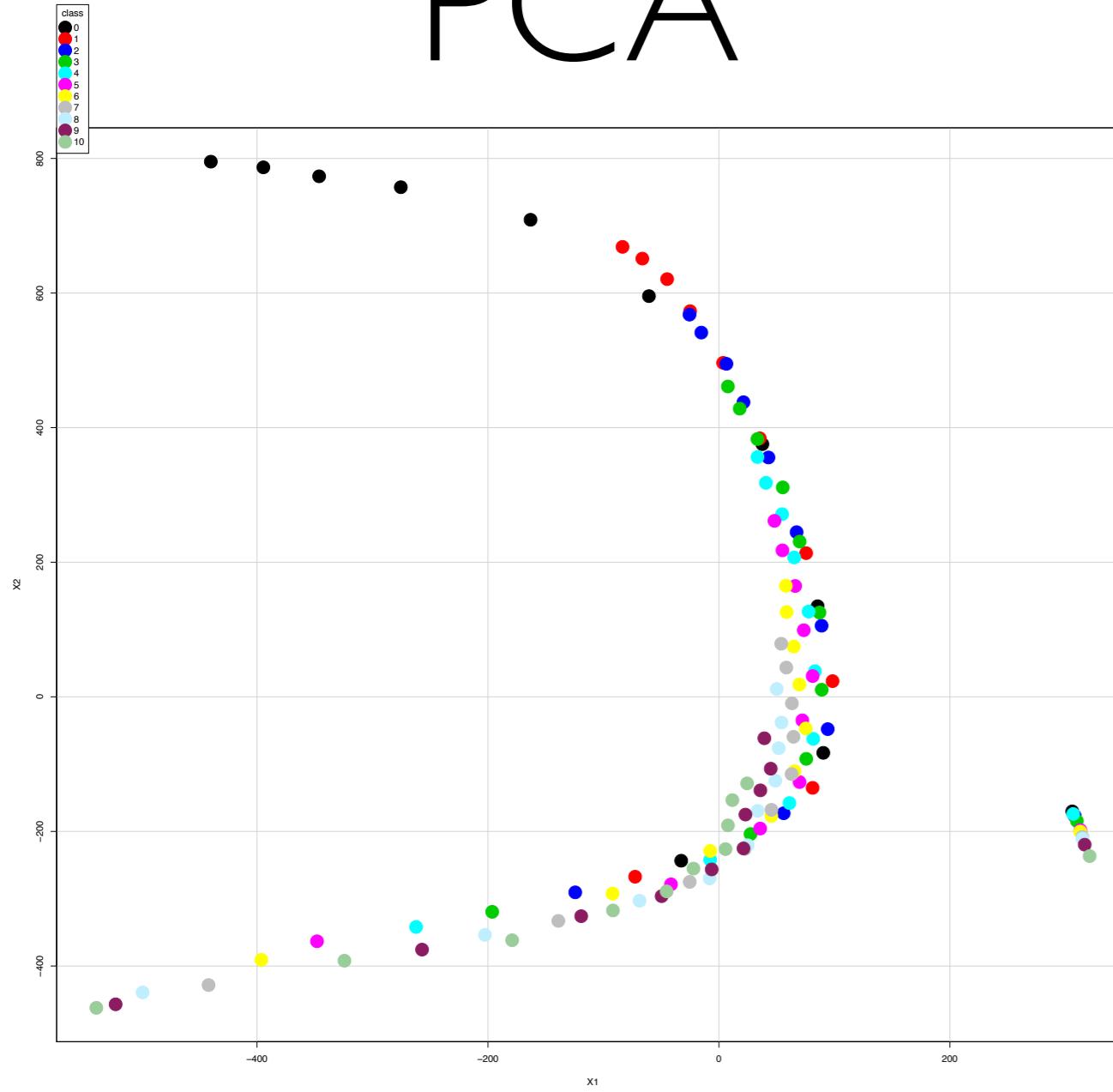
t-SNE



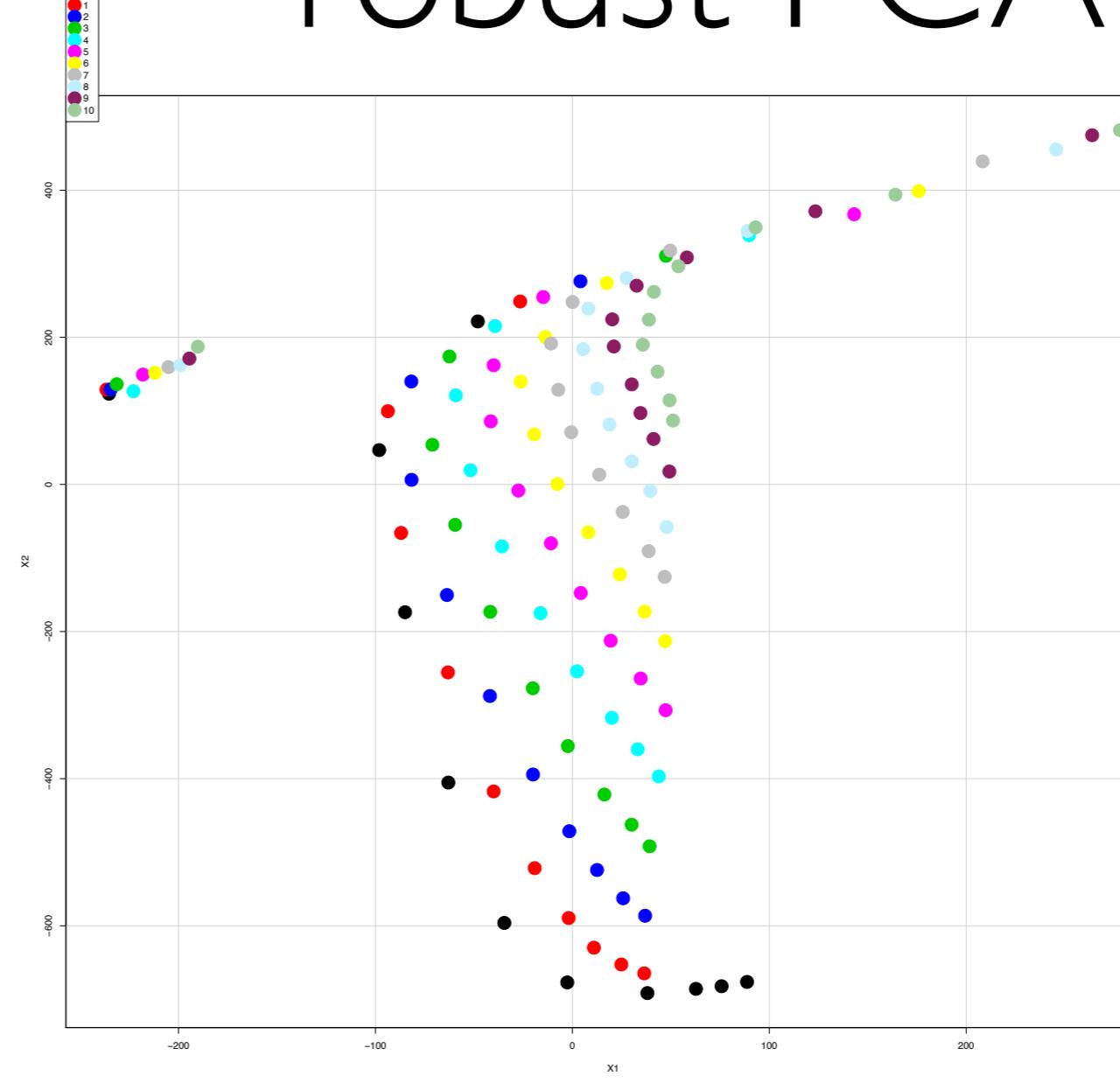
# C3. Between-DR (not in the paper):

dataset: fisheriesEscapementTarget

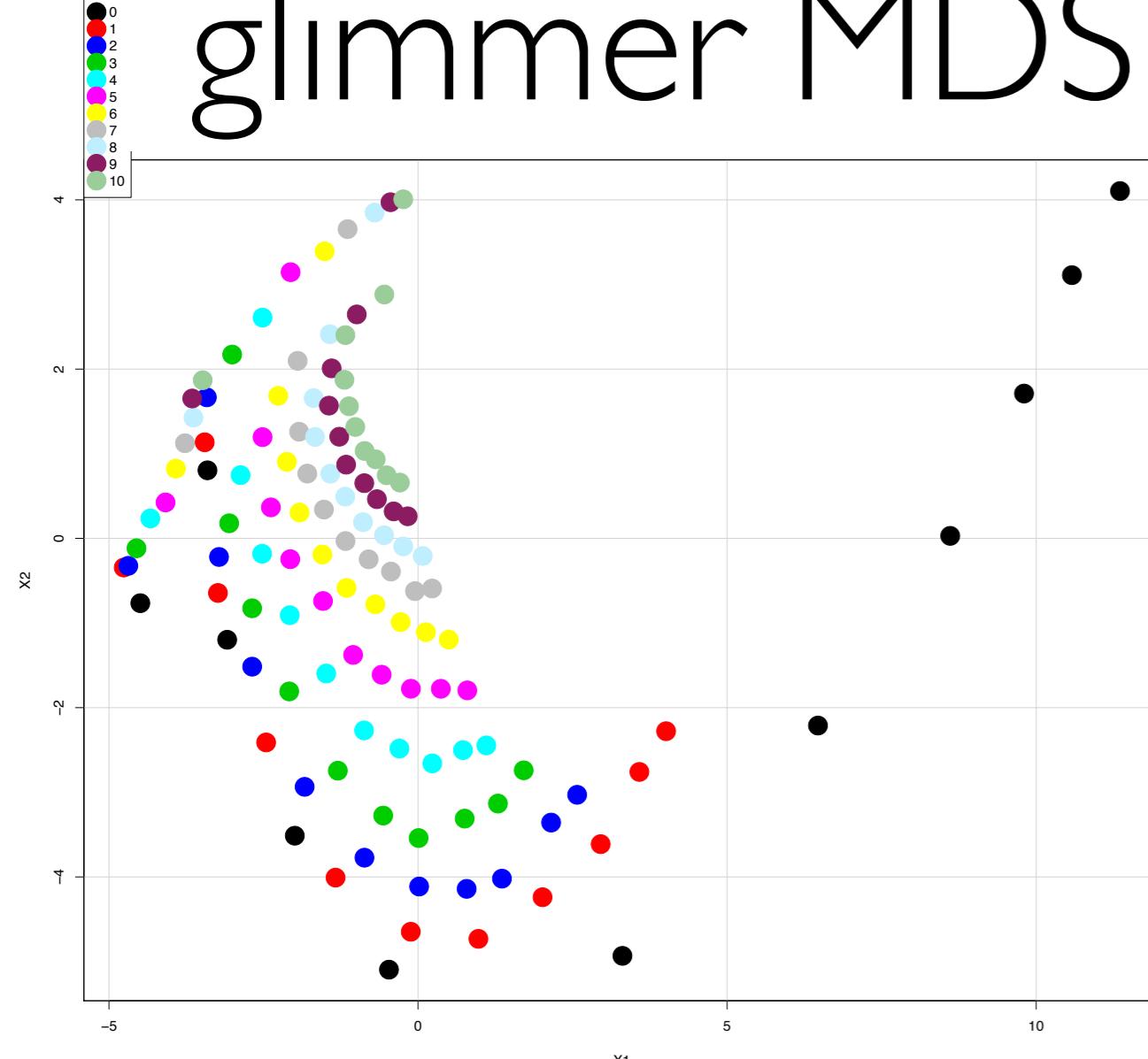
## PCA



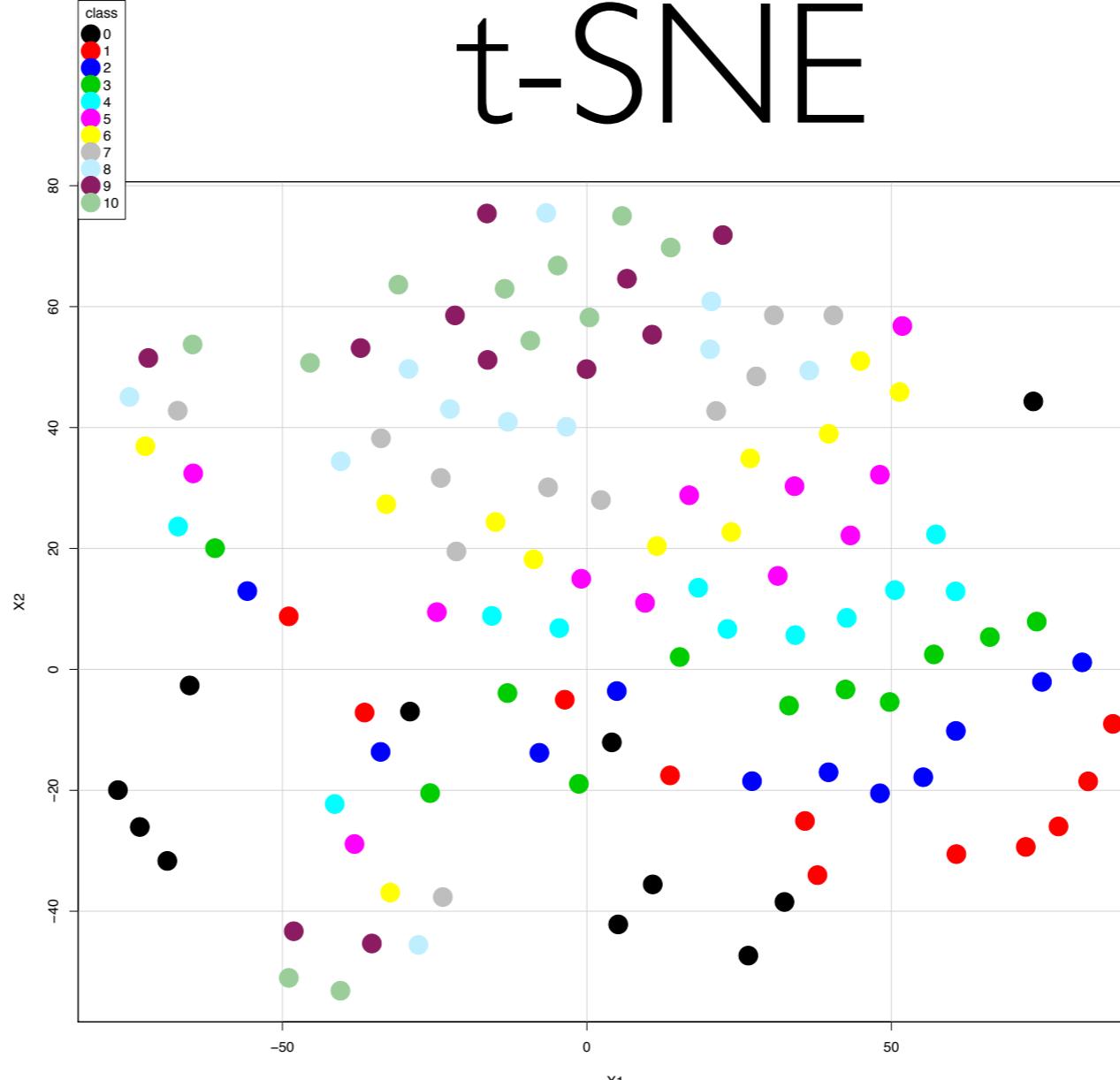
## robust PCA



## glimmer MDS



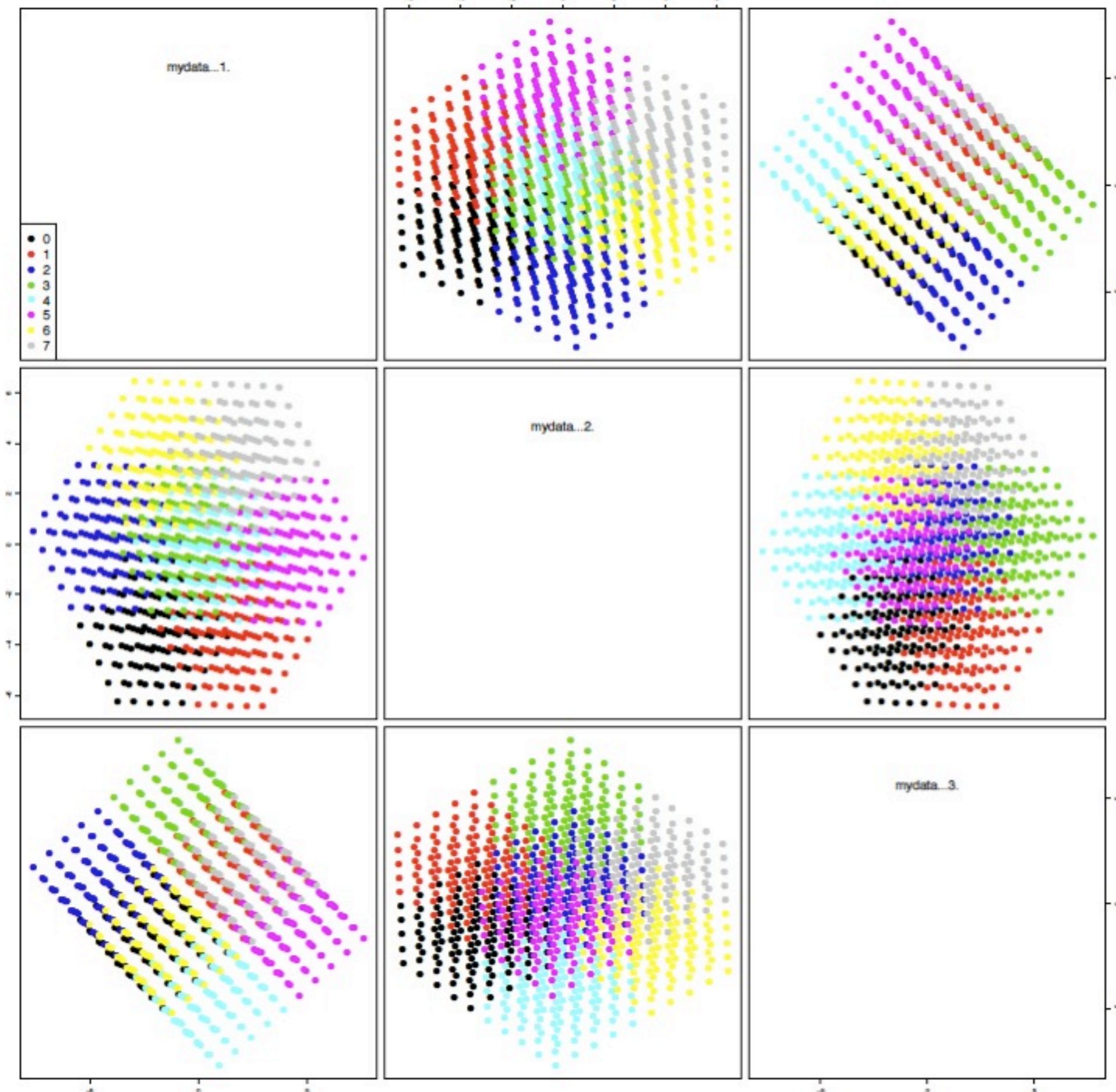
## t-SNE



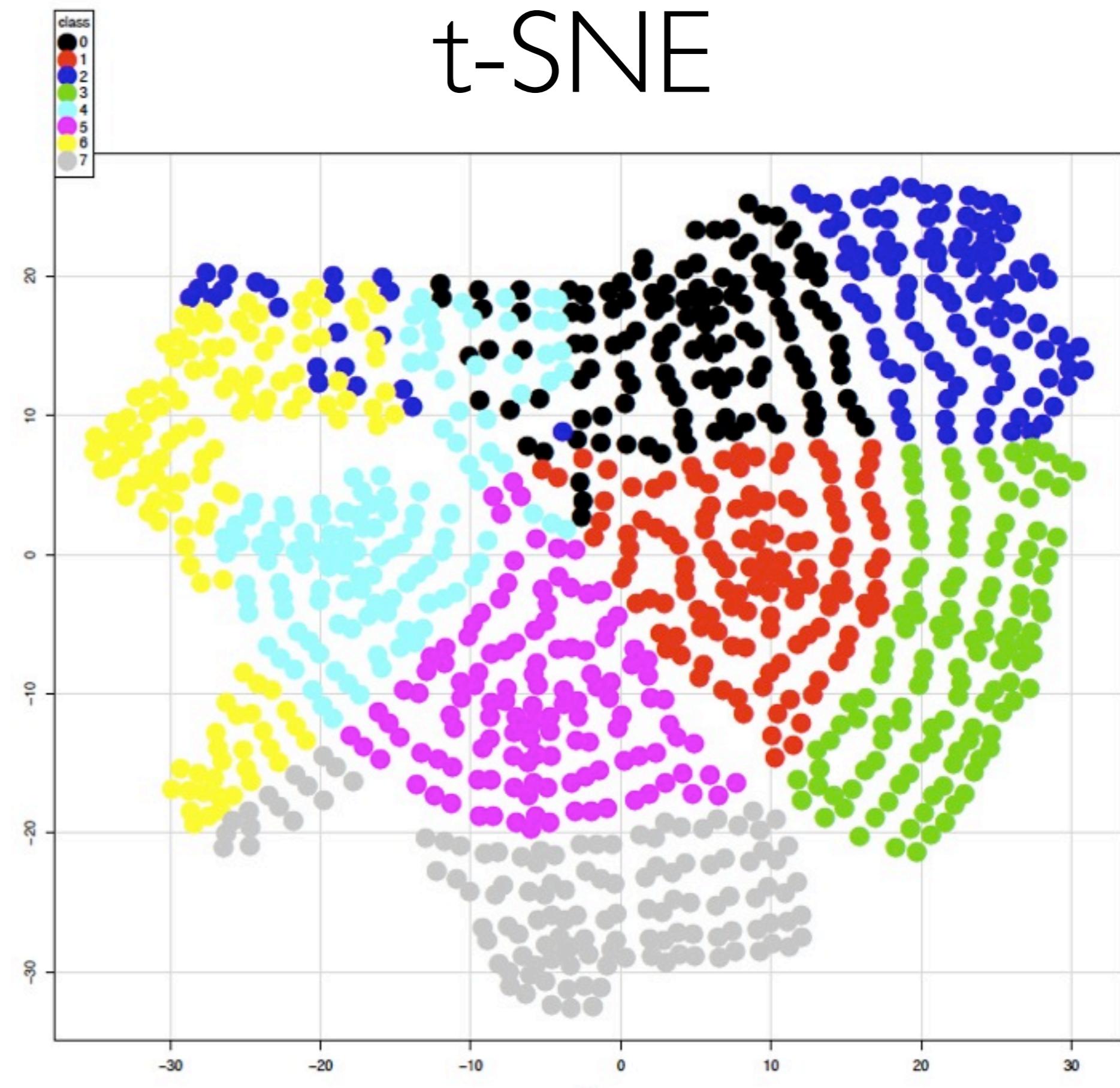
### C3. Between-DR (not in the paper):

dataset: grid10\_3d

robust PCA



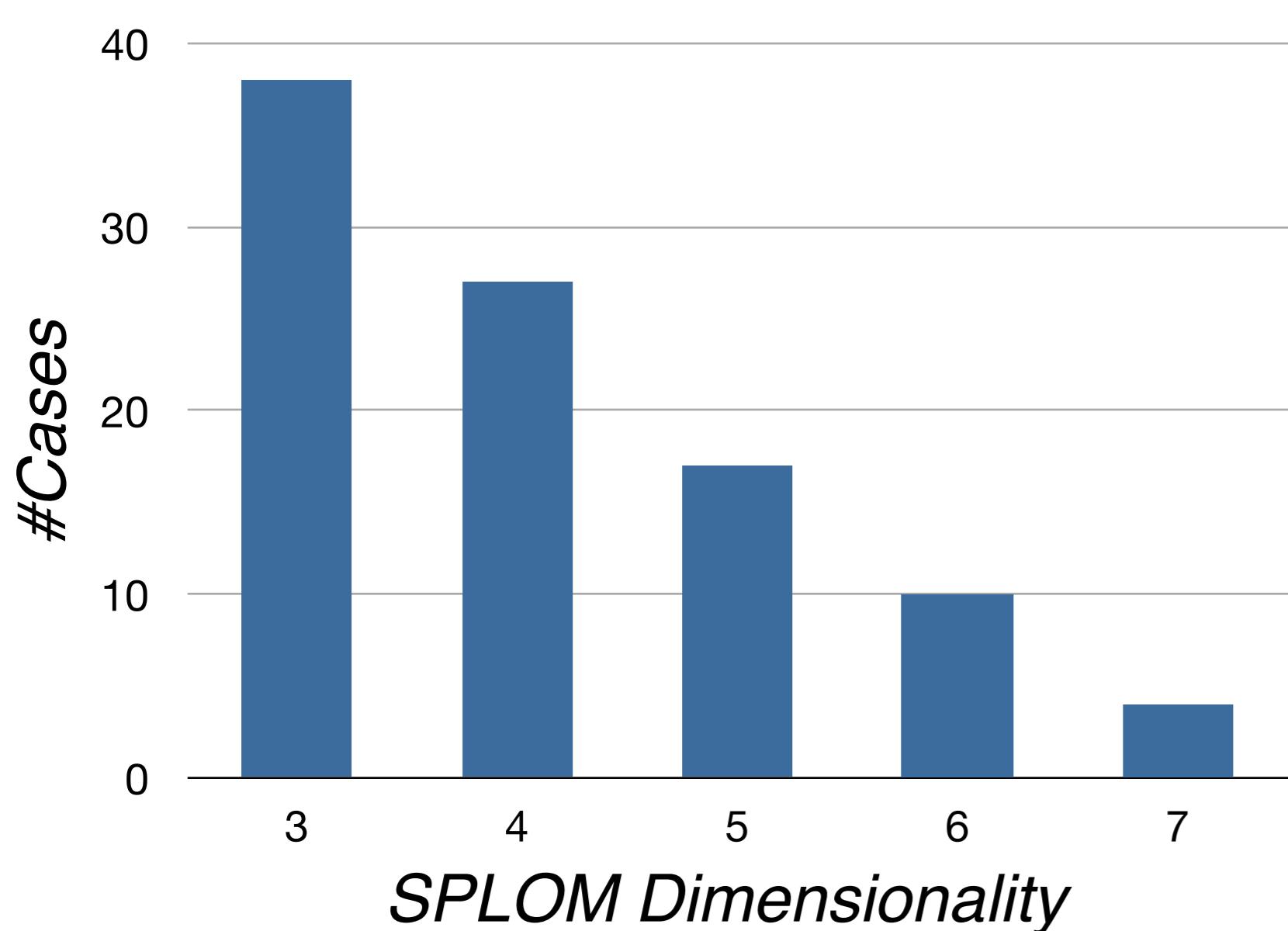
t-SNE



# **D. SPLOM Evaluation**

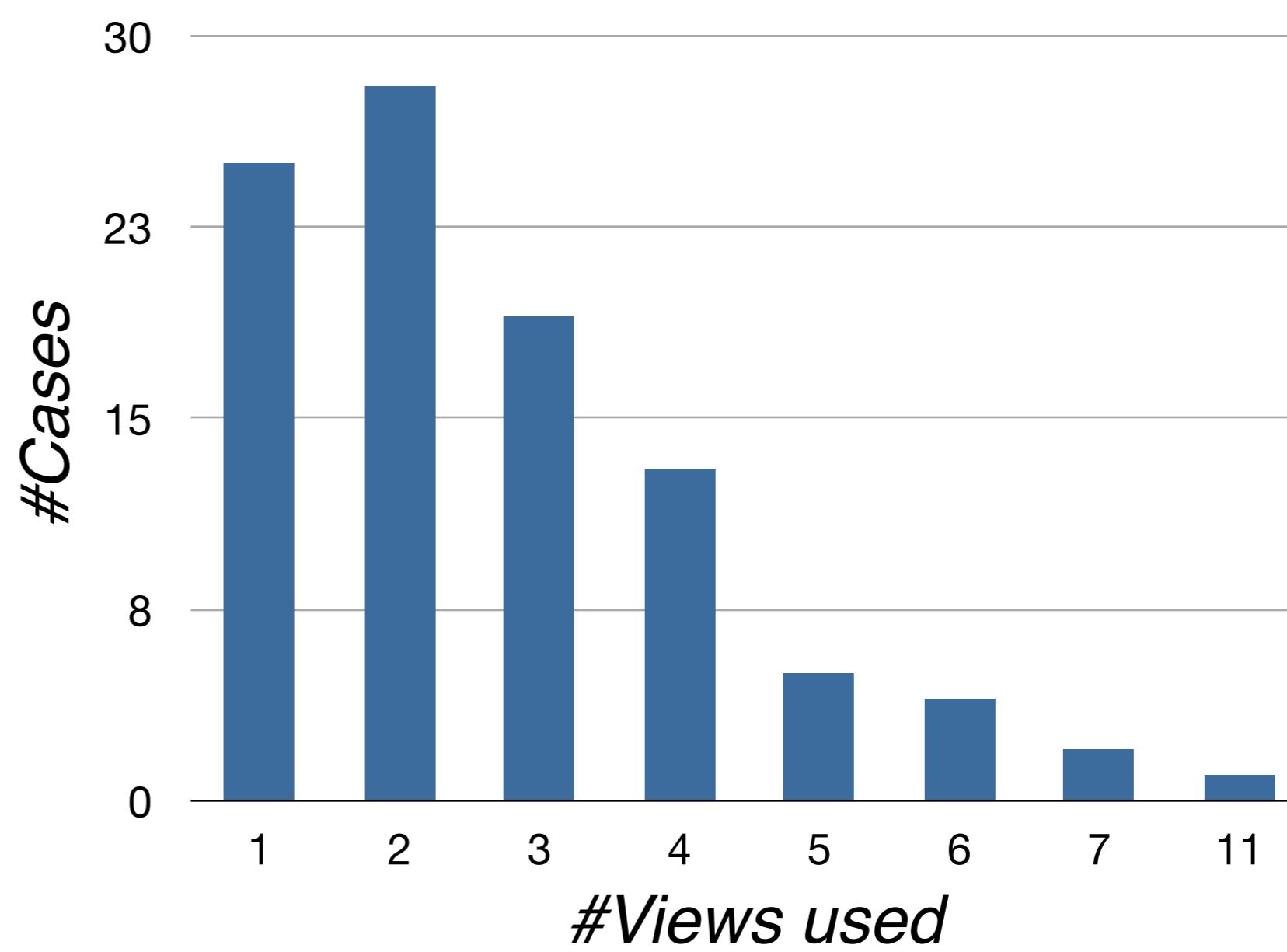
extending upon Section 5.5.  
of the paper.

Both coders kept track of the dimensionality of the SPLOMs as well as the number of views they used for the cluster verification task. We analyzed the data for which SPLOMs were marked by the coders marked as the ‘winner’ for a DR x dataset combination. In doing so, we exclude cases where a linear (PCA, robust PCA) SPLOM was equally good as the 2D version. In such cases, a single component of the SPLOM would be enough, and so this case should not be represented as a true count of how SPLOMs are used because it would bias towards lower dimensionality and fewer views.



**Figure 1:** Dimensionality of SPLOMS used (max=7, mean=4.1);

**Figure 1** shows the dimensionality of the winners: there is a linear decrease in dimensionality, with a maximum of 7 and a mean of 4.1. **Figure 2** shows the number of views used within a SPLOM. In 88% of the cases, between 1 and 4 views revealed the necessary class structure. The average number of used views is only 2.7, and the maximum is 11 views. These findings indicate that when using SPLOMs for cluster verification in DR data their size is limited and thus they are indeed usable. In particular, these findings show that our usability cost assumptions are not violated.



**Figure 2:** The numbers of views used for cluster verification (max=11, mean=2.7).

# **E. Coding Guidelines**

The following table summarizes guidelines that both coders used for judging class separation.

Acronym	Description	Guideline
	Goodness Factors	
no-overlap	Class has no overlap with any other class.	
connected	Points of a class are visually connected to each other. In particular, they should not be interrupted by points of another class.	Good classes, ie 5, are classes that have no overlap and which points are connected to each other. (Use concept of convex hull as rough estimator for overlap judgments)
	Badness Factors	
full-mixture	Cluster is heavily intermixed and fully overlapped with one or more other ones. No visible class structure is visible.	Those classes should be rated as bad, ie 1.
partial-mixture	Cluster is partially overlapped with one or more other clusters.	According to the degree of partial overlap, classes should be judged between 4 and 1.
disconnected	Points of a cluster appear disconnected.	According to the degree of disconnectedness, classes should be judged between 4 and 1.
multiple-outliers	Multiple outliers exist.	While single and obvious outliers should not have a negative impact on class separability judgements, multiple outliers can decrease the score of a class.
	Robustness Factors	
adjacent	Adjacent classes: no physical distance between classes.	As long as there is no overlap between classes, adjacent, connected classes should be judged as good, ie 5.
equidistant-mixed	Dataset with partly or fully overlapped classes and equidistant point structure.	Equidistant point structures should not lead to any other negative impact as compared to randomly distributed class structures.
single-outlier	A single outlier exists.	Single and obvious outliers should not have a negative impact on class separability judgements. Whether a single outlier is obvious also depends on the number of points of that class.
shape	The shape of a cluster is important for its detection.	The shape of a cluster should not negatively impact class separation judges.
sparse	Data of a class is sparsely distributed in view.	Different densities should not negatively impact class separation judges .
unbalanced-classes	Classes differ strongly in no. of points / class.	Differences of no. of points should not negatively impact class separation judges.
periphery	Clusters at the periphery might be easier to spot.	All visual encodings should be carefully checked for such phenomena in order not to incorrectly judge class separation.
z-depth	Z-depth might influence your decision.	
3D-move	Movement of points in 3D helps to detect a cluster (Gestalt law: Common fate).	
3D-bg-noise	The background noise in 3D makes detectability of clusters harder.	i3D should be carefully checked for such phenomena in order not to incorrectly judge class separation.
3D-inner-class	Inner clusters usually okay in 2D and SPLOM but not in 3D.	
	Specific Class Scores	
1-point-class	Class with only 1 point.	1-point classes get a special label, they can not be used in the study as the separation of a 1-point class is not meaningful. Mark with "99".
no-DR-computation	DR embedding couldn't be computed for this dataset.	Mark all classes with "NA".