Multivariate exploratory data analysis of kink role affinity scores

Dr. Julie C. La Corte
Georgia State University

First draft: Jul. 15, 2020
Last revised: Nov. 24, 2020

Overview

The dataset

Variables

- Each variable measures affinity for a different role
  - Range of each variable: 0 to 100 points (given as percentage)
  - 20 of the 25 variables are paired (e.g. *Domi*/*Subm*)

Agep	Ageplayer	*Maso*	Masochist
BBot	Bondage Bottom	*MaMi*	Master/Mistress
BTop	Bondage Top	*NoMo*	Non-Monogamist
BoGi	Boy/Girl	*Ownr*	Owner
Brat	Brat	*Pet*	Pet
BrTa	Brat Tamer	*Prey*	Prey (Primal)
DaMo	Daddy/Mommy	*Sadi*	Sadist
Dgee	Degradee	*Slav*	Slave
Dger	Degrader	*Subm*	Submissive
Domi	Dominant	*Swit*	Switch
Exhi	Exhibitionist	*Vani*	Vanilla
Expe	Experimentalist	*Voye*	Voyeur
Hunt	Hunter (Primal)		

Overview

Three stages

Preliminary stages

- **Stage 1:** One-variable EDA
  - Classify distributions by shape
  - Normalize each variable
- **Stage 2:** Two-variable EDA
  - Test assumptions for classical linear regression
  - Cluster variables by correlation

Overview

Three stages

Stage 3: Multivariate analysis (clustering)

- What's the "best" choice of algorithm parameters?
  - How do we quantify a clustering's stability?
  - How do we measure intracluster consistency?
- Characterize the clusters
  - Are the clusters significantly statistically different?

Ultimate goal: Create a classification of individuals based on empirical data, not theoretical assumptions.

Stage 1: One-variable EDA

How do the univariate distributions compare?

Apples to apples...

- Do the distributions vary in shape, or just in location?
- Can we homogenize the distributions?

Stage 1: One-variable EDA

Classify distributions by shape

Center and spread

- Means, medians, and IQRs vary widely
- Standard deviations are all similar

score	Q_1	M	Q_3	$\bar{x}$	s
Ageplayer	.09	.26	.56	.34	.28
Bondage Bottom	.39	.74	.93	.64	.32
Bondage Top	.16	.51	.79	.49	.32
Boy/Girl	.11	.29	.60	.37	.30
Brat	.24	.47	.69	.47	.27
Brat Tamer	.09	.30	.58	.35	.28
Daddy/Mommy	.08	.23	.49	.31	.27
Degradee	.05	.27	.69	.37	.34
Degrader	.03	.16	.50	.29	.30
Dominant	.25	.57	.80	.52	.32
Exhibitionist	.15	.40	.70	.43	.30
Experimentalist	.46	.68	.83	.63	.25
Hunter	.07	.25	.59	.34	.30
Masochist	.23	.52	.76	.50	.30
Master/Mistress	.13	.35	.65	.40	.30
Non-Monogamist	.13	.36	.66	.40	.30
Owner	.04	.19	.49	.29	.29
Pet	.05	.14	.50	.29	.31
Prey	.12	.36	.65	.40	.30
Sadist	.10	.30	.62	.37	.30
Slave	.10	.31	.61	.37	.30
Submissive	.54	.79	.93	.69	.29
Switch	.31	.63	.87	.57	.32
Vanilla	.29	.49	.68	.49	.24
Voyeur	.16	.50	.78	.48	.32

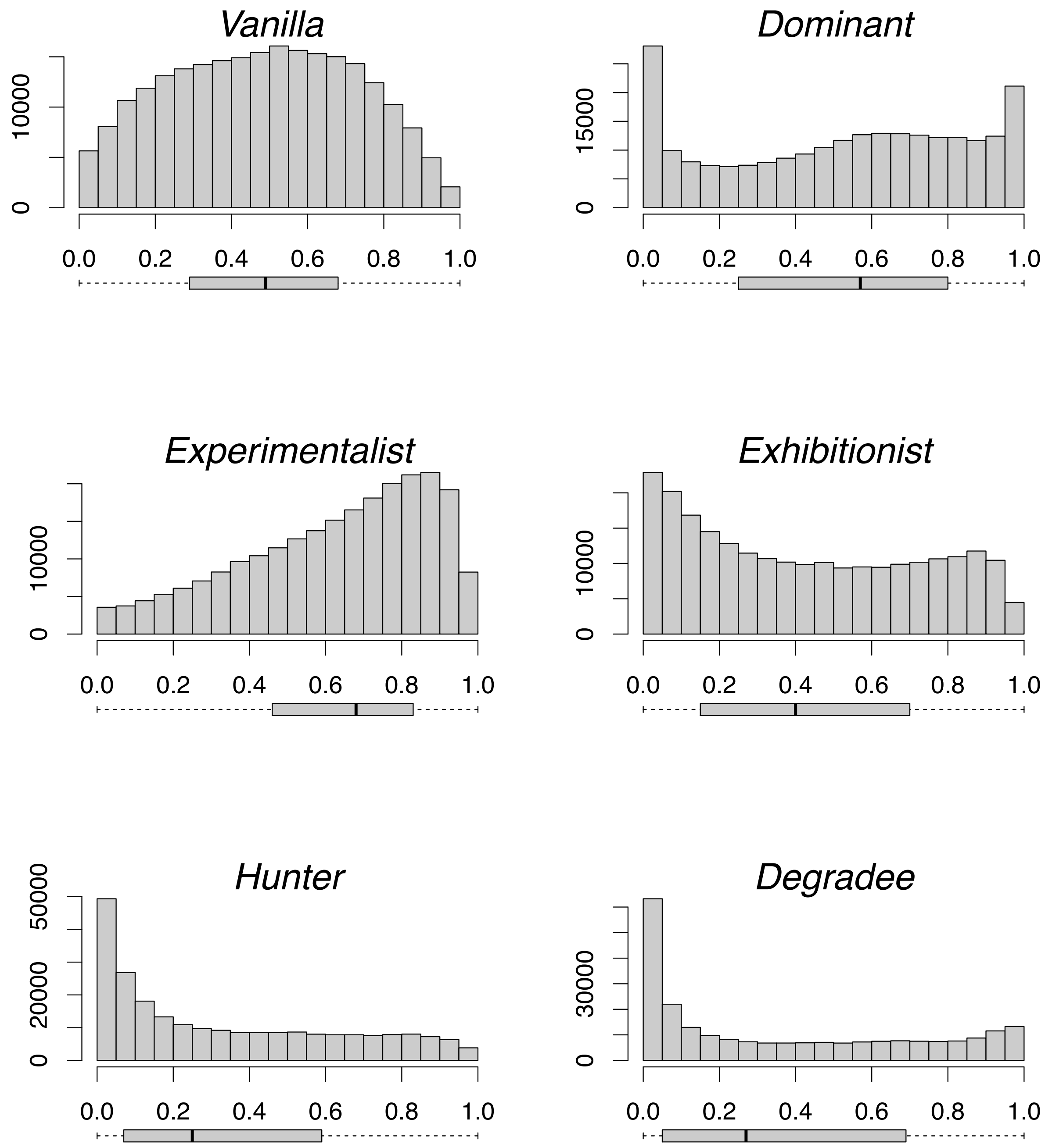
Five-number summaries

Stage 1: One-variable EDA

Classify distributions by shape

Shape

- No normal curves
  - Bounded domains
  - Flat-tailed ($Kurt < 3$)
  - One, two, or three peaks
- Some mildly random, some wildly random



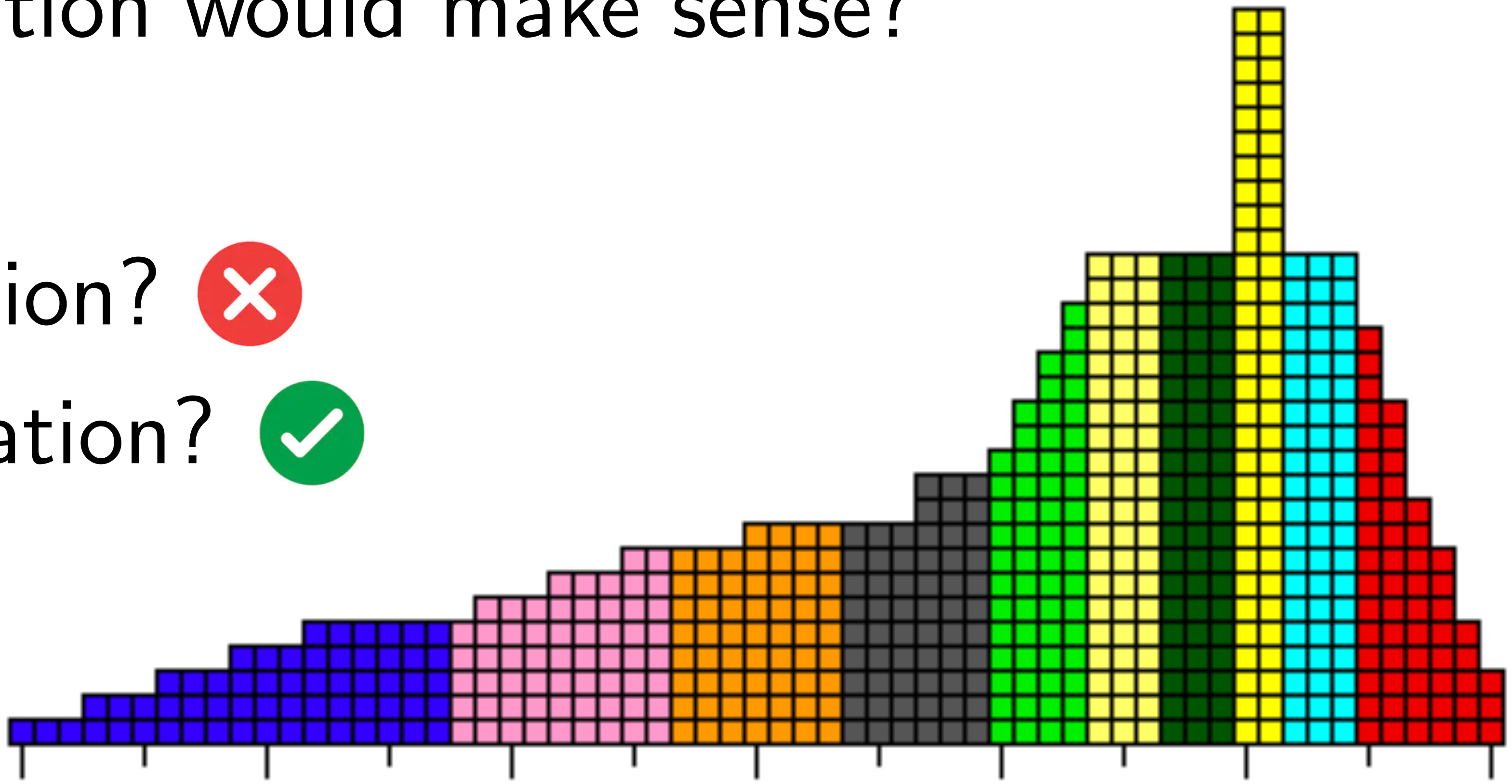
Stage 1: One-variable EDA

Normalize each variable

Normalization

- No single family of standard parametric distributions describes all 25 variables

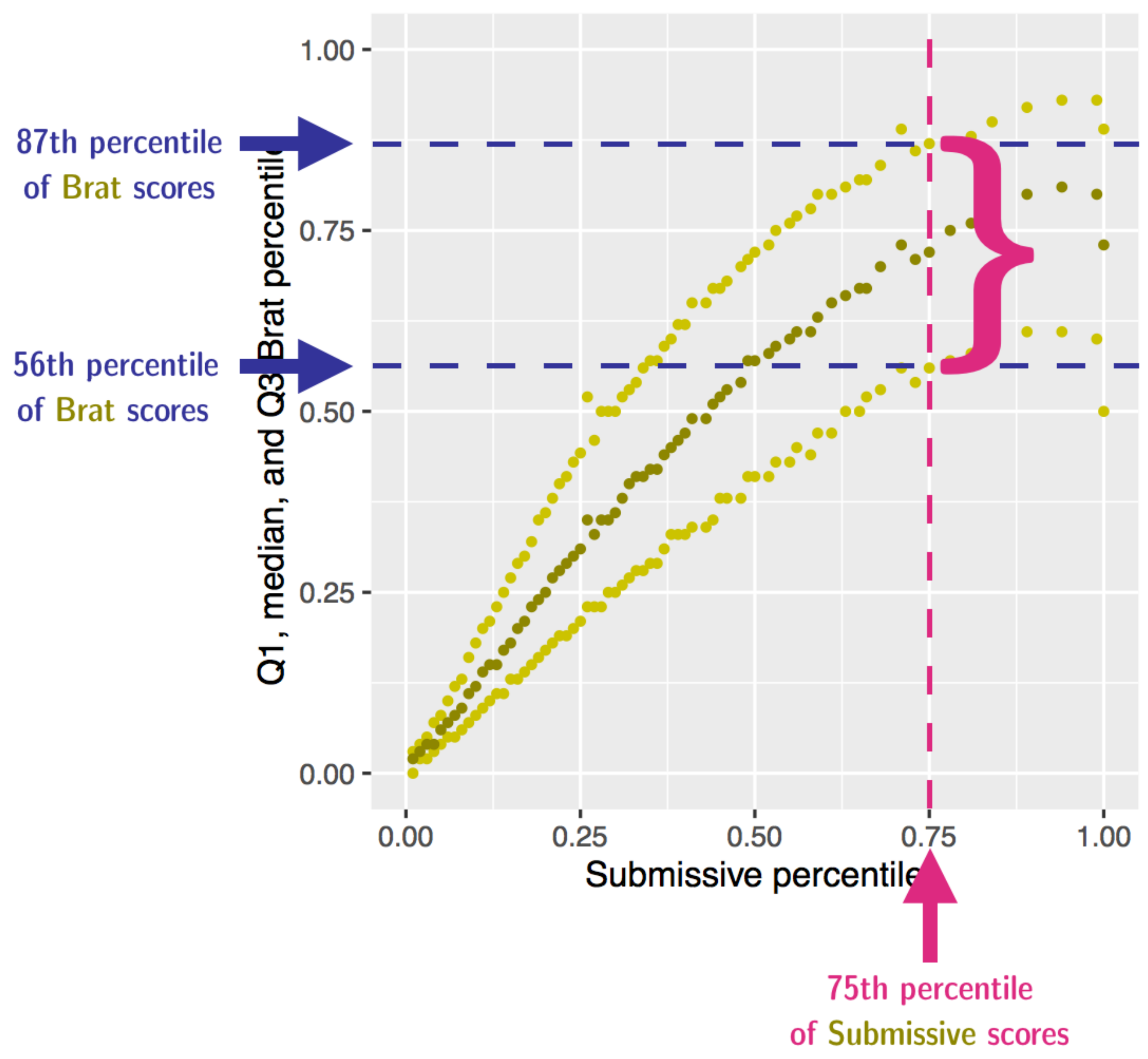
- So what transformation would make sense?
  - *z*-scores? ✗
  - Log-transformation? ✗
  - Rank transformation? ✓



Stage 1: One-variable EDA

Normalize each variable

Percentile rank transformation



The q^{th} **percentile** ($0 \leq q \leq 1$) is the score below which ($100q$)% of the data lies.

We'll call q the **percentile rank**.

Stage 1: One-variable EDA

Normalize each variable

After rank-transforming each variable...

- Comparisons between variables are more meaningful
- 2nd, 3rd, and 4th moments are approximately equal

Before rank-transformation

score	Kurt	Skew
Submissive	2.87	-1.01
Bondage Bottom	2.08	-0.68
Experimentalist	2.45	-0.61
Switch	1.82	-0.39
Dominant	1.78	-0.22
Masochist	1.73	-0.10
Bondage Top	1.56	-0.07
Voyeur	1.57	-0.02
Vanilla	2.03	-0.02
Brat	1.89	-0.01
Exhibitionist	1.71	0.23
Prey	1.81	0.31
Non-Monogamist	1.80	0.32
Master/Mistress	1.89	0.37
Brat Tamer	1.97	0.46
Degradee	1.71	0.48
Slave	1.98	0.49
Sadist	1.90	0.49
Hunter	1.99	0.58
Boy/Girl	2.10	0.61
Ageplayer	2.14	0.62
Daddy/Mommy	2.52	0.80
Owner	2.49	0.86
Degrader	2.44	0.87
Pet	2.53	0.99

After rank-transformation

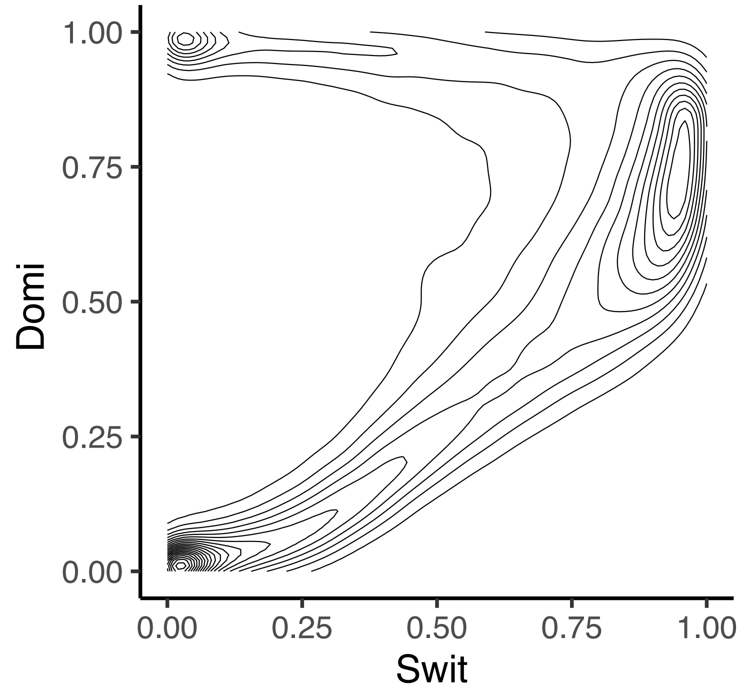
rank	s	Kurt	Skew
Ageplayer	0.28	1.80	-0.01
Bondage Bottom	0.29	1.81	-0.03
Bondage Top	0.29	1.81	-0.02
Boy/Girl	0.29	1.79	-0.01
Brat	0.29	1.80	0
Brat Tamer	0.28	1.81	-0.02
Daddy/Mommy	0.28	1.79	-0.01
Degradee	0.28	1.79	-0.05
Degrader	0.28	1.78	-0.04
Dominant	0.29	1.80	-0.02
Exhibitionist	0.29	1.81	-0.01
Experimentalist	0.29	1.81	0.01
Hunter	0.28	1.80	-0.02
Masochist	0.29	1.81	0
Master/Mistress	0.29	1.81	-0.01
Non-Monogamist	0.29	1.80	-0.01
Owner	0.28	1.80	-0.05
Pet	0.28	1.76	-0.01
Prey	0.29	1.81	-0.01
Sadist	0.28	1.80	-0.02
Slave	0.28	1.81	-0.02
Submissive	0.30	1.80	-0.03
Switch	0.29	1.81	-0.02
Vanilla	0.29	1.80	0.01
Voyeur	0.29	1.80	-0.01

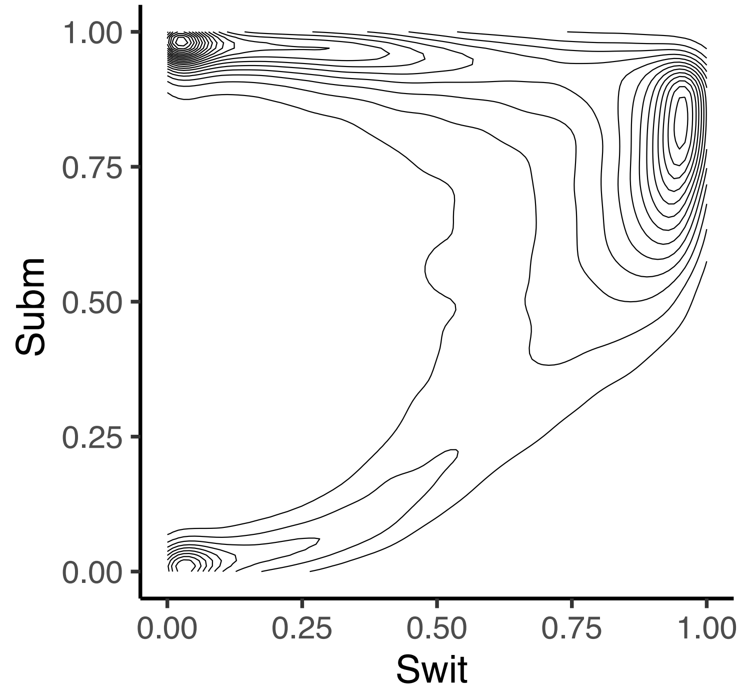
Stage 2: Two-variable EDA

Assumptions for linear regression

Why the assumptions matter:

- Linear correlation coefficients can't be trusted for nonlinear data
  - Should we expect similar values of Pearson's r for the two density plots shown?



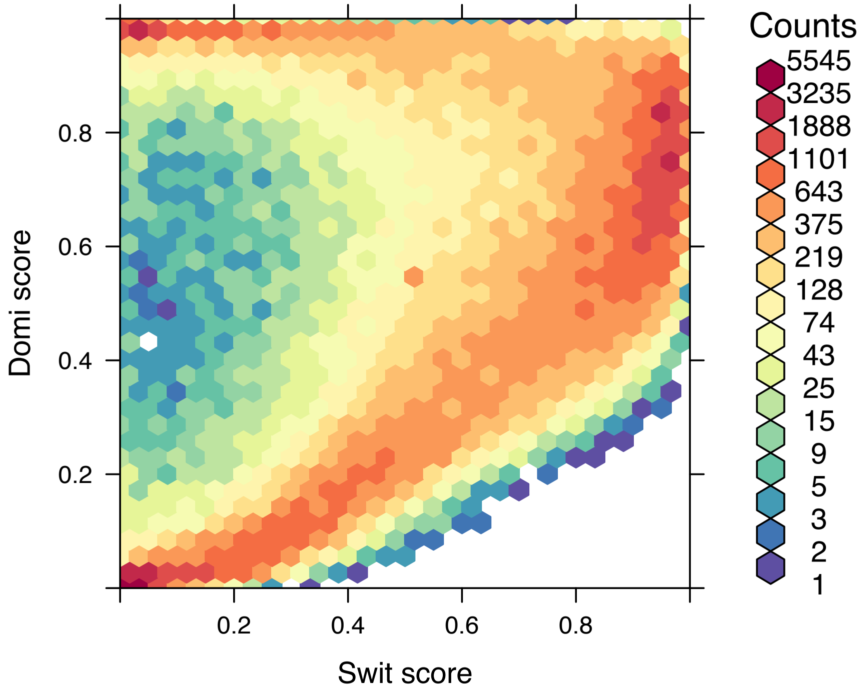


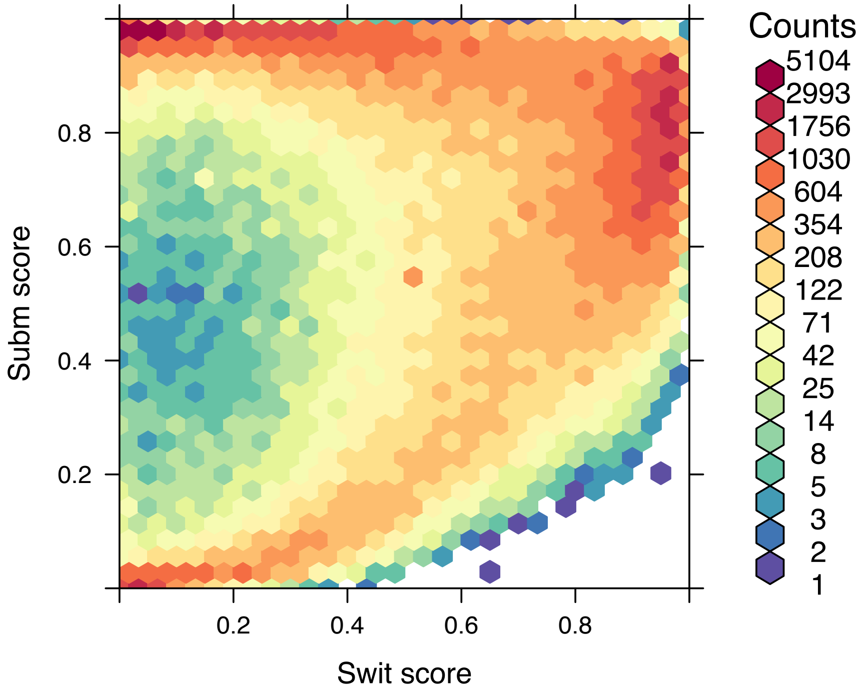
Stage 2: Two-variable EDA

Assumptions for linear regression

Why the assumptions matter:

- Linear correlation coefficients can't be trusted for nonlinear data
  - Should we expect similar values of Pearson's *r* for the two hexbin plots shown?
  - Would it help to transpose *x* and *y*?

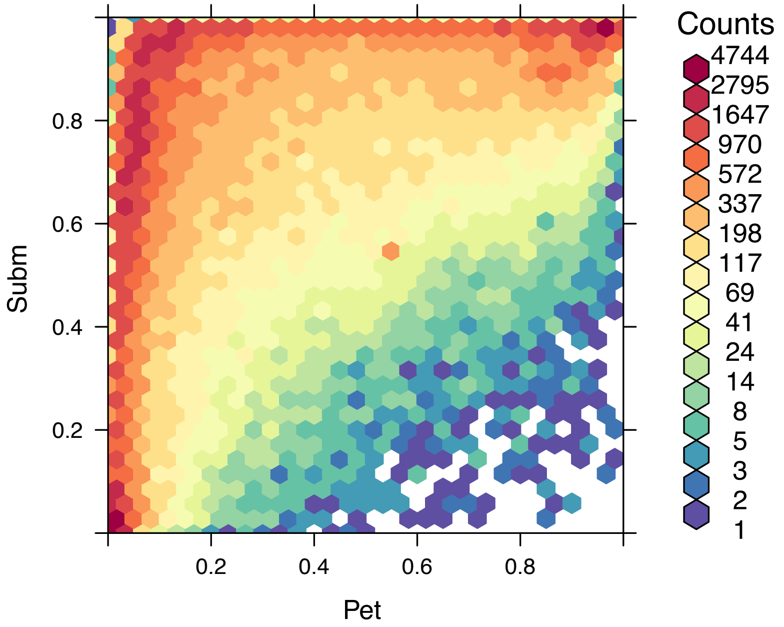




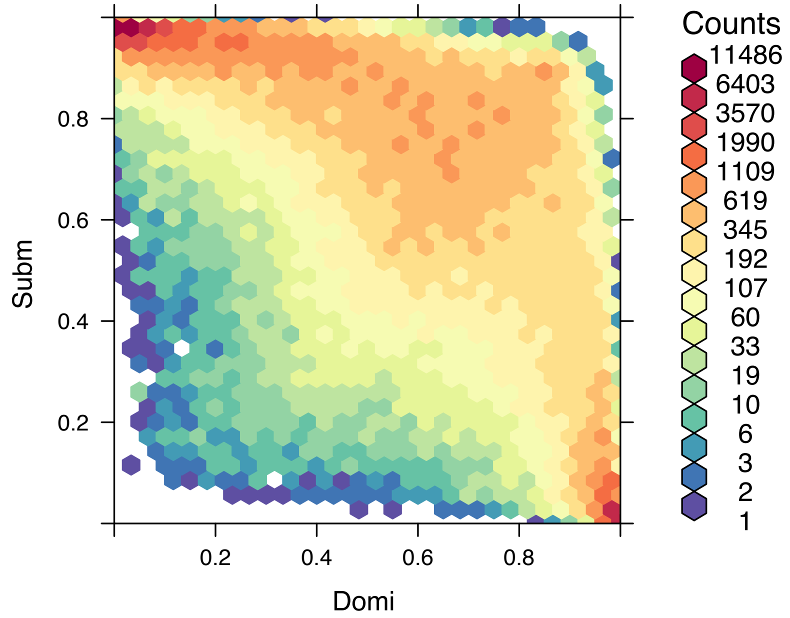
Stage 2: Two-variable EDA

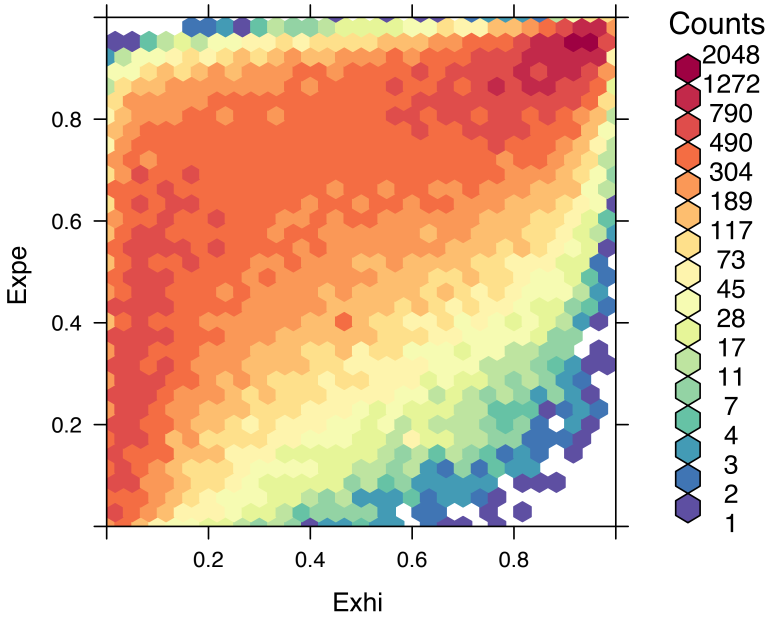
Assumptions for linear regression

How the assumptions appeared to be violated:

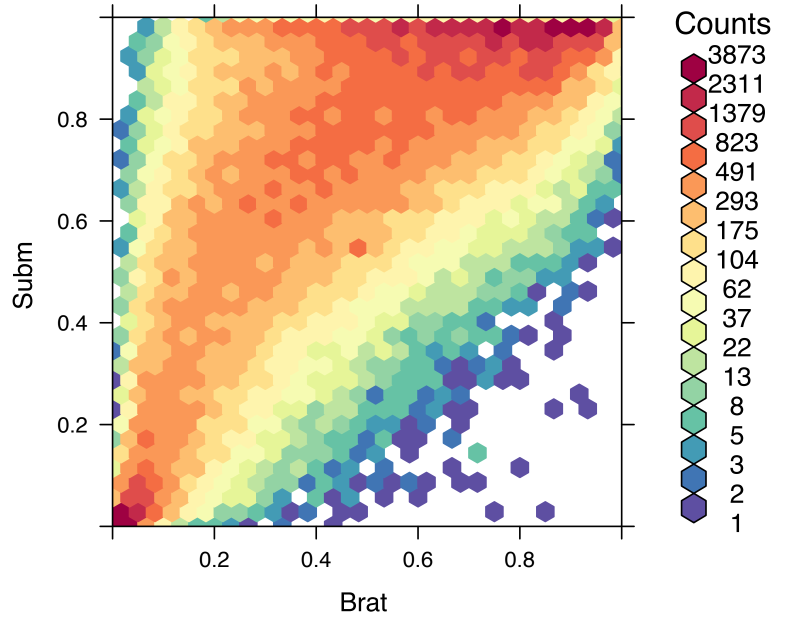


- Most bivariate projections were *very* nonlinear...





- ...and extreme values tended to occur at very high frequencies

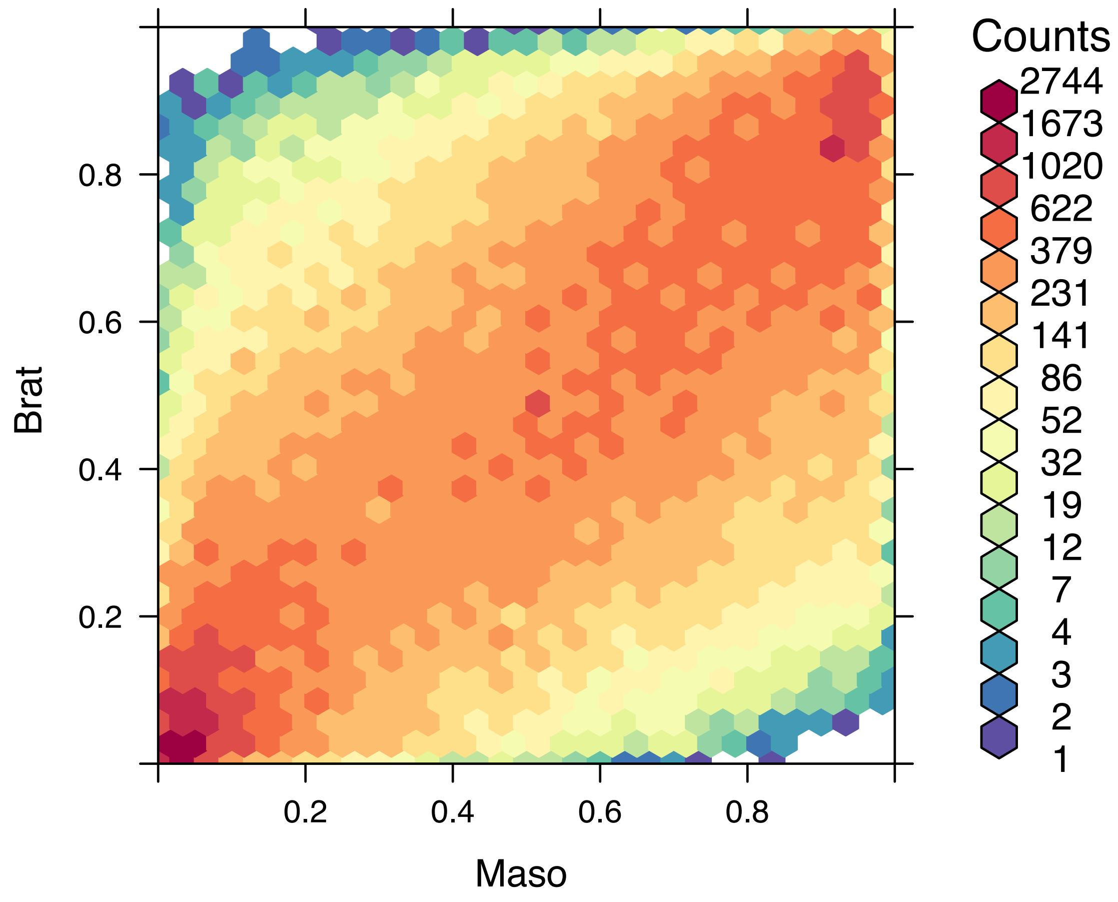


Stage 2: Two-variable EDA

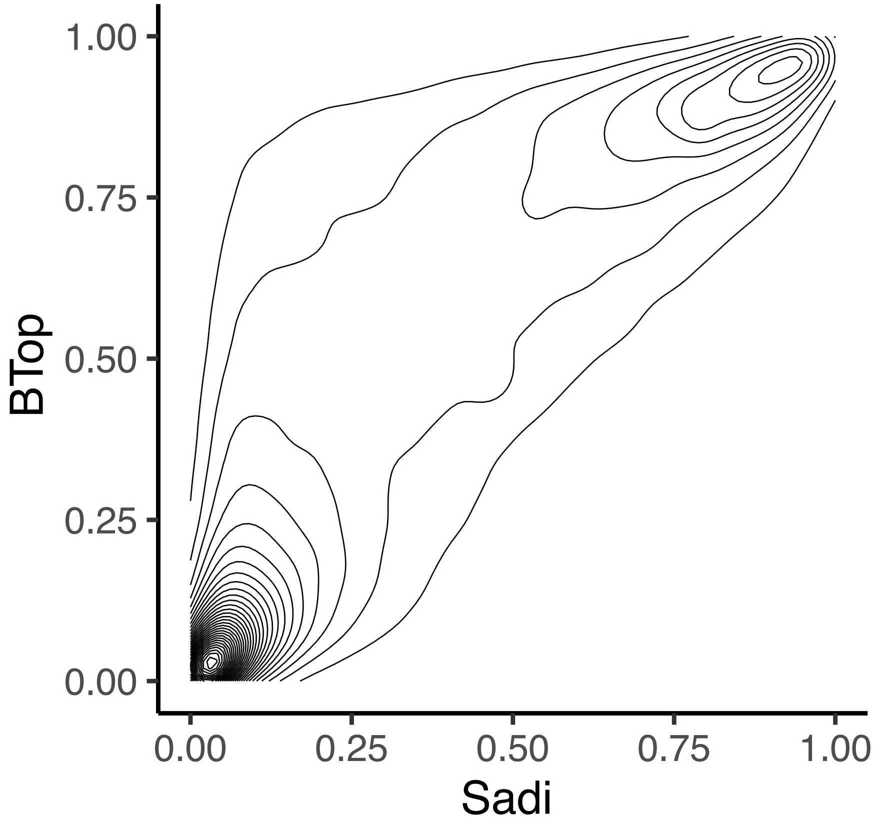
Assumptions for linear regression

How the assumptions appeared to be violated:

- Large variation in y for fixed x



- Strong heteroscedasticity



Stage 2: Two-variable EDA

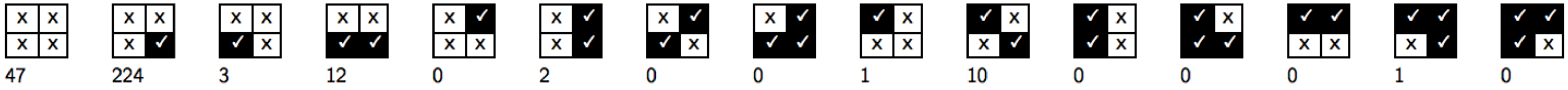
Assumptions for linear regression

Formally checking the assumptions

- R function `gvlma()` tests assumptions for classical linear regression

  ✗ Every pair of variables (raw scores) failed (228 of 300 pairs failed 3 of 4 tests)

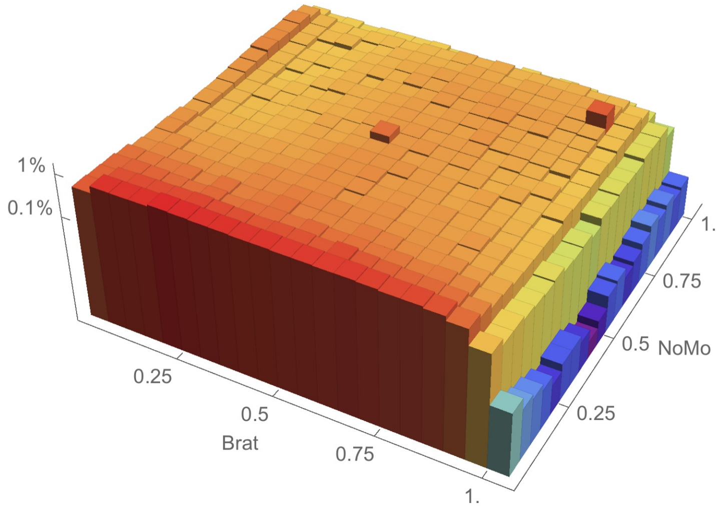
Figure 3: Tallied results of failed GVLMA tests for each pair of variables (clockwise from the top left of each grid: skewness, kurtosis, link function, and heteroscedasticity).

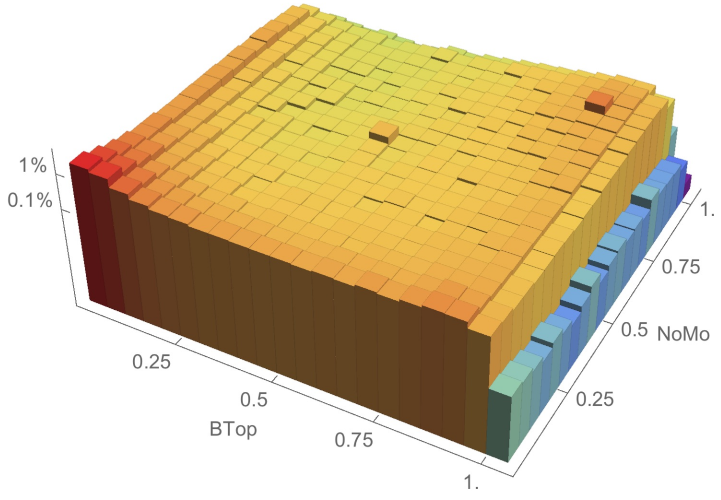


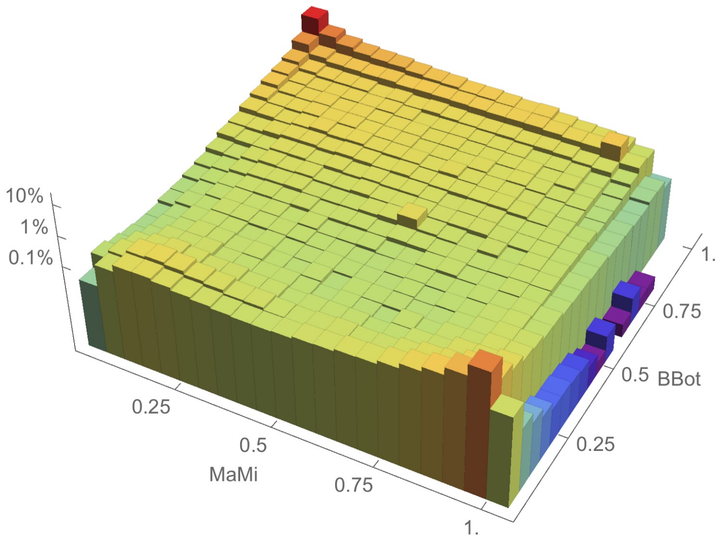
Stage 2: Two-variable EDA

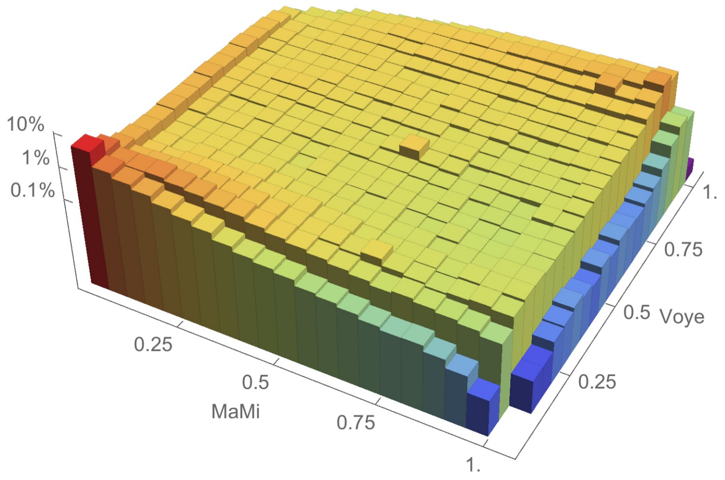
Assumptions for linear regression

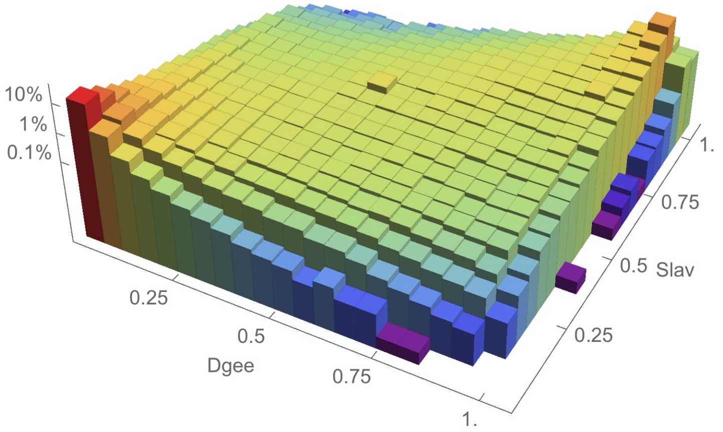
- 3D histograms for pairs (rank-transformed scores) that only failed one of `gvlma()`'s tests:









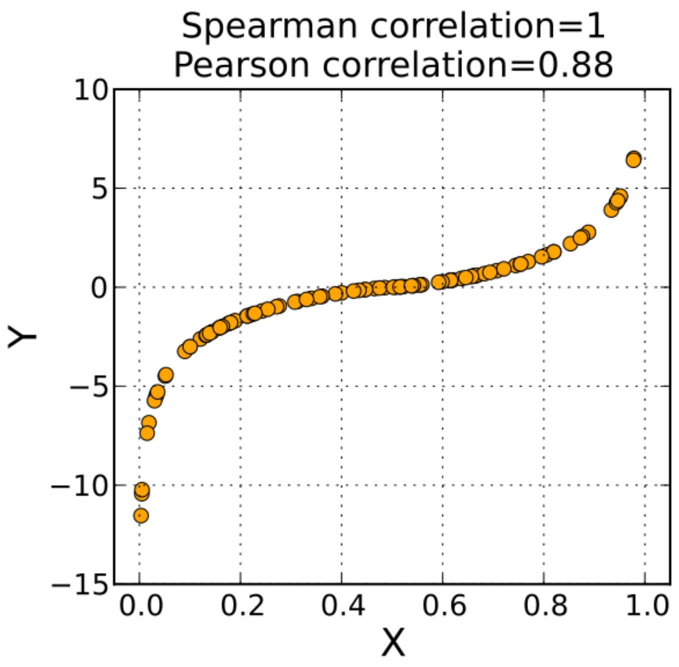


Stage 2: Two-variable EDA

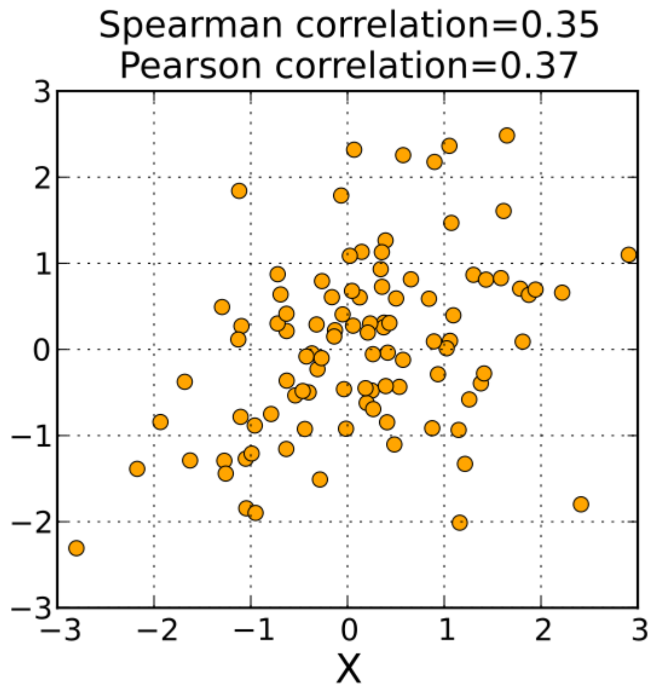
Nonparametric correlation coefficients

Spearman's rank correlation coefficient ρ

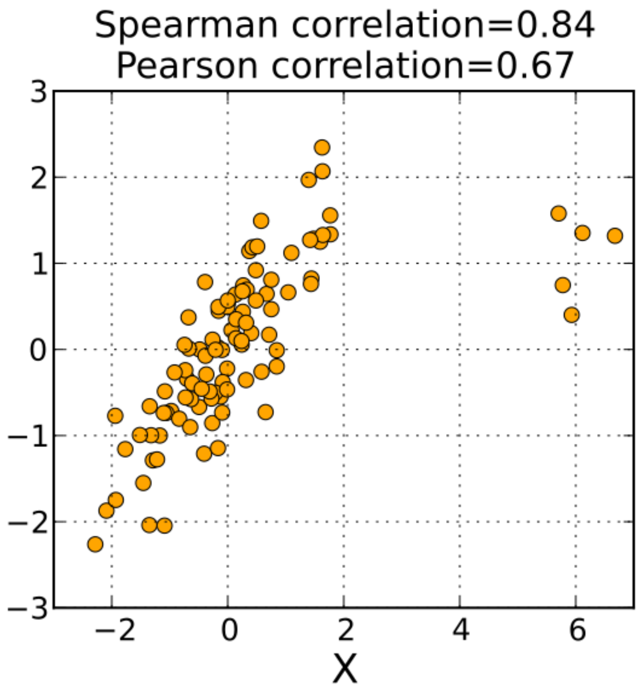
- Measures ordinal, not linear association ✓
- More resistant to outliers than Pearson's r ✓
- Does not handle ties well ✗



Spearman correlation=1
Pearson correlation=0.88



Spearman correlation=0.35
Pearson correlation=0.37



Spearman correlation=0.84
Pearson correlation=0.67

Image source: Wikipedia

Stage 2: Two-variable EDA

A nonparametric correlation coefficient

Kendall's rank correlation coefficient τ_b

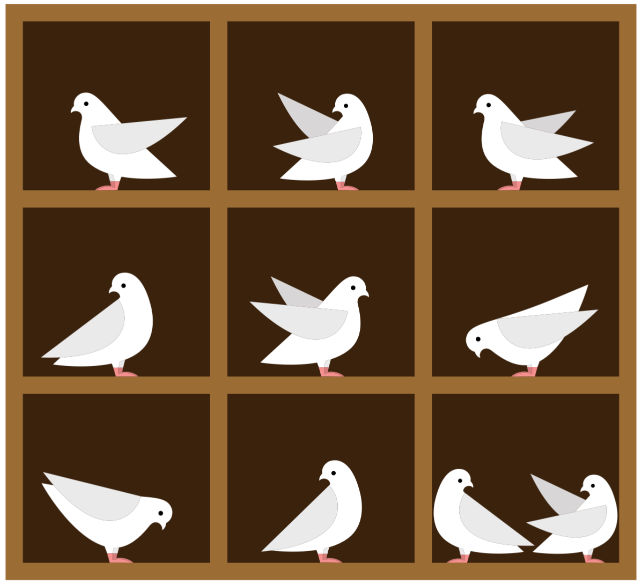
- Measures ordinal, not linear association ✓
- More resistant to outliers than Pearson's r ✓
- Corrects for ties ✓

Sample size: 236,353

# of possible values for each variable: 101

$\Downarrow$

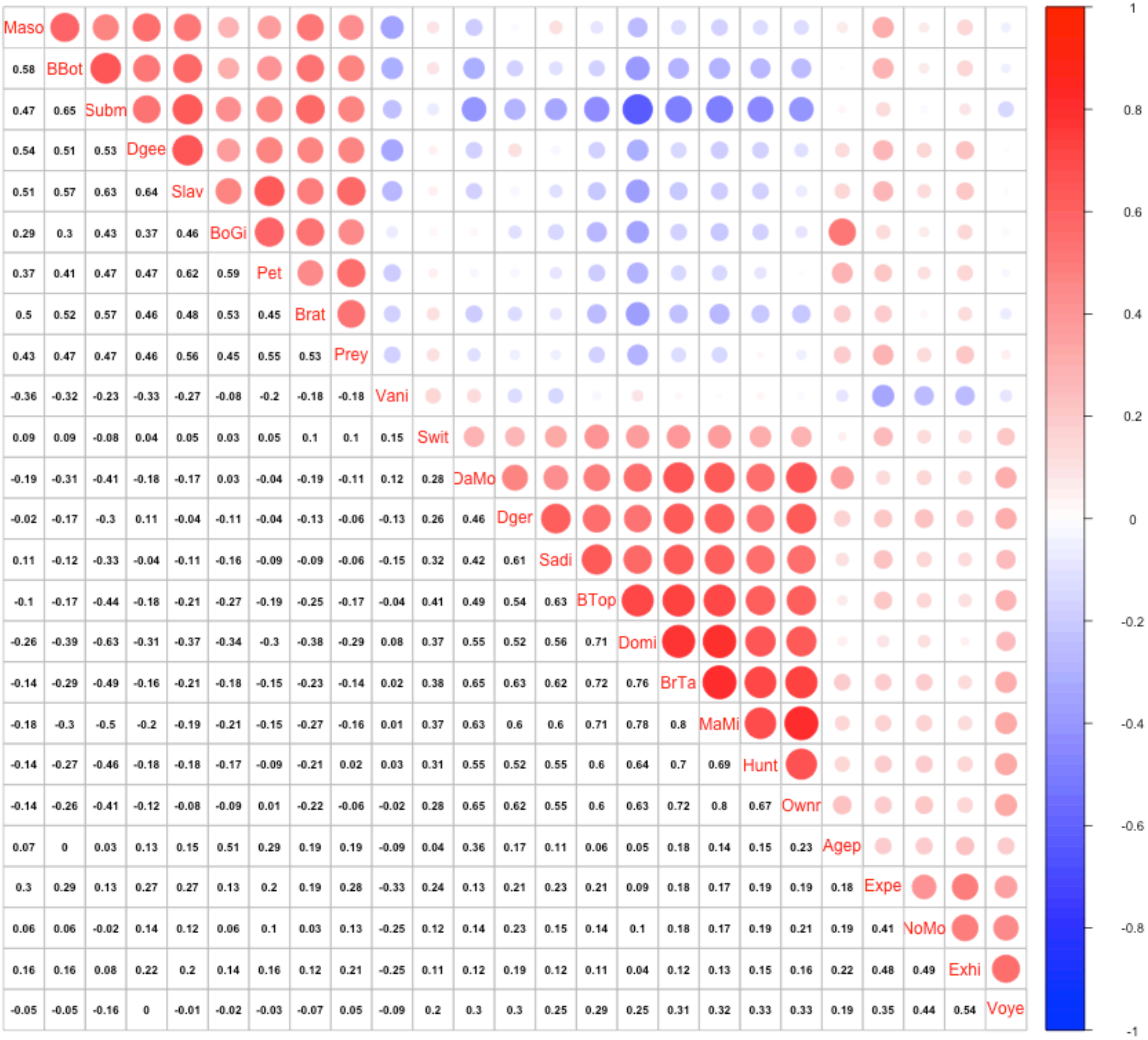
lots of ties!



Stage 2: Two-variable EDA

Cluster the variables by correlation

A **corrgram** summarizes the correlation coefficients between the variables.



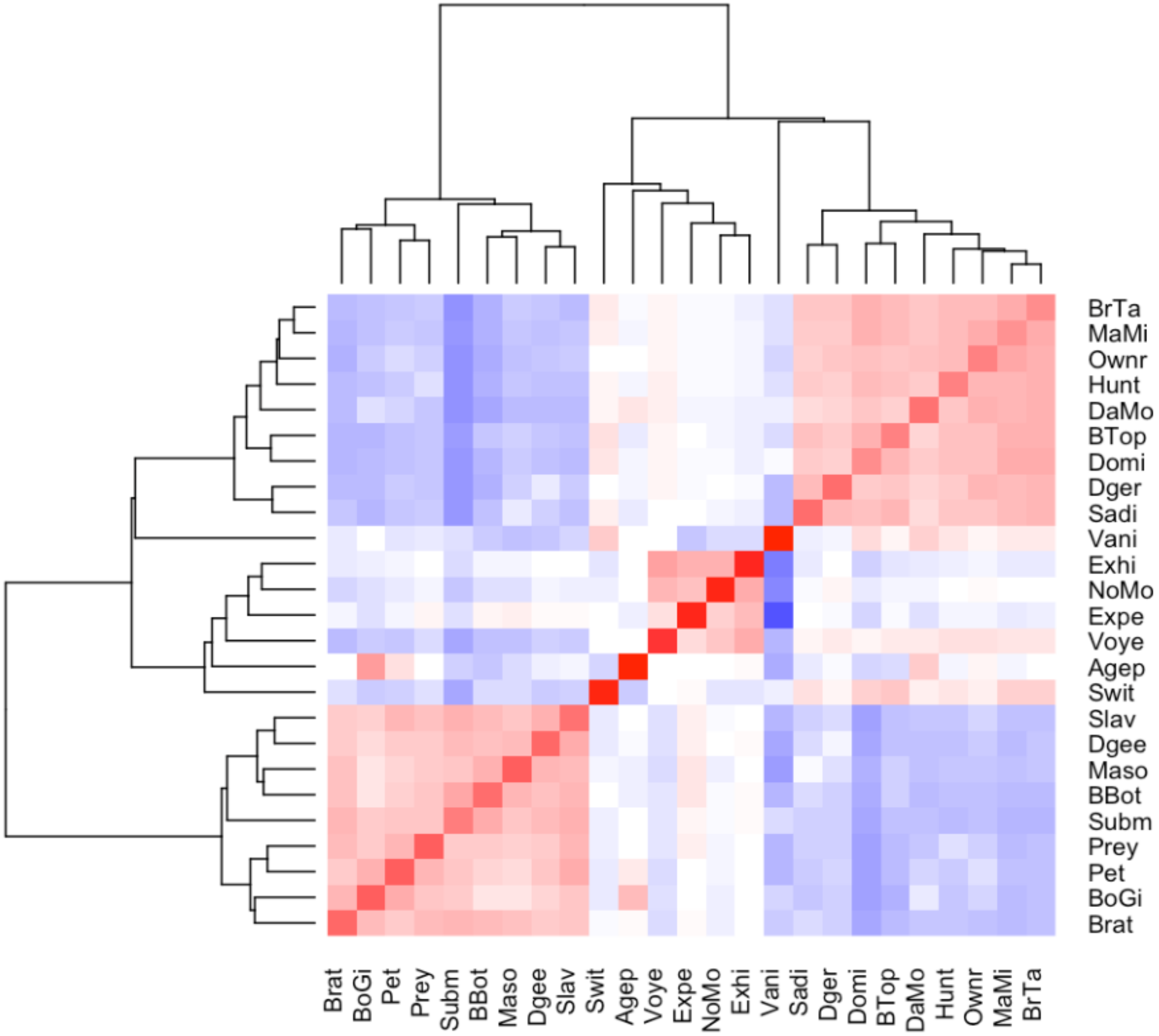
Maso																								
0.58	BBot																							
0.47	0.65	Subm																						
0.54	0.51	0.53	Dgee																					
0.51	0.57	0.63	0.64	Slav																				
0.29	0.3	0.43	0.37	0.46	BoGi																			
0.37	0.41	0.47	0.47	0.62	0.59	Pet																		
0.5	0.52	0.57	0.46	0.48	0.53	0.45	Brat																	
0.43	0.47	0.47	0.46	0.56	0.45	0.55	0.53	Prey																
-0.36	-0.32	-0.23	-0.33	-0.27	-0.08	-0.2	-0.18	-0.18	Vani															
0.09	0.09	-0.08	0.04	0.05	0.03	0.05	0.1	0.1	0.15	Swit														
-0.19	-0.31	-0.41	-0.18	-0.17	0.03	-0.04	-0.19	-0.11	0.12	0.28	DaMo													
-0.02	-0.17	-0.3	0.11	-0.04	-0.11	-0.04	-0.13	-0.06	-0.13	0.26	0.46	Dger												
0.11	-0.12	-0.33	-0.04	-0.11	-0.16	-0.09	-0.09	-0.06	-0.15	0.32	0.42	0.61	Sadi											
-0.1	-0.17	-0.44	-0.18	-0.21	-0.27	-0.19	-0.25	-0.17	-0.04	0.41	0.49	0.54	0.63	BTop										
-0.26	-0.39	-0.63	-0.31	-0.37	-0.34	-0.3	-0.38	-0.29	0.08	0.37	0.55	0.52	0.56	0.71	Domi									
-0.14	-0.29	-0.49	-0.16	-0.21	-0.18	-0.15	-0.23	-0.14	0.02	0.38	0.65	0.63	0.62	0.72	0.76	BrTa								
-0.18	-0.3	-0.5	-0.2	-0.19	-0.21	-0.15	-0.27	-0.16	0.01	0.37	0.63	0.6	0.6	0.71	0.78	0.8	MaMi							
-0.14	-0.27	-0.46	-0.18	-0.18	-0.17	-0.09	-0.21	0.02	0.03	0.31	0.55	0.52	0.55	0.6	0.64	0.7	0.69	Hunt						
-0.14	-0.26	-0.41	-0.12	-0.08	-0.09	0.01	-0.22	-0.06	-0.02	0.28	0.65	0.62	0.55	0.6	0.63	0.72	0.8	0.67	Ownr					
0.07	0	0.03	0.13	0.15	0.51	0.29	0.19	0.19	-0.09	0.04	0.36	0.17	0.11	0.06	0.05	0.18	0.14	0.15	0.23	Agep				
0.3	0.29	0.13	0.27	0.27	0.13	0.2	0.19	0.28	-0.33	0.24	0.13	0.21	0.23	0.21	0.09	0.18	0.17	0.19	0.19	0.18	Expe			
0.06	0.06	-0.02	0.14	0.12	0.06	0.1	0.03	0.13	-0.25	0.12	0.14	0.23	0.15	0.14	0.1	0.18	0.17	0.19	0.21	0.19	0.41	NoMo		
0.16	0.16	0.08	0.22	0.2	0.14	0.16	0.12	0.21	-0.25	0.11	0.12	0.19	0.12	0.11	0.04	0.12	0.13	0.15	0.16	0.22	0.48	0.49	Exhi	
-0.05	-0.05	-0.16	0	-0.01	-0.02	-0.03	-0.07	0.05	-0.09	0.2	0.3	0.3	0.25	0.29	0.25	0.31	0.32	0.33	0.33	0.19	0.35	0.44	0.54	Voye

Stage 2: Two-variable EDA

Cluster the variables by correlation

Hierarchical clustering of the variables by τ_b:

- D-types and *Swit*
- s-types
- non-D/s kink roles
- *Vani*



Stage 3: Multivariate analysis

Choosing the algorithm parameters

We now seek to classify individual survey responses.

We'll divide them up into groups of "similar" responses.

Each group of similar responses is called a **cluster**.

A division into groups is called a **clustering**.

The computational technique we'll use is called **cluster analysis** (specifically, agglomerative hierarchical cluster analysis).

Stage 3: Multivariate analysis

Choosing the algorithm parameters

Parameters for hierarchical clustering

- dissimilarity metric d: {pairs of survey responses} → $[0,\infty)$
- number J of clusters
- linkage method ℓ

How do we pick d, J, and ℓ ?

Stage 3: Multivariate analysis

Choosing the algorithm parameters

Replication technique

- Fix a choice of d, J, ℓ and subsample size n
- Draw K random subsamples of size n from the given sample
- Cluster each subsample
- Compare the clusterings of the K subsamples
    - Are the characteristics of the clusters consistent across all K subsamples?
    - Do the clusters tend to be meaningfully separated?

Stage 3: Multivariate analysis

Choosing the algorithm parameters

What we want from our clustering

- Some cluster should contain all respondents who have high-ranked *Domi* scores and low-ranked *Subm* scores
  - Similarly for respondents who have high-ranked *Subm* scores and low-ranked *Domi* scores
- The median intracluster score in *Domi* should lie in a narrow range of values across all clusterings
  - Similarly for *Subm* and *Swit*

Stage 3: Multivariate analysis

Choosing the algorithm parameters

Visualizing a clustering in terms of our objectives

- For the *j*th cluster of the *k*th subsample, let

$$M_{j,k} = \left(M_{j,k}^{(i)}\right)_{i=1}^{3} \qquad (1 \leq j \leq J,\ 1 \leq k \leq K)$$

be the triple of component-wise medians

$$M_{j,k}^{(i)} = \text{median}\left(x_i \,|\, C_{j,k}\right) \qquad (1 \leq i \leq 3,\, 1 \leq j \leq J,\, 1 \leq k \leq K)$$

where

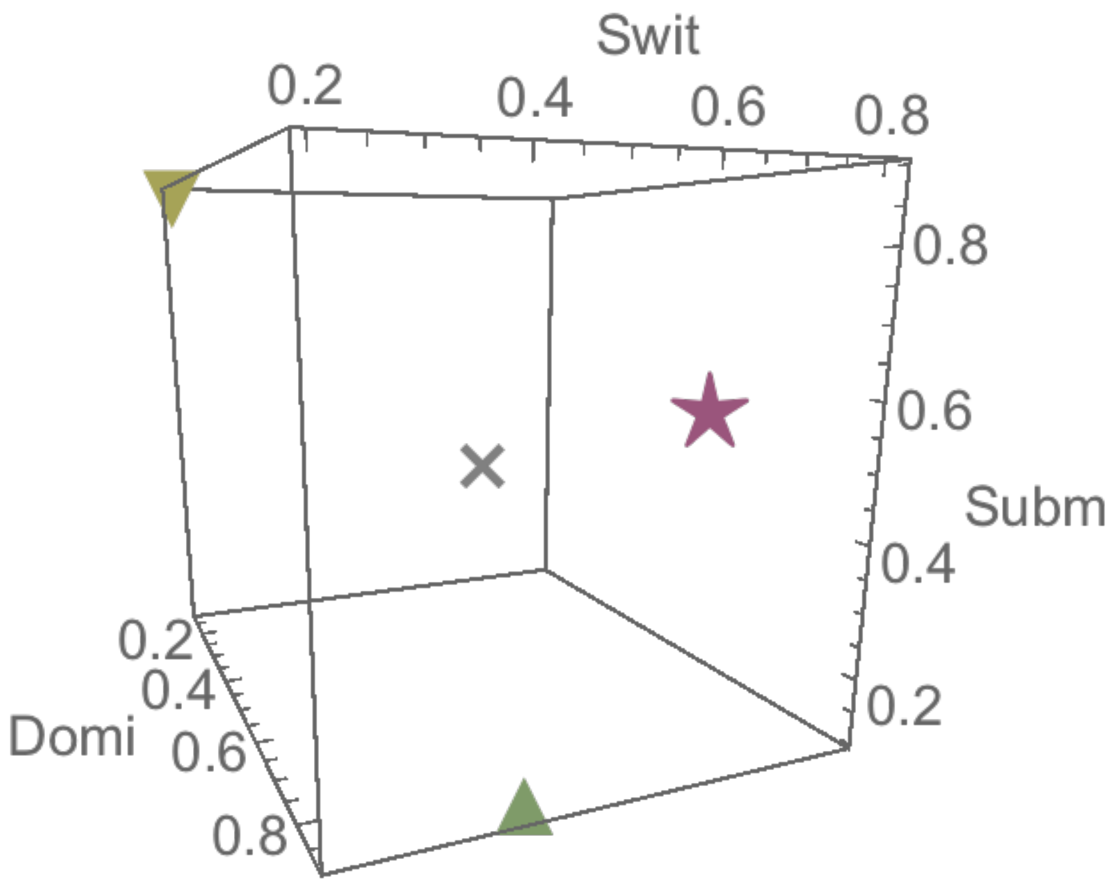
x_1 = (*Domi* rank), x_2 = (*Swit* rank), and x_3 = (*Subm* rank)

Stage 3: Multivariate analysis

Choosing the algorithm parameters

Visualizing a clustering in terms of our objectives

- Each clustering can thus be represented visually as a set of "**summary points**" in $\mathbb{R}^3$.
  - The picture shows intracluster medians for 4 clusters.

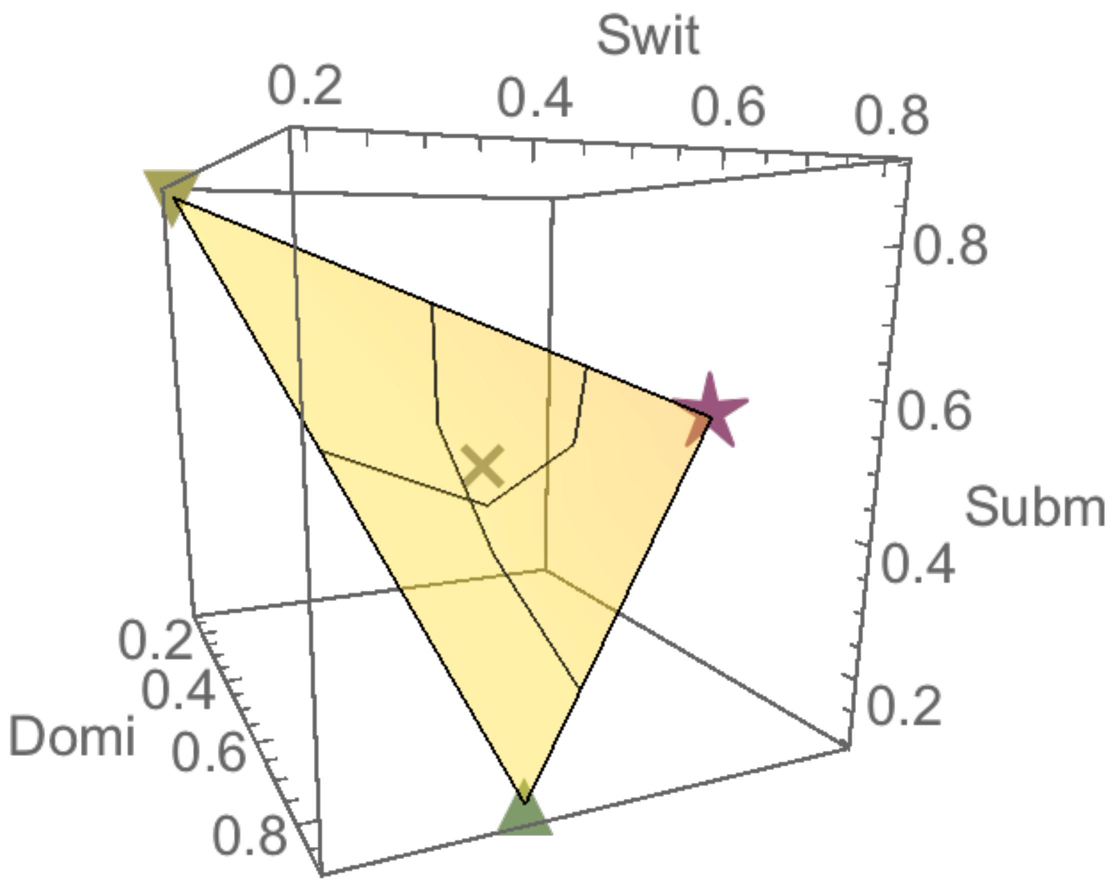


Stage 3: Multivariate analysis

Choosing the algorithm parameters

Visualizing a clustering in terms of our objectives

- Each clustering can thus be represented visually as a set of "**summary points**" in $\mathbb{R}^3$.
  - The picture shows intracluster medians for 4 clusters.
  - The curved surface clarifies position in 3D.



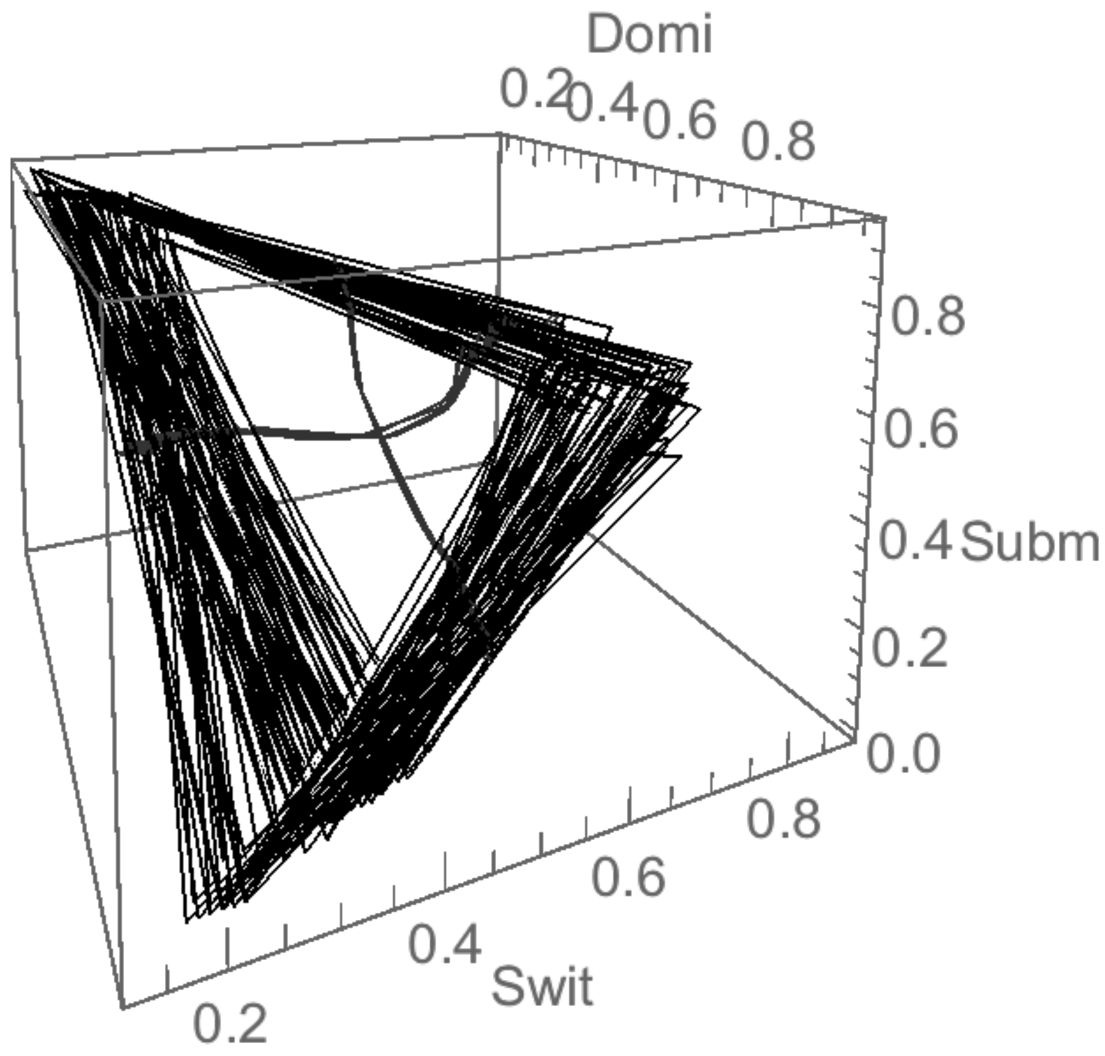
Stage 3: Multivariate analysis

Choosing the algorithm parameters

Visualizing a clustering in terms of our objectives

- We can compare the clusterings of different subsamples (for a fixed choice of parameters) by plotting the surfaces together.

Did we choose our parameters for clustering well?



```
ranks n=10000 numClusters=4 roles=Ds
metric=manhattan linkage=ward.D2
```

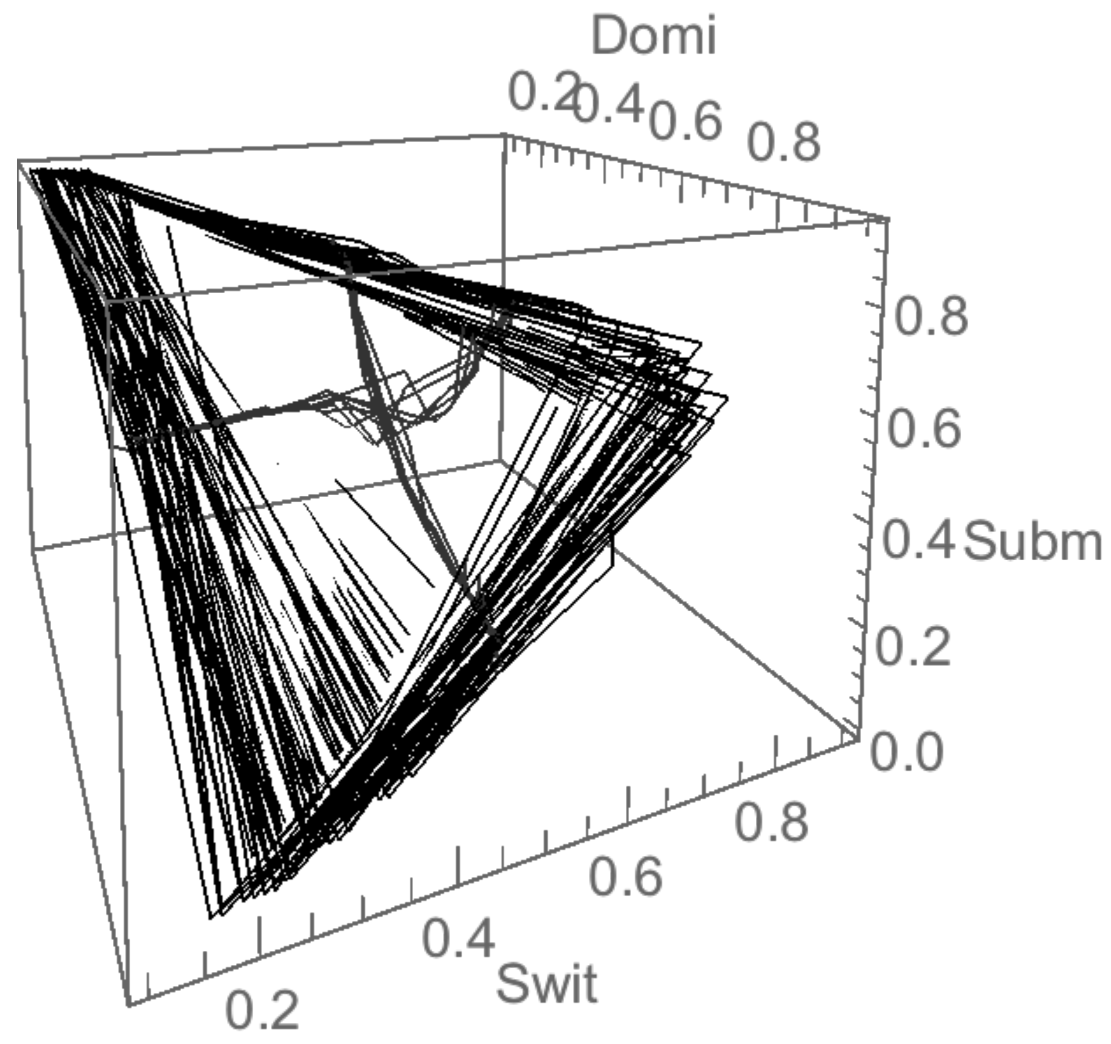
Stage 3: Multivariate analysis

Choosing the algorithm parameters

Visualizing a clustering in terms of our objectives

- We can compare the clusterings of different subsamples (for a fixed choice of parameters) by plotting the surfaces together.

More stable, or less stable?



```
ranks n=10000 numClusters=3 roles=Ds
 metric=manhattan linkage=ward.D2
```

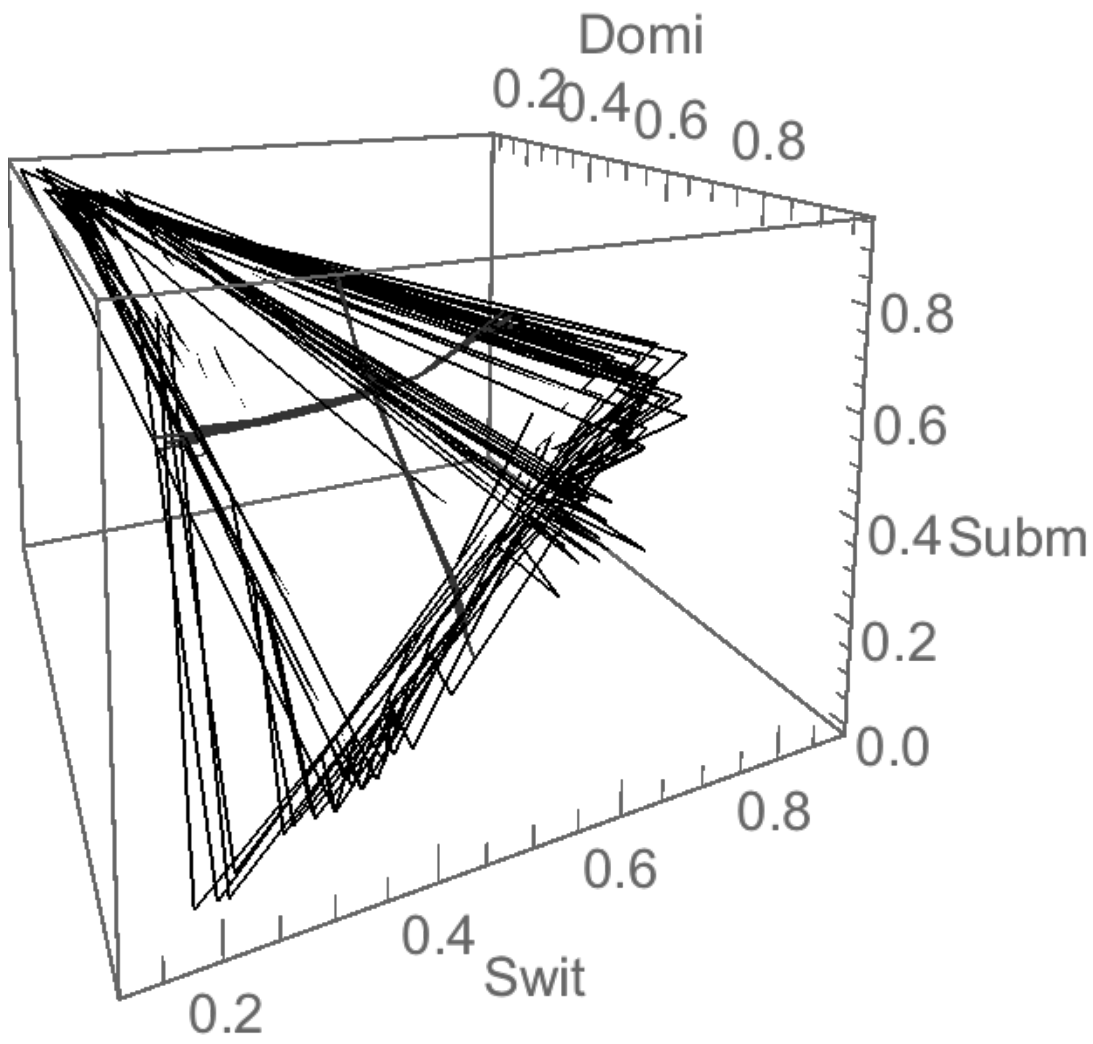
Stage 3: Multivariate analysis

Choosing the algorithm parameters

Visualizing a clustering in terms of our objectives

- We can compare the clusterings of different subsamples (for a fixed choice of parameters) by plotting the surfaces together.

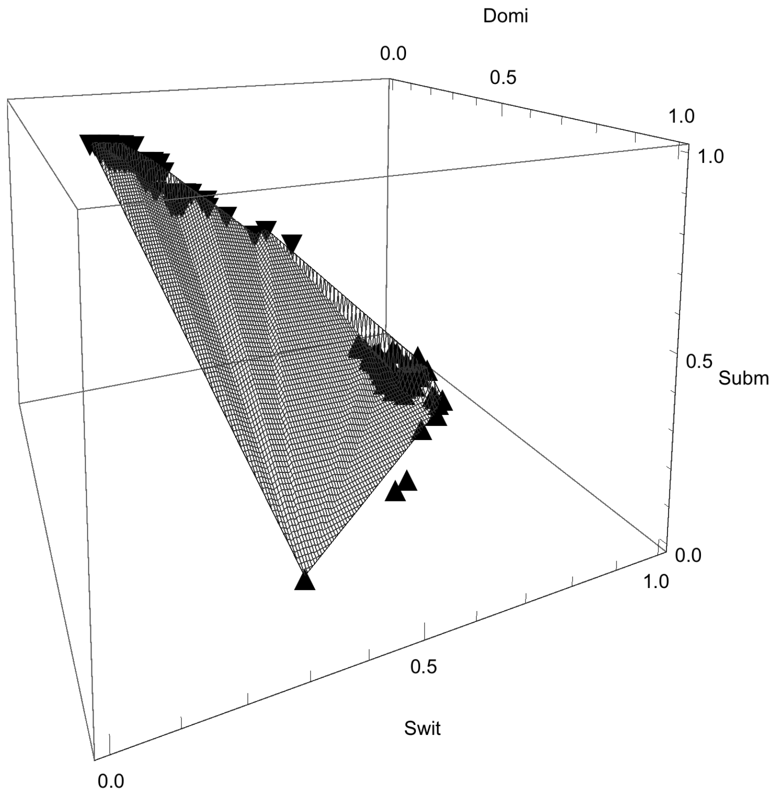
Better, or worse?



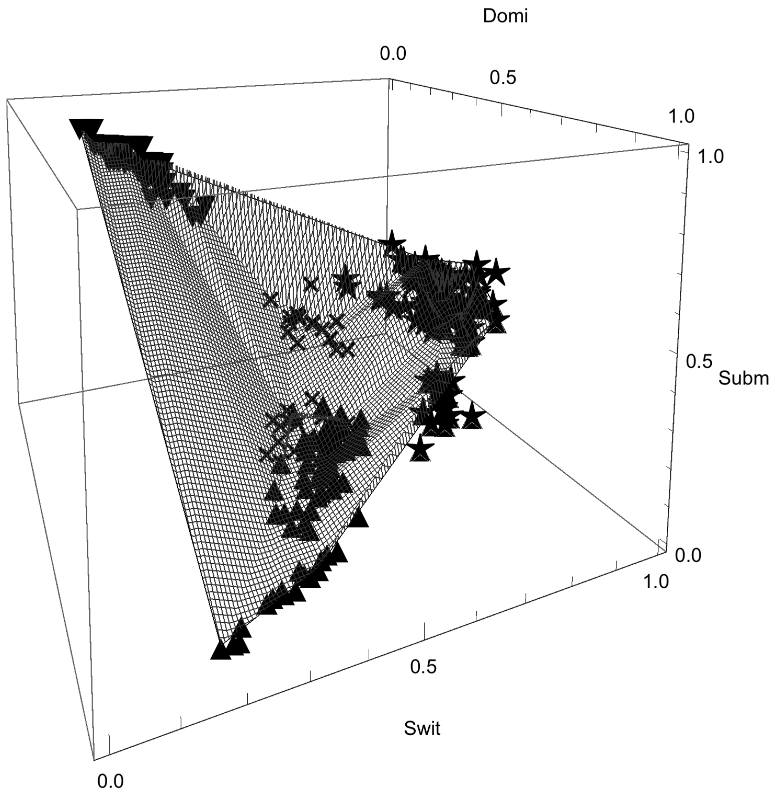
```
ranks n=10000 numClusters=5 roles=Ds
 metric=manhattan linkage=ward.D2
```

Stage 3: Multivariate analysis

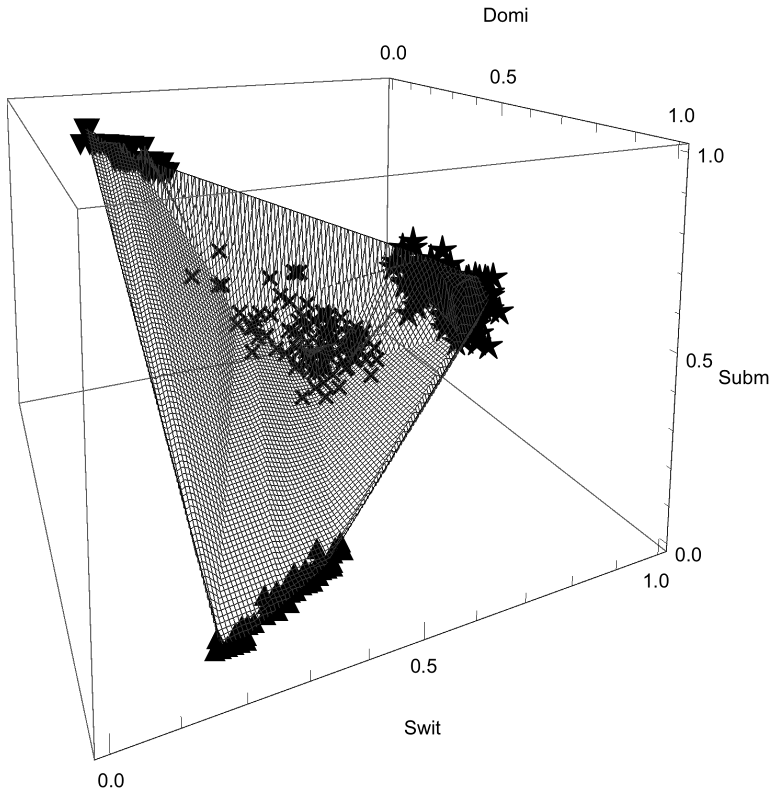
2 clusters



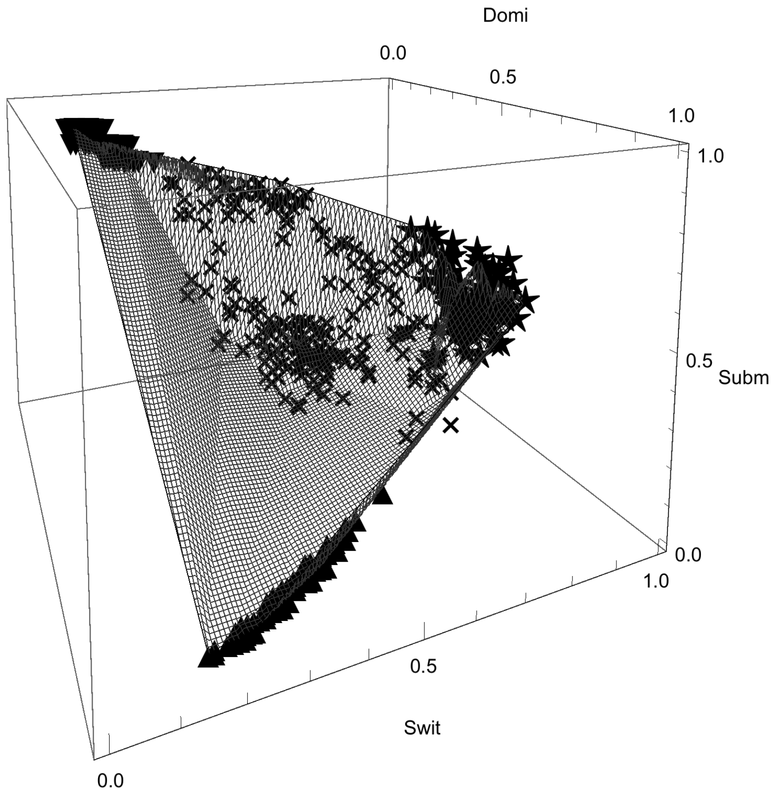
3 clusters



4 clusters



5 clusters



We chose

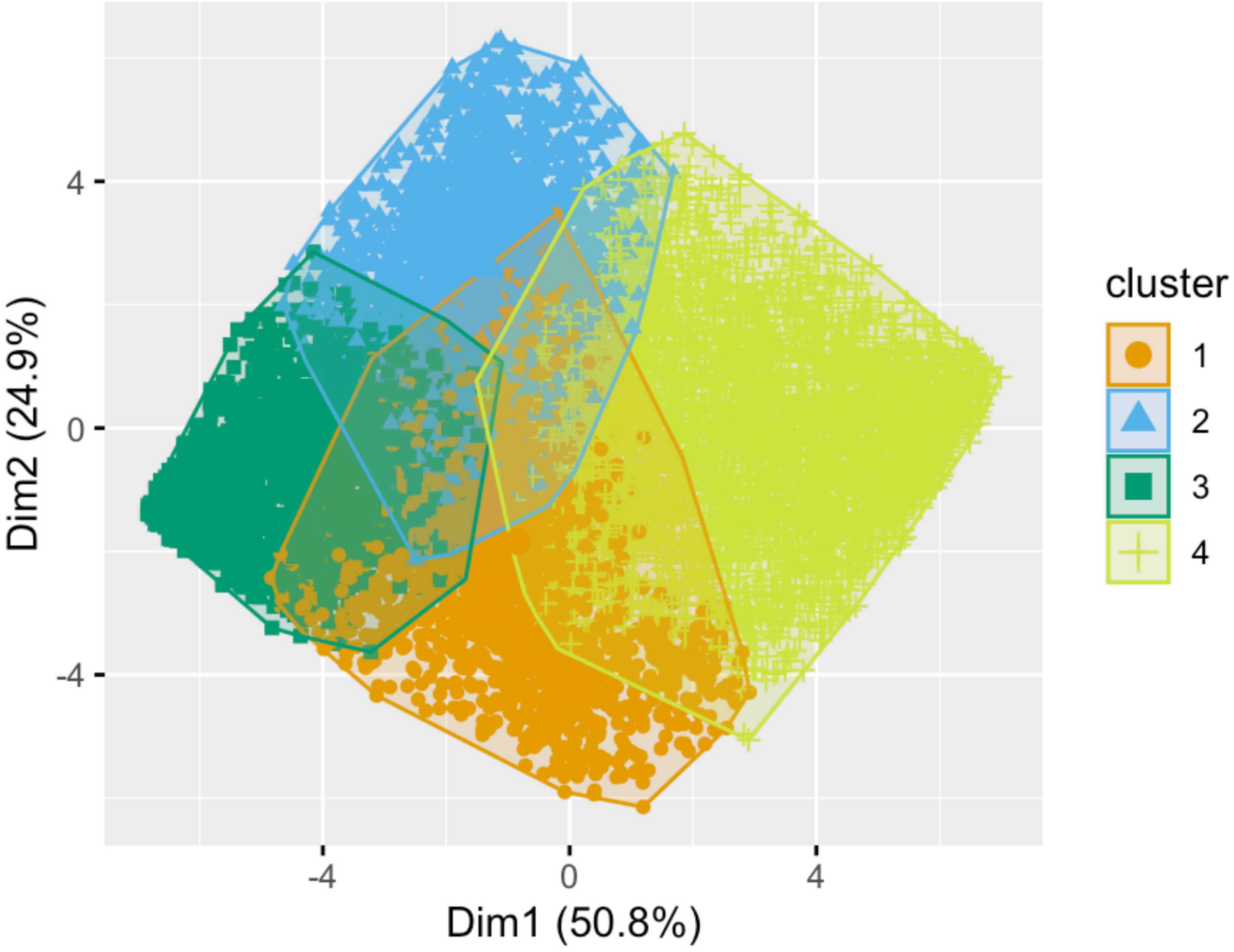
- $J = 4$ clusters
- $d = \mathrm{L}_1$ metric
- $\ell = \texttt{ward.D2}$

Stage 3: Multivariate analysis

Dimensional reduction

A **cluster plot** is a low-dimensional representation of how much the clusters overlap or are separated.

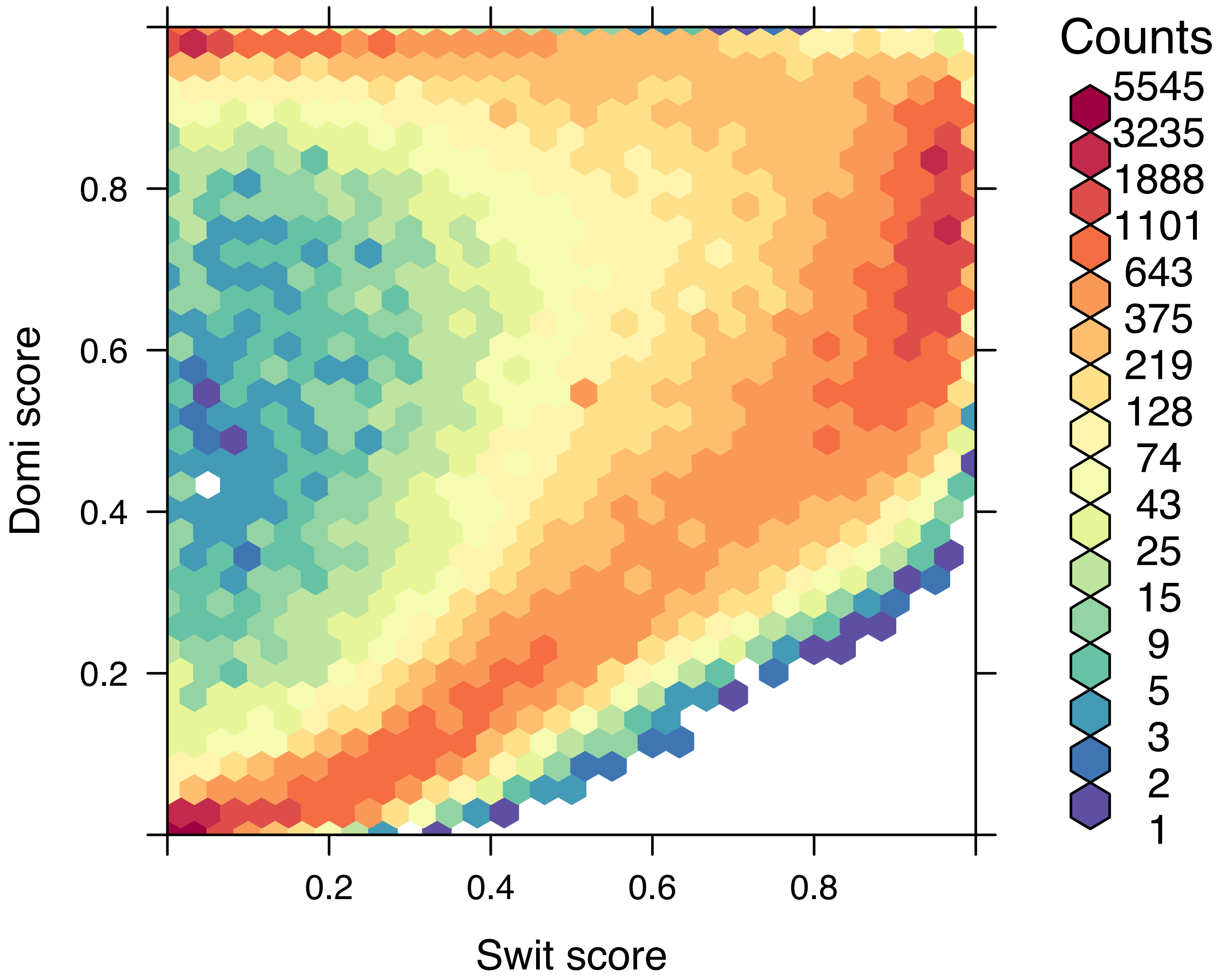
- The axes are the first two **principal components**. Each axis accounts for some proportion of the variance in all variables.

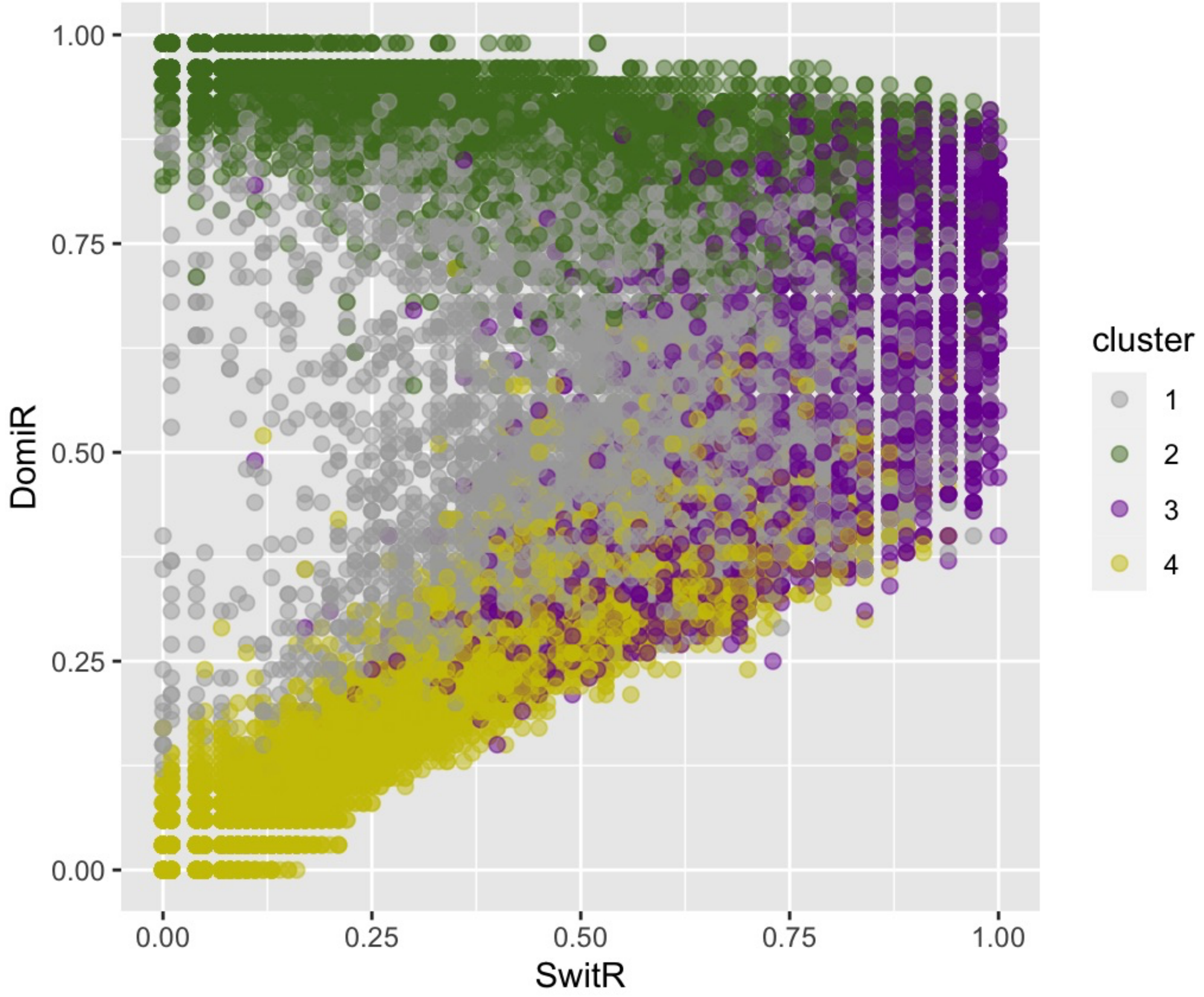


Stage 3: Multivariate analysis

Correlation within each cluster

Intracluster correlation may be more meaningful than correlation across the entire sample.

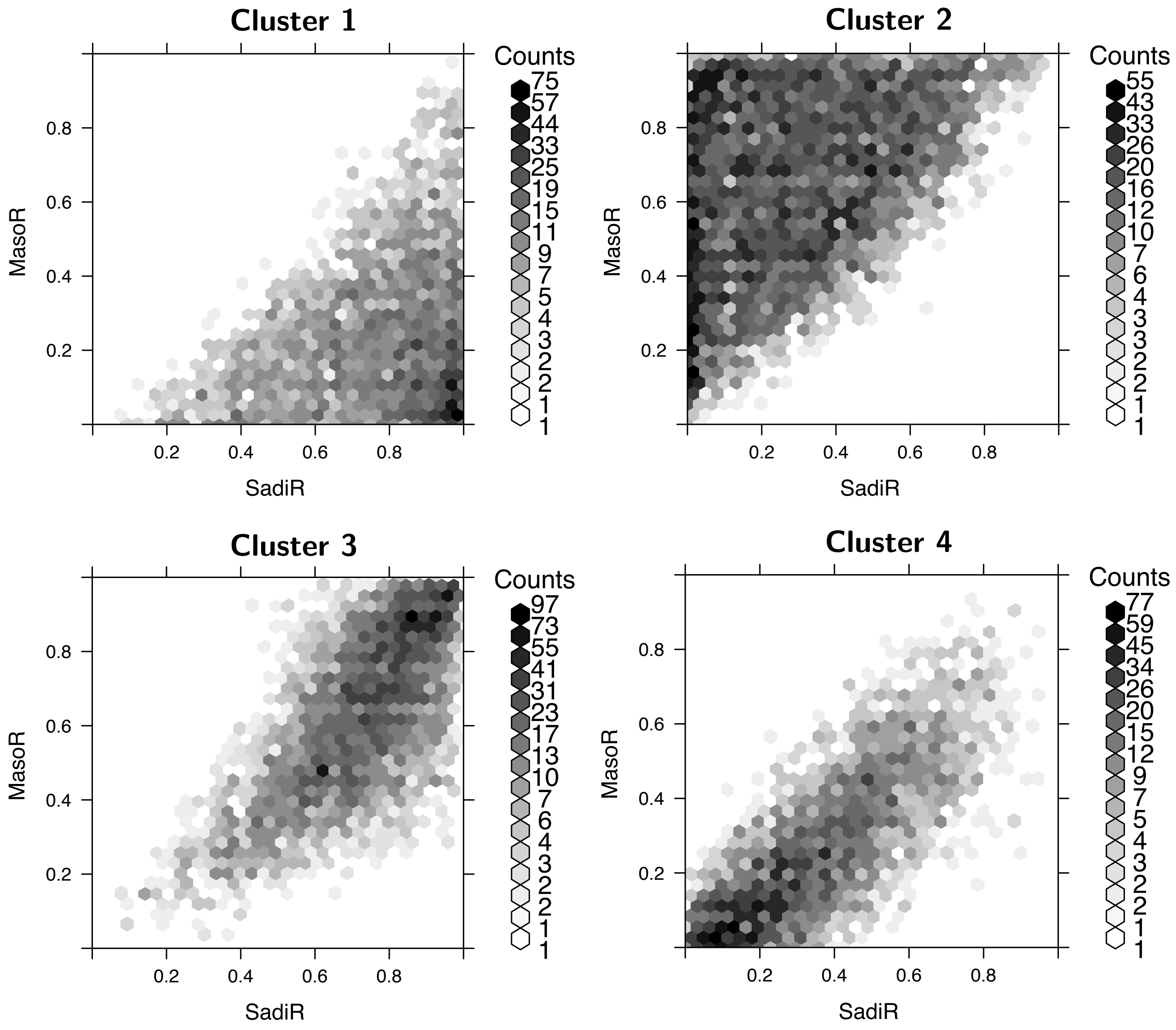




Stage 3: Multivariate analysis

Correlation within each cluster

Figure 8: The relationship between *Sadi* and *Maso* ranks in each cluster. (Clockwise from top: polar dominant, polar submissive, non-polar kinky, non-polar vanilla.)

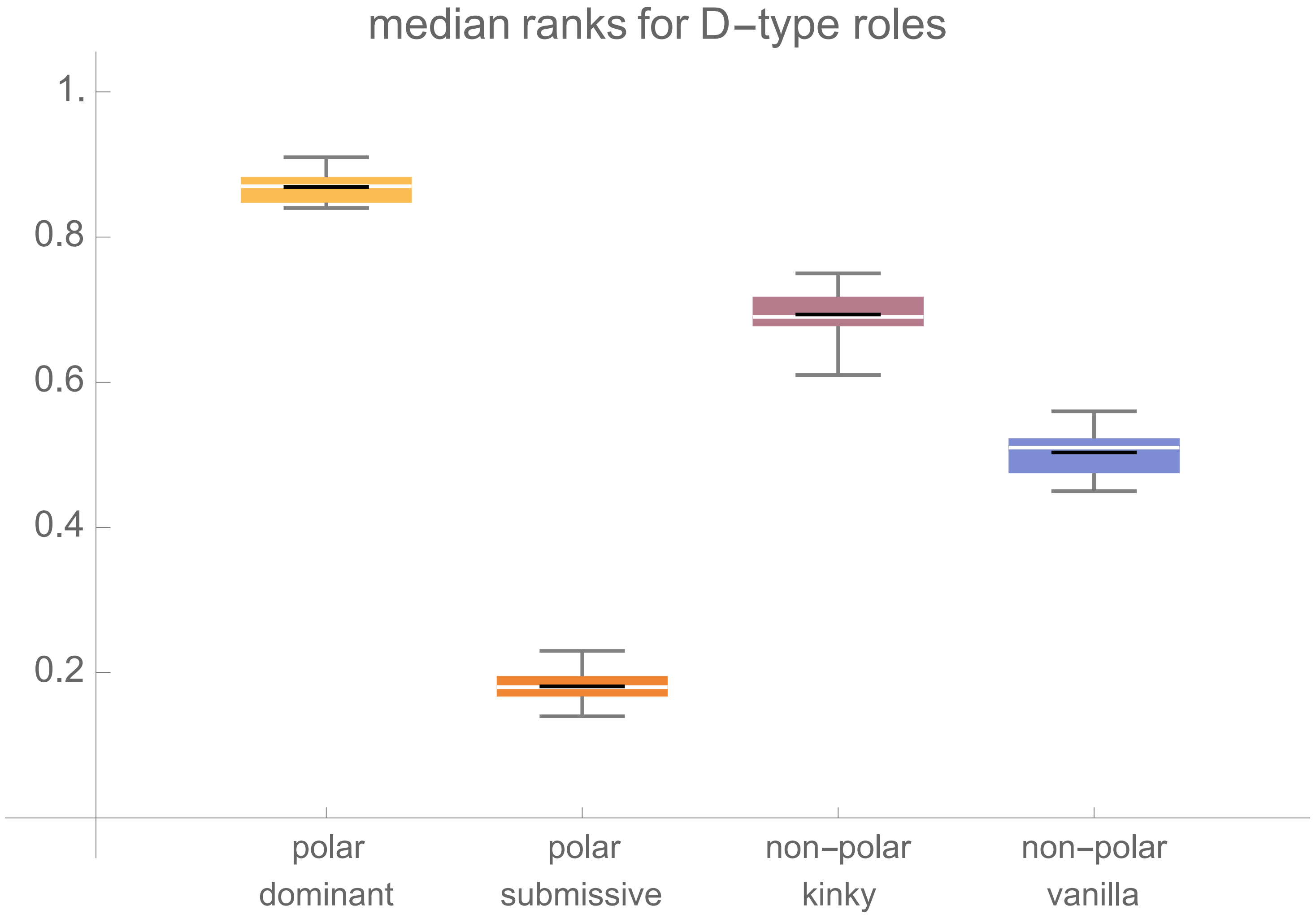


Stage 3: Multivariate analysis

Characterize the clusters

Clustering of survey respondents:

- polar dominant
- polar submissive
- non-polar kinky
- non-polar vanilla

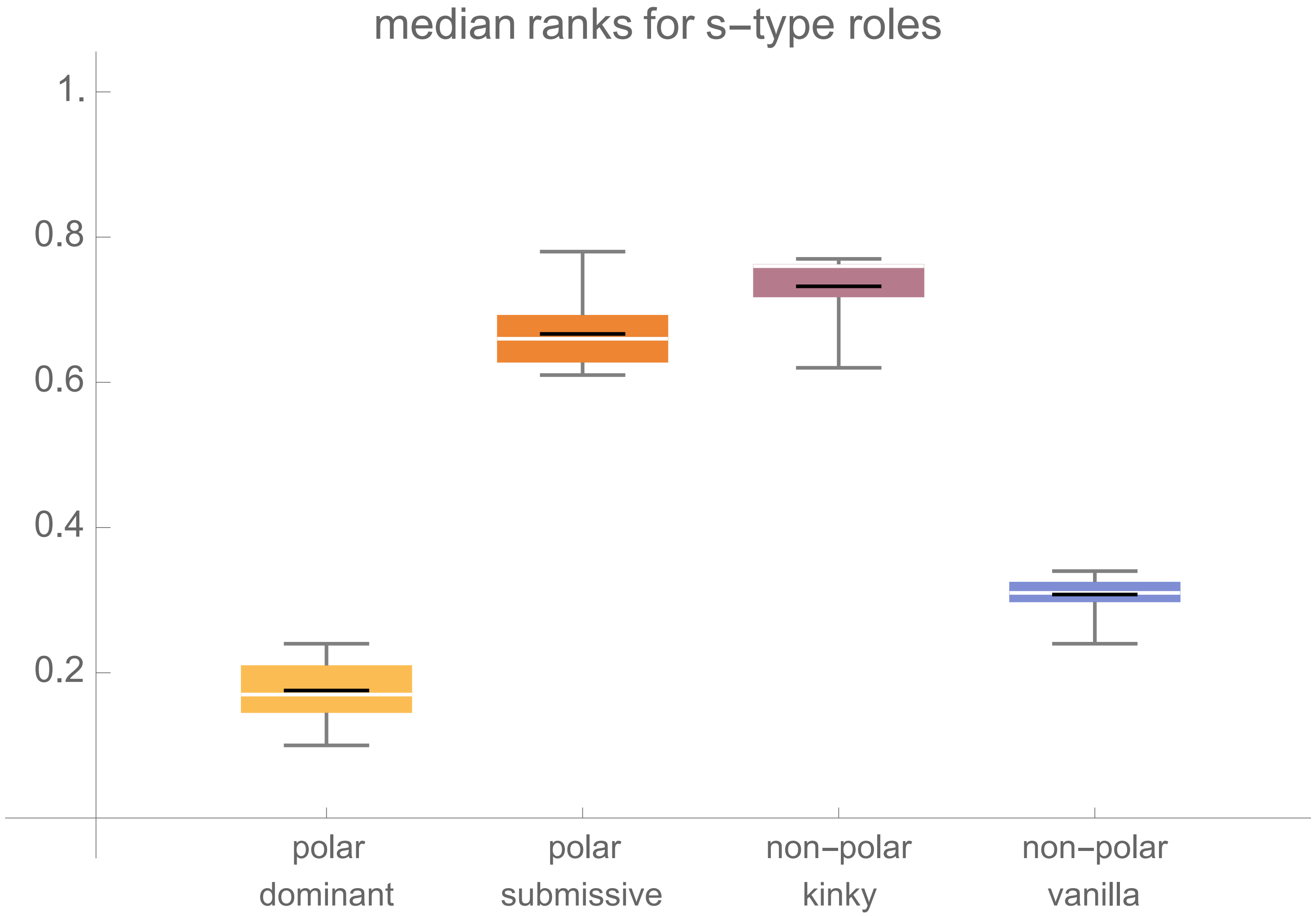


Stage 3: Multivariate analysis

Characterize the clusters

Clustering of survey respondents:

- polar dominant
- polar submissive
- non-polar kinky
- non-polar vanilla

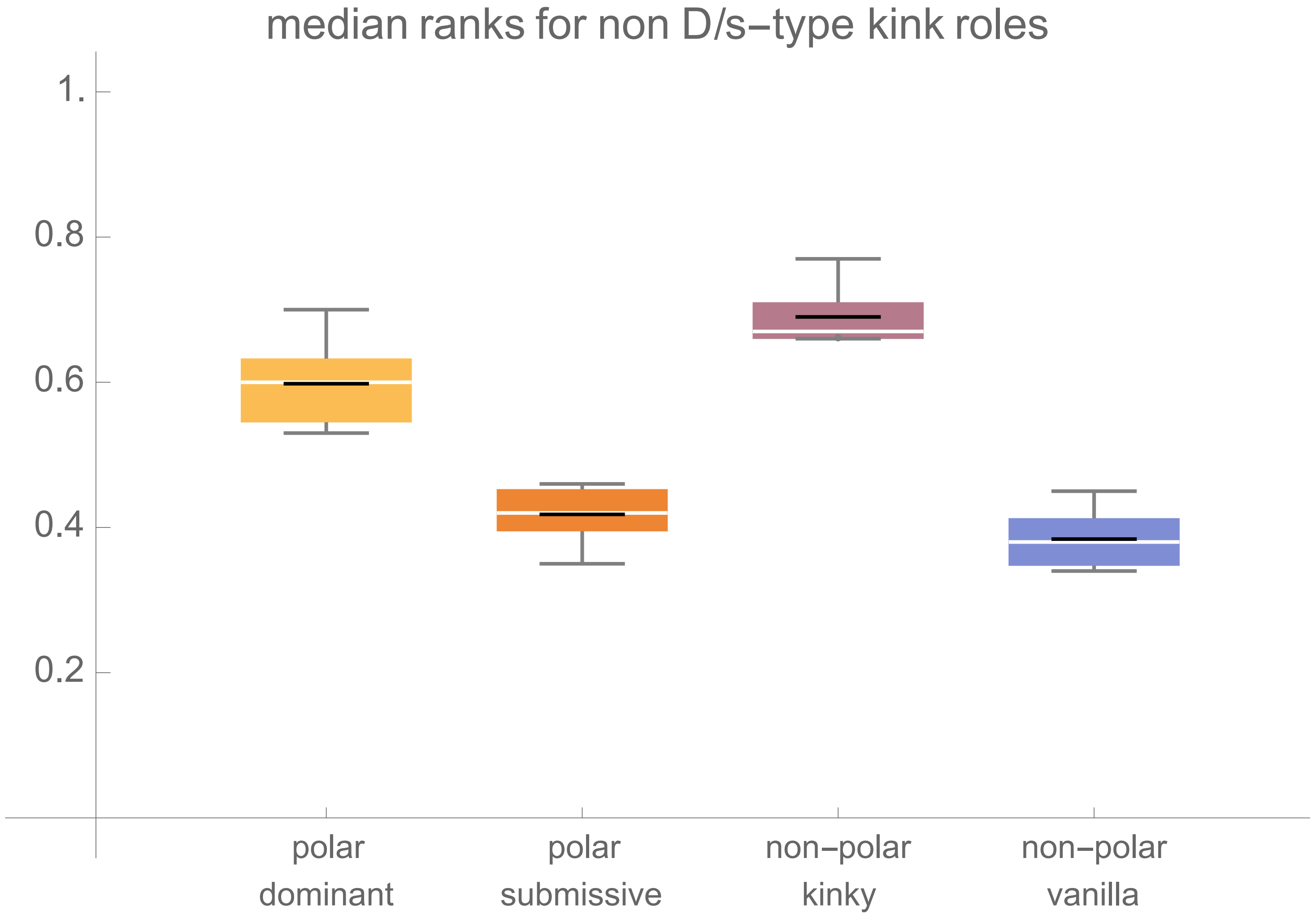


Stage 3: Multivariate analysis

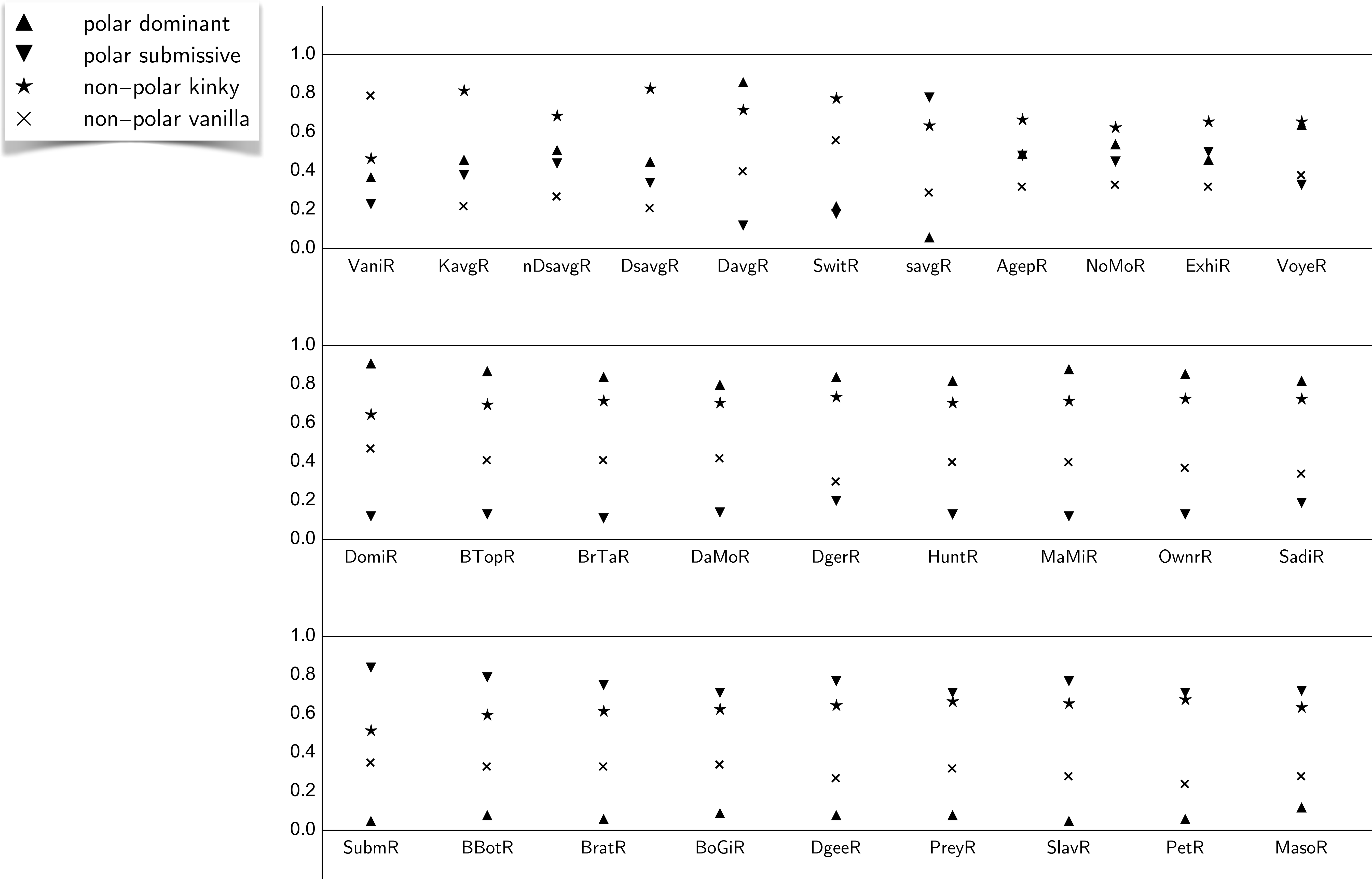
Characterize the clusters

Clustering of survey respondents:

- polar dominant
- polar submissive
- non-polar kinky
- non-polar vanilla



Stage 3: Multivariate analysis

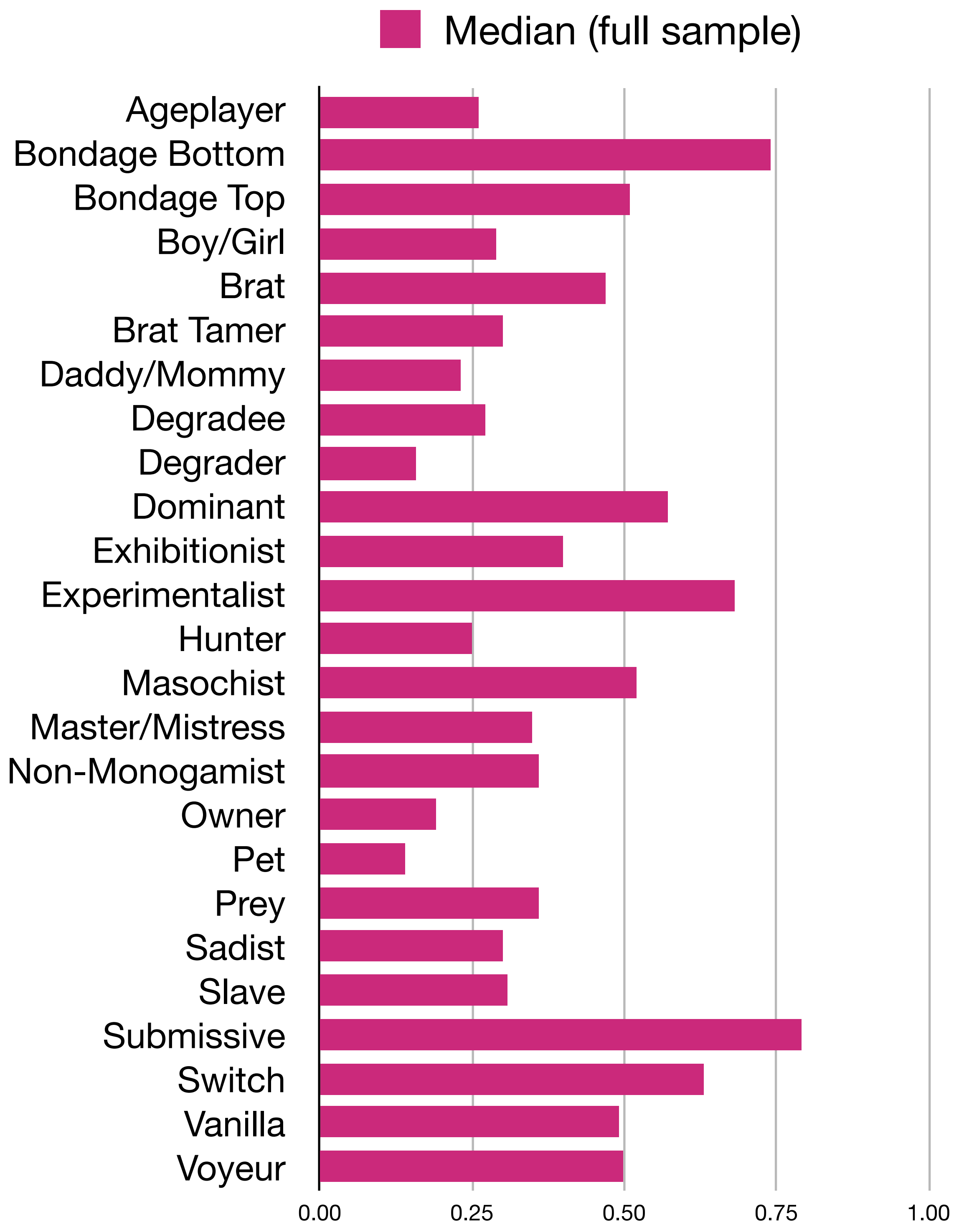


Toward a conceptual model

What topological shape should our model have?

As a practical matter, we often do a kind of dimensional reduction in everyday life.

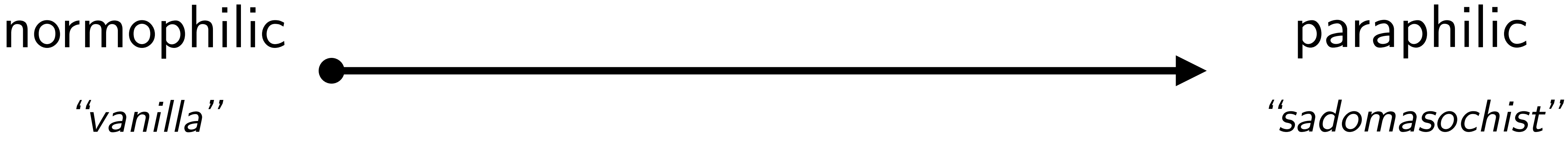
- discrete categories (0D)
- spectrum (1D)
- How many variables can *you* think about simultaneously varying without straining?



Toward a conceptual model

What topological shape should our model have?

Is it useful to conceive of "kinkiness" as one-dimensional?

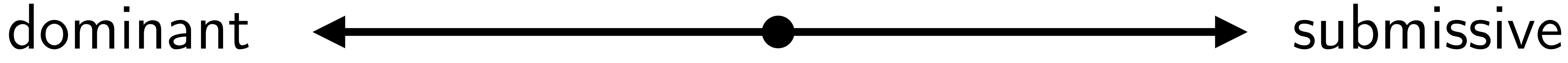

normophilic

"vanilla"

paraphilic

"sadomasochist"

...or
maybe...

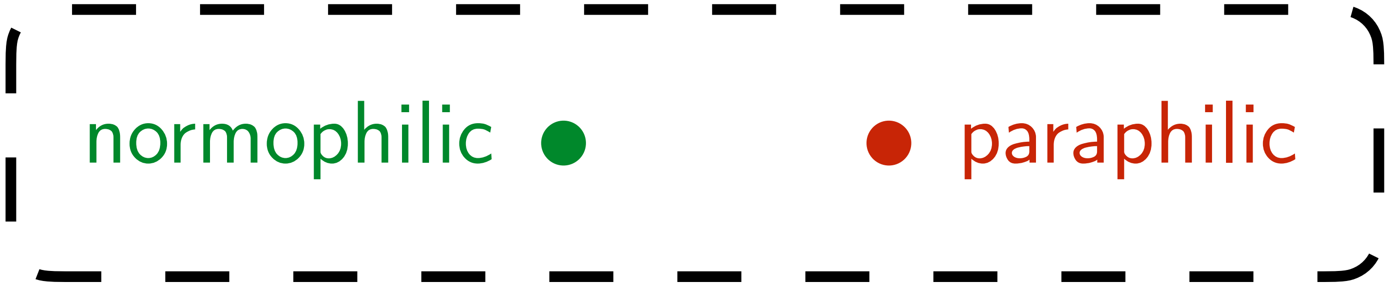

dominant

submissive

Toward a conceptual model

What topological shape should our model have?

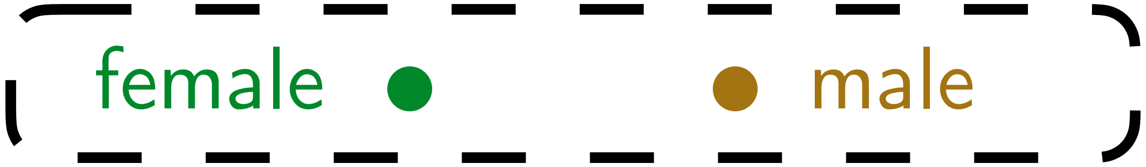
Maybe “kinkiness” is zero-dimensional?



Toward a conceptual model

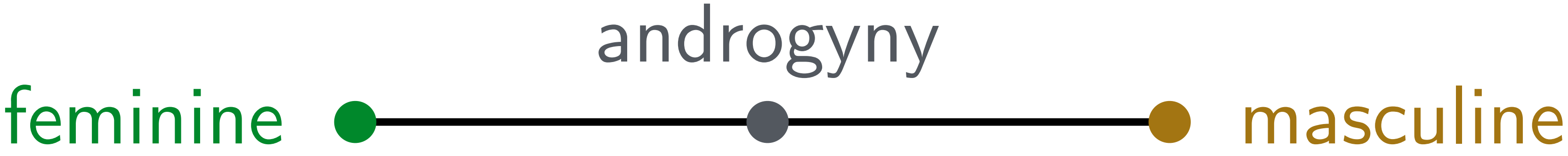
What topological shape should our model have?

"Don't be silly—no one believes sexual diversity is one-dimensional or zero-dimensional."



Topologically equivalent to $\mathbb{S}^0$ = *0-dimensional sphere*

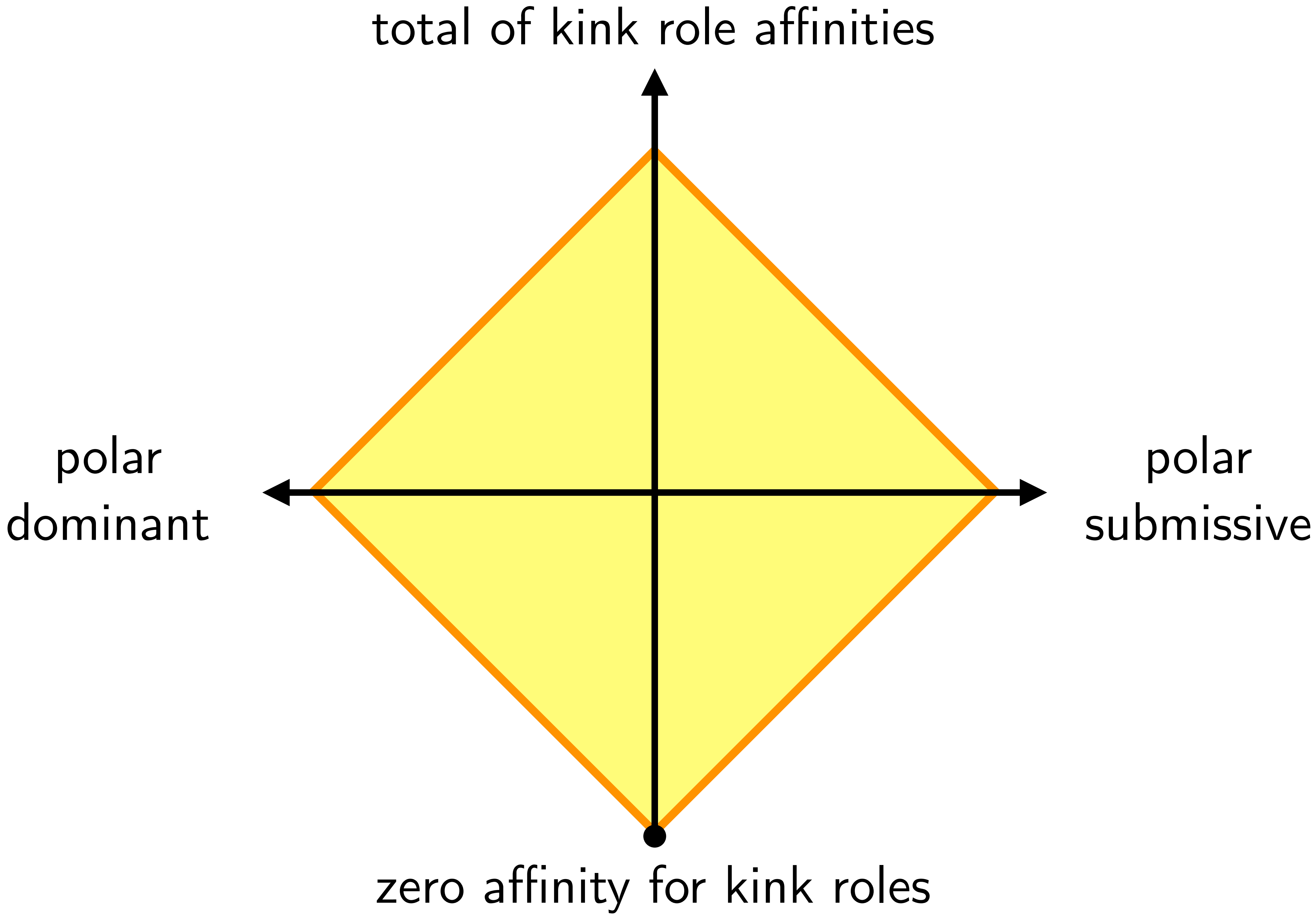
GET YELLED AT
GET BEAT UP



Topologically equivalent to $\mathbb{I}^1$ = *1-dimensional cell*

Toward a conceptual model

A two-dimensional model: ▯×▯



total of kink role affinities

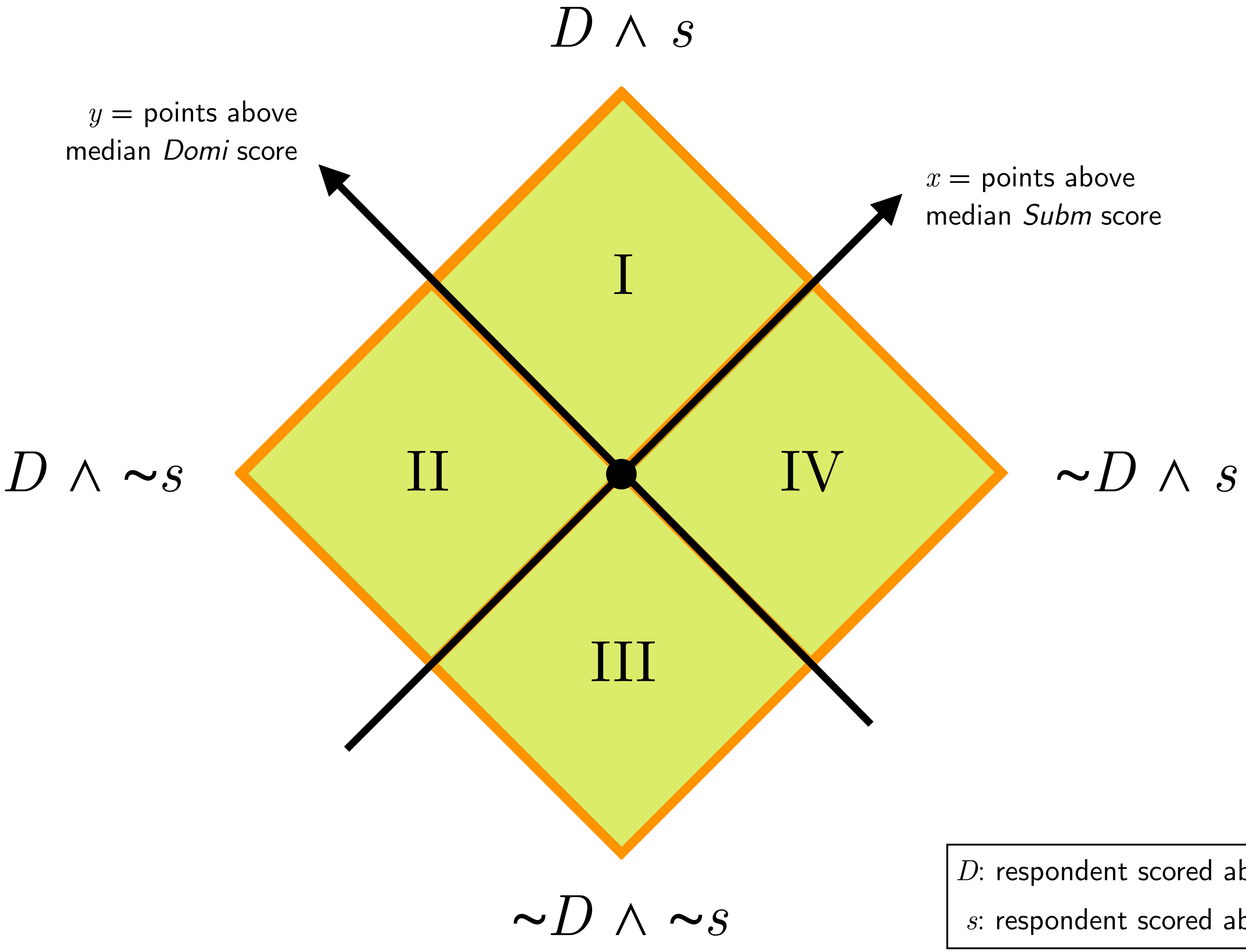
polar dominant

polar submissive

zero affinity for kink roles

Toward a conceptual model

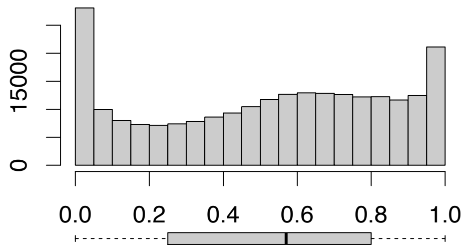
A two-dimensional model: ▯×▯


$D \wedge s$
$y =$ points above median *Domi* score
$x =$ points above median *Subm* score
I
II
IV
III
$D \wedge \sim s$
$\sim D \wedge s$
$\sim D \wedge \sim s$

D: respondent scored above median in *Domi*

s: respondent scored above median in *Subm*

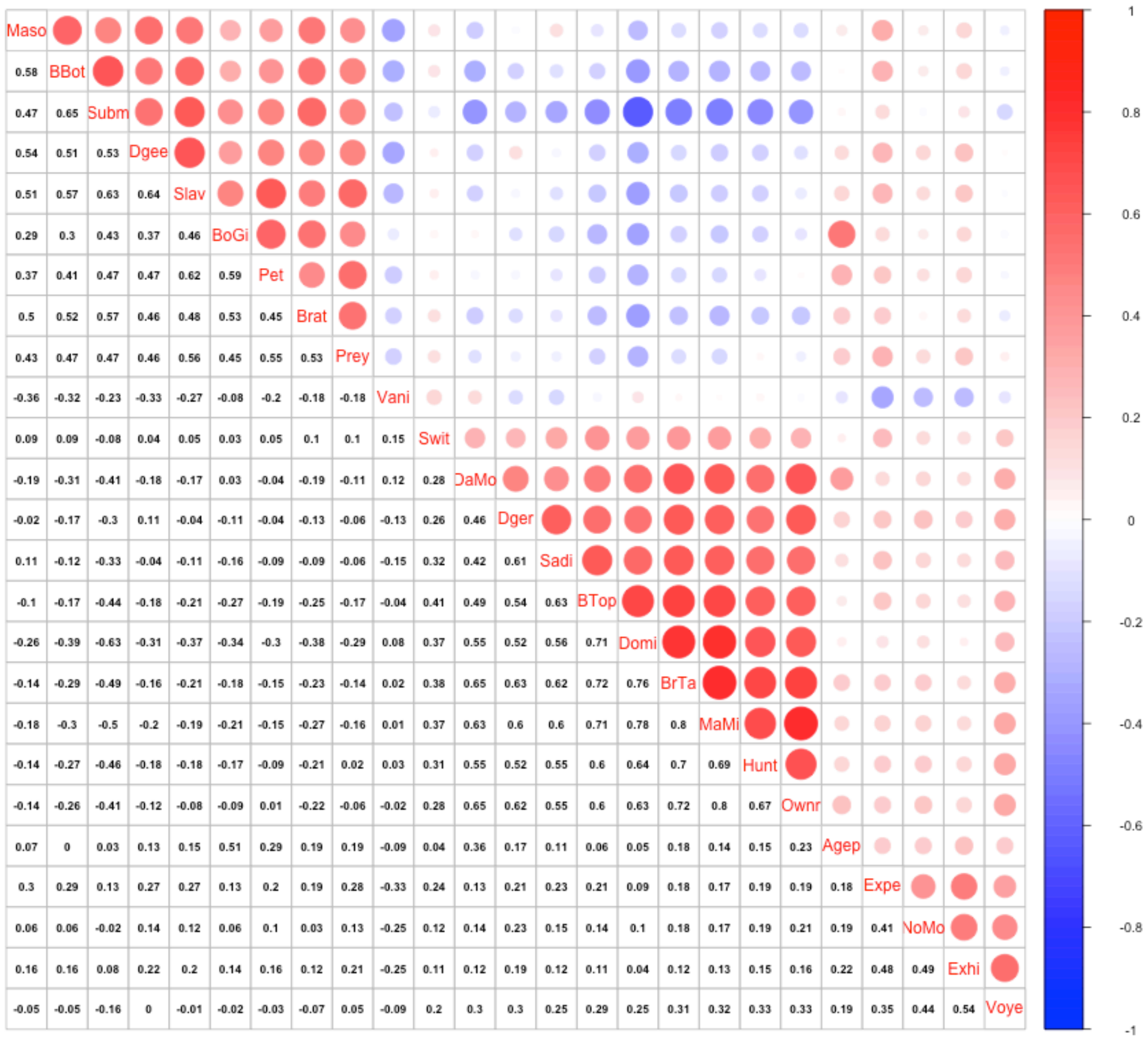
Summary of univariate and bivariate exploratory methodology



Summarize raw univariate distributions

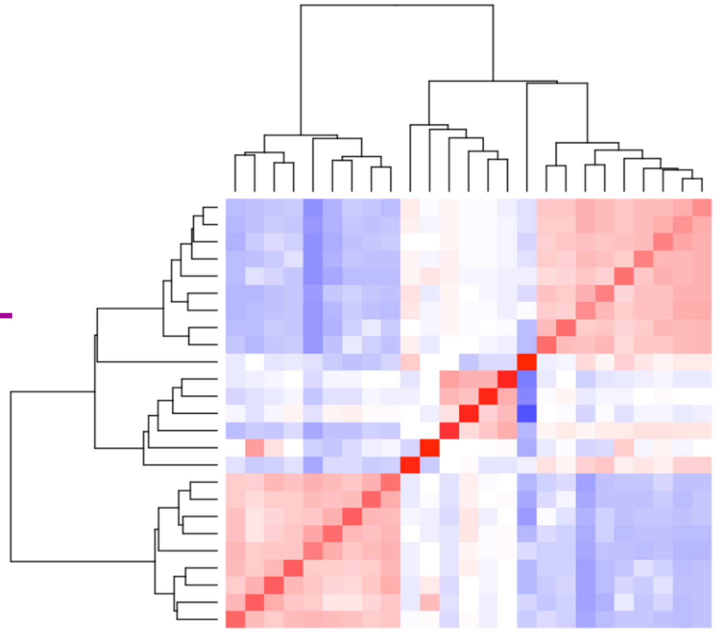
- 5-number summary
- classify distributions by shape

Normalize univariate distributions



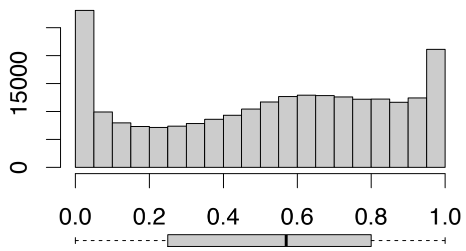
Bivariate correlations

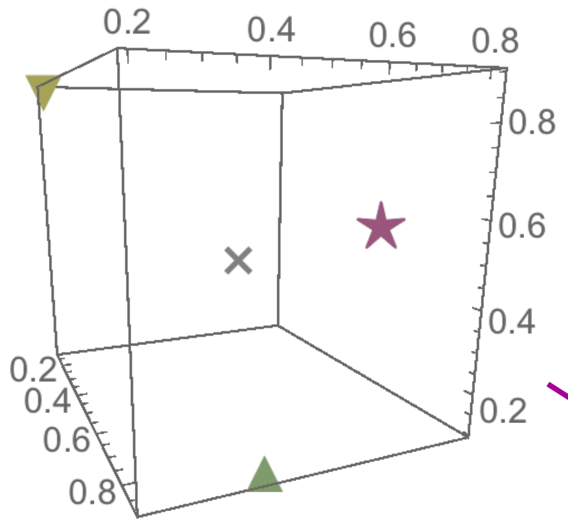
Clustering of variables

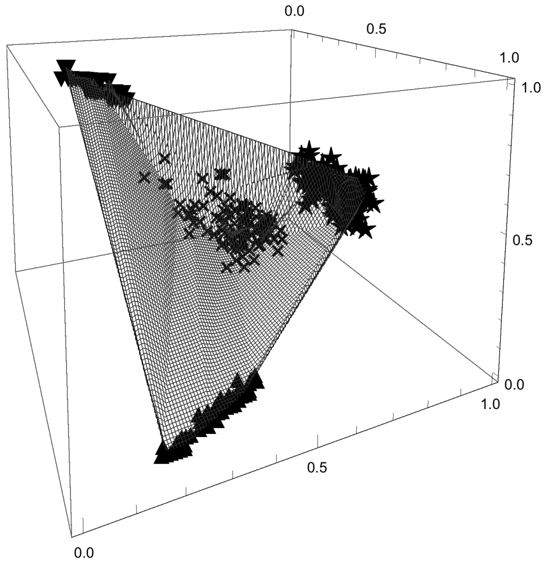


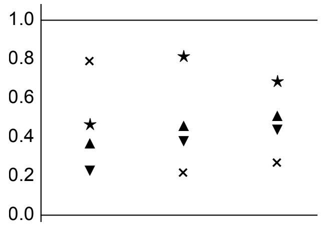
Characterize each cluster of variables

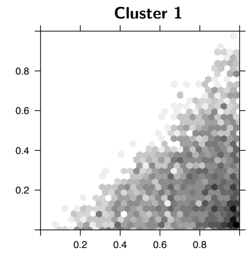
Summary of multivariate exploratory methodology



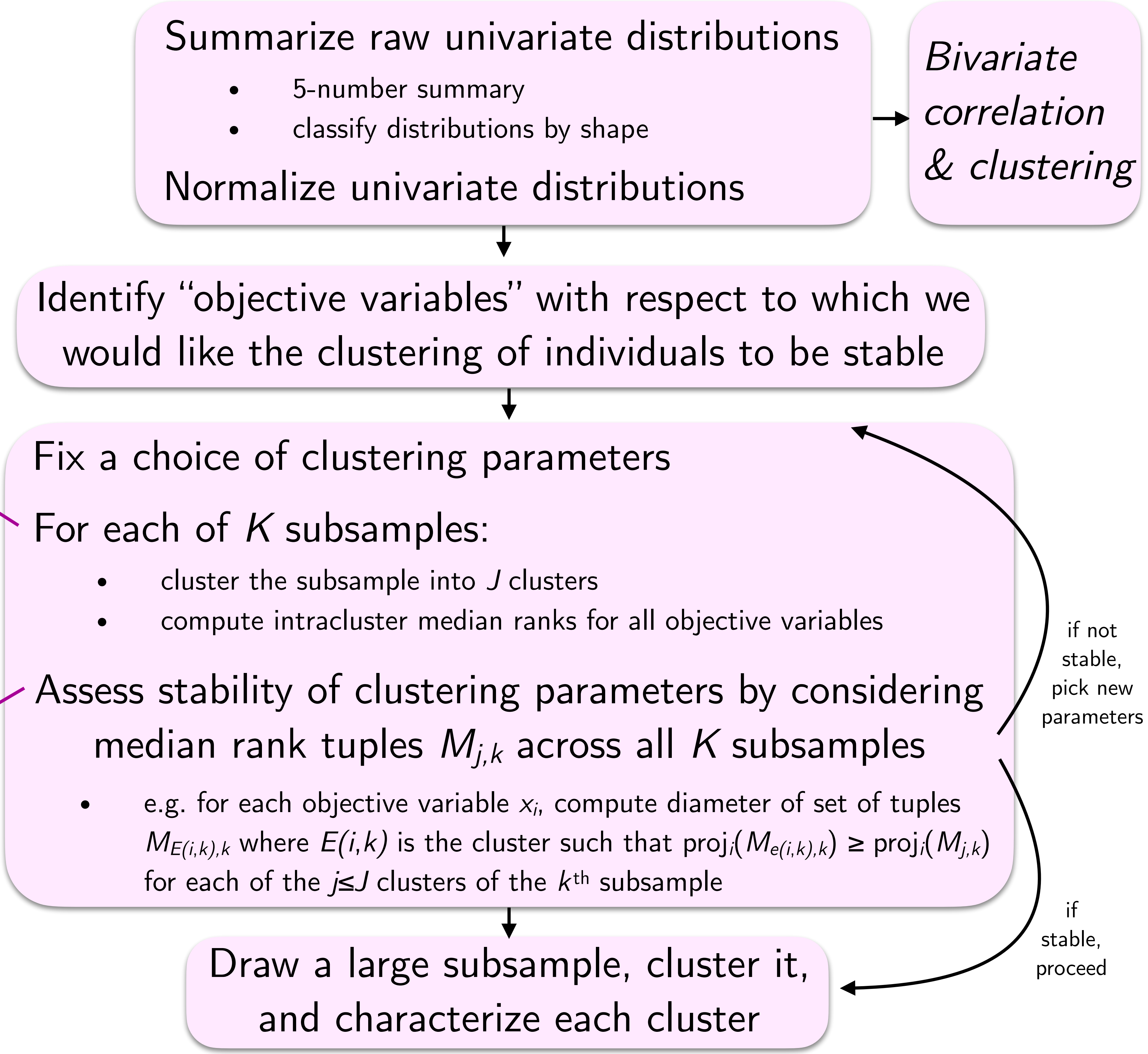








…



Summarize raw univariate distributions

- 5-number summary
- classify distributions by shape

Normalize univariate distributions

→ *Bivariate correlation & clustering*

Identify "objective variables" with respect to which we would like the clustering of individuals to be stable

Fix a choice of clustering parameters

For each of K subsamples:

- cluster the subsample into J clusters
- compute intracluster median ranks for all objective variables

Assess stability of clustering parameters by considering median rank tuples $M_{j,k}$ across all K subsamples

- e.g. for each objective variable x_i, compute diameter of set of tuples $M_{E(i,k),k}$ where $E(i,k)$ is the cluster such that $\mathrm{proj}_i(M_{e(i,k),k}) \geq \mathrm{proj}_i(M_{j,k})$ for each of the $j \leq J$ clusters of the kth subsample

if not stable, pick new parameters

if stable, proceed

Draw a large subsample, cluster it, and characterize each cluster