Nonparametric MANOVA via Independence Testing

Sambit Panda

Cencheng Shen, Ronan Perry, Jelle Zorn, Antoine Lutz, Carey E. Priebe, Joshua T. Vogelstein

19 January, 2022

Johns Hopkins University

Department of Biomedical Engineering

INSTITUTE *for* COMPUTATIONAL MEDICINE

JOHNS HOPKINS BIOMEDICAL ENGINEERING

Motivation

- Understand the relationship between k groups (*i.e.* control vs. disease)
- Question: Are they related? How?

One often desires to test groups

$\boldsymbol{U}$	$\boldsymbol{V}$
my grass	neighbor’s grass
human brain connectivity	alien brain connectivity
control	disease
cancer risk group 1	cancer risk group 2

One often desires to test groups

$\boldsymbol{U}$	$\boldsymbol{V}$
my grass	neighbor’s grass
human brain connectivity	alien brain connectivity
control	disease
cancer risk group 1	cancer risk group 2
any group	any other group

Statistics Background

- X is a random variable (some measurement)
- F_X is the distribution of X
- This means $F_X(a) = P(X \leq a)$
- This is denoted:

$$X \sim F_X$$

Statistics Background

- For two random variables X and Y, F_{XY} is called the joint distribution
- This means $F_{XY}(a, b) = P(X \leq a \text{ and } Y \leq b)$

  or,

$$(X, Y) \sim F_{XY}$$

Informal Definition of Hypothesis Testing

- **Null Hypothesis**: The conventional belief about a phenomenon of interest, written H_0.
- **Alternative Hypothesis**: An alternate belief about the same phenomenon, written H_A.
- **p-value**: The probability (under the null) of measurements more extreme than what was observed.

Formal Definition of *K*-Sample Testing

$$U_i^j \sim F_j, \qquad j \in 1, \ldots, k, \qquad i \in 1, \ldots, n_j$$

$$H_0 : F_1 = F_2 = \cdots = F_k$$

$$H_A : \exists\, j \neq j' \; s.t. \; F_j \neq F_{j'}$$

Note: These ideas and notation generalize for multivariate X and Y.

Outline

1. Intuition
2. Simulations
3. Multiway and Multilevel
4. Real Data
5. Conclusion

Intuition

Intuitive Desiderata of Testing Procedure

- Performant under *any* distribution
  - low- and high-dimensional
  - Euclidean and structured data (eg, sequences, images, networks, shapes)
  - linear and nonlinear relationships
- Is computational efficient

Provides a tractable algorithm that addresses the motivating question:

Are they related?

Analysis of Variance (ANOVA)

$$MST = \frac{\sum_{i=1}^{k} \left(T_i^2 / n_i \right) - G^2 / n}{k - 1}$$

$$MSE = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}^2 - \sum_{i=1}^{k} \left(T_i^2 / n_i \right)}{n - k}$$

$$ANOVA = \frac{MST}{MSE}$$

Multivariate ANOVA (MANOVA)

$$\mathbf{W} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_{i\cdot})(\mathbf{x}_{ij} - \bar{\mathbf{x}}_{i\cdot})^T$$

$$\mathbf{B} = \sum_{i=1}^{k} n_i (\bar{\mathbf{x}}_{i\cdot} - \bar{\mathbf{x}}_{\cdot\cdot})(\bar{\mathbf{x}}_{i\cdot} - \bar{\mathbf{x}}_{\cdot\cdot})^T$$

$$MANOVA = \sum_{i=1}^{s} \frac{\lambda_i}{1 + \lambda_i} = \mathrm{tr}\left(\mathbf{B}(\mathbf{B} + \mathbf{W})^{-1}\right)$$

MANOVA Assumptions

- Data is derived from a multivariate Gaussian distribution
- Each group has the same covariance matrix

There must be a better test out there

Slight Tangent - Independence Testing

- X and Y are independent if neither contains information about the other
- In other words,

$$F_{XY} = P(X \leq a \text{ and } Y \leq b) = P(X \leq a) \times P(Y \leq b) = F_X F_Y$$

$$F_{XY} = F_X F_Y$$

Note: These ideas and notation generalize for multivariate X and Y.

Distance Correlation (Dcorr)

$$\widehat{Dcov}_{xy} = \frac{1}{n^2}\operatorname{tr}(\mathbf{H}\mathbf{D}^{\mathbf{x}}\mathbf{H}\mathbf{D}^{\mathbf{y}}\mathbf{H})$$

$$\widehat{Dcorr}_{xy} = \frac{\widehat{Dcov}_{xy}}{\sqrt{\widehat{Dcov}_{xx} \times \widehat{Dcov}_{yy}}}$$

$$\mathbf{C}^{\mathbf{x}}_{ij} = \mathbb{I}_{i \neq j}\left(\mathbf{D}^{\mathbf{x}}_{ij} - \frac{1}{n-2}\sum_{t=1}^{n}\mathbf{D}^{\mathbf{x}}_{it} - \frac{1}{n-2}\sum_{t=1}^{n}\mathbf{D}^{\mathbf{x}}_{tj} + \frac{1}{(n-1)(n-2)}\sum_{t=1}^{n}\mathbf{D}^{\mathbf{x}}_{tt}\right)$$

$$Dcov_{xy} = \frac{1}{n(n-3)}\operatorname{tr}(\boldsymbol{C}^{\mathbf{x}}\boldsymbol{C}^{\mathbf{y}})$$

$$Dcorr_{xy} = \frac{Dcov_{xy}}{\sqrt{Dcov_{xx} \times Dcov_{yy}}}$$

Multiscale Graph Correlation (MGC)

- Compute local Dcorr **at all scales**
- Find scale with **max** smoothed test statistic
- Permutation test to determine p-value

Kernel Mean Embedding Random Forest (KMERF)

- Train random forest on X, compute kernel matrix
- Transform similarity kernel matrix to distance matrix
- Permutation test to determine p-value

Great, what now?

- Can reduce the k-sample testing problem to the independence problem

$$\mathbf{x} = \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_k \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} \mathbf{1}_{n_1 \times 1} & \mathbf{0}_{n_1 \times 1} & \cdots & \mathbf{0}_{n_1 \times 1} \\ \mathbf{0}_{n_2 \times 1} & \mathbf{1}_{n_2 \times 1} & \cdots & \mathbf{0}_{n_2 \times 1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{n_k \times 1} & \mathbf{0}_{n_k \times 1} & \cdots & \mathbf{1}_{n_k \times 1} \end{bmatrix}$$

- Run any independence test

- ***Note: This process does not add any additional computational complexity to the independence testing algorithm***

Simulations

Definitions

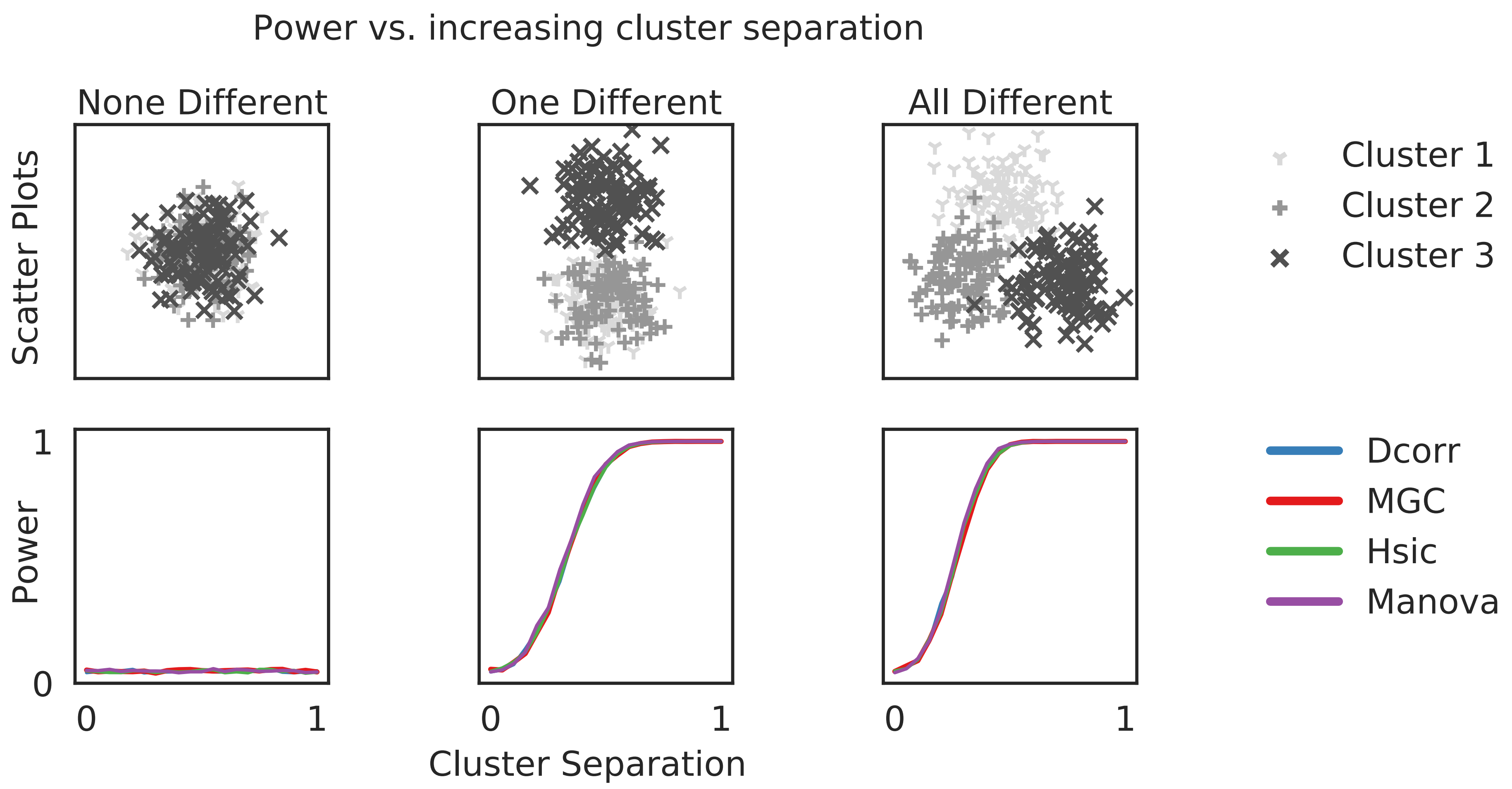
- **power** is the probability of rejecting the null when the alternative is true

$\beta_n(t)$: power of test statistic t given n samples

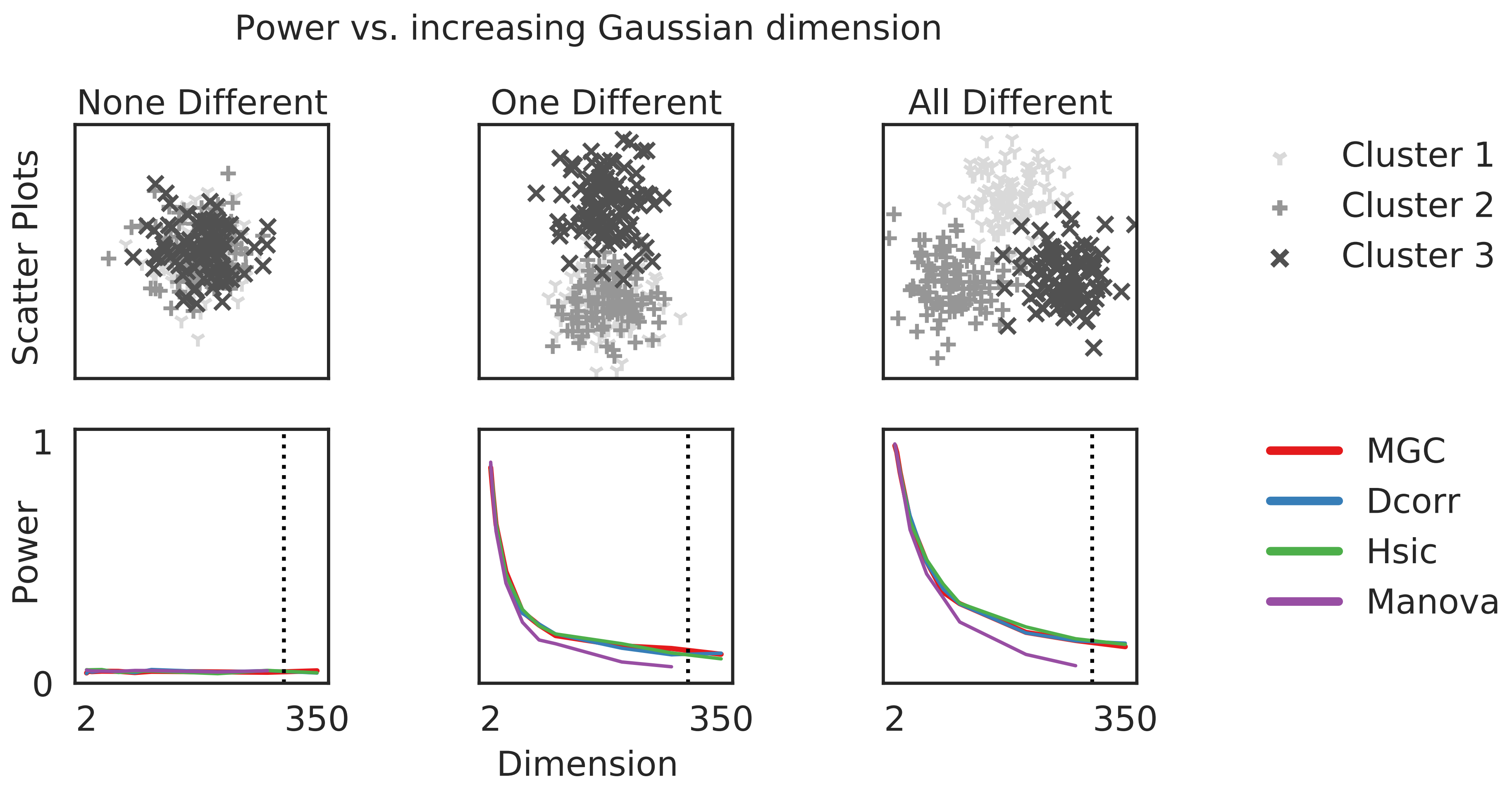
- **relative power** power of one approach minus power of another

$$\beta_n(t) - \beta_n(manova)$$

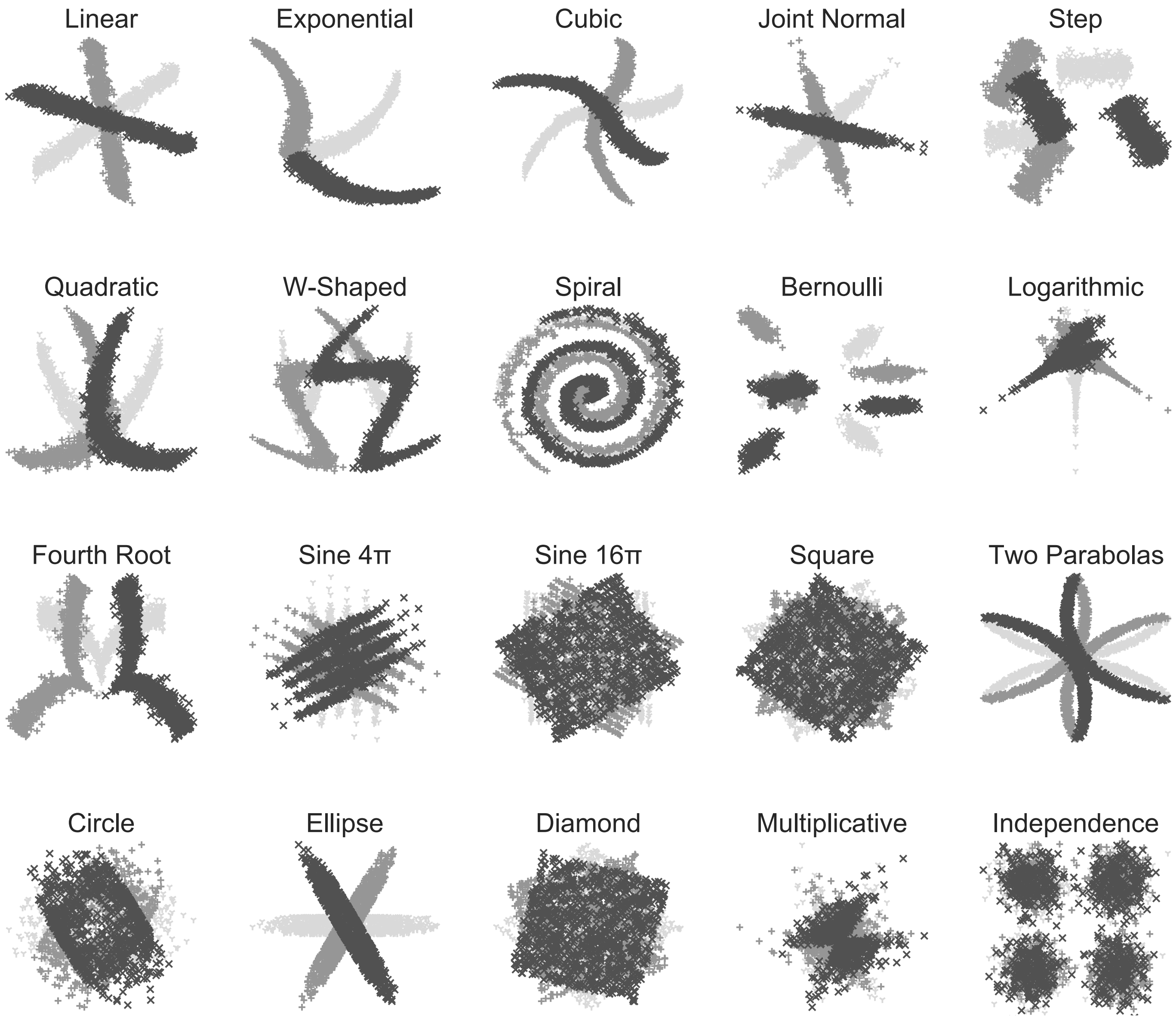
Optimal Settings for MANOVA (1D)

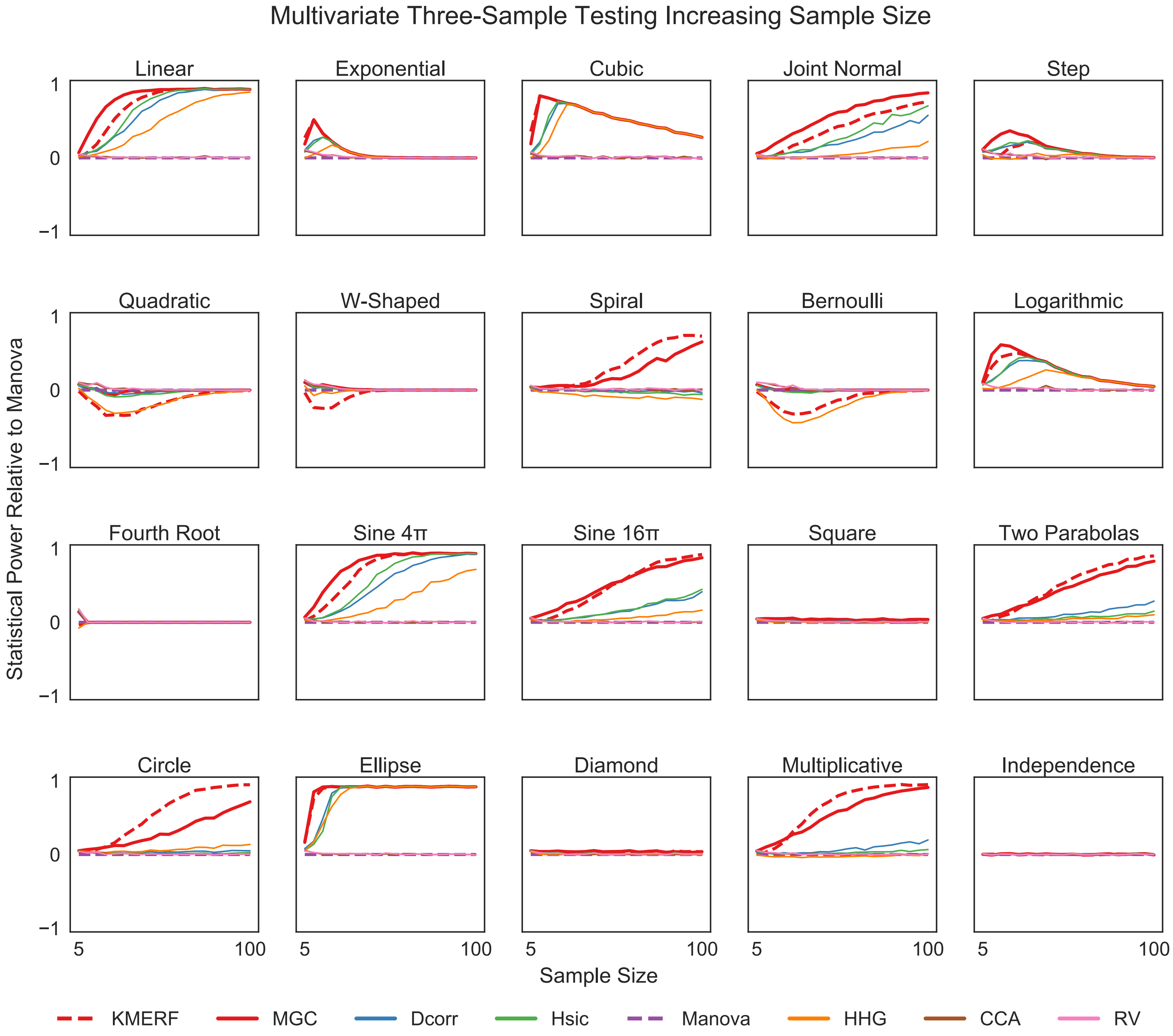


Optimal Settings for MANOVA (HD)



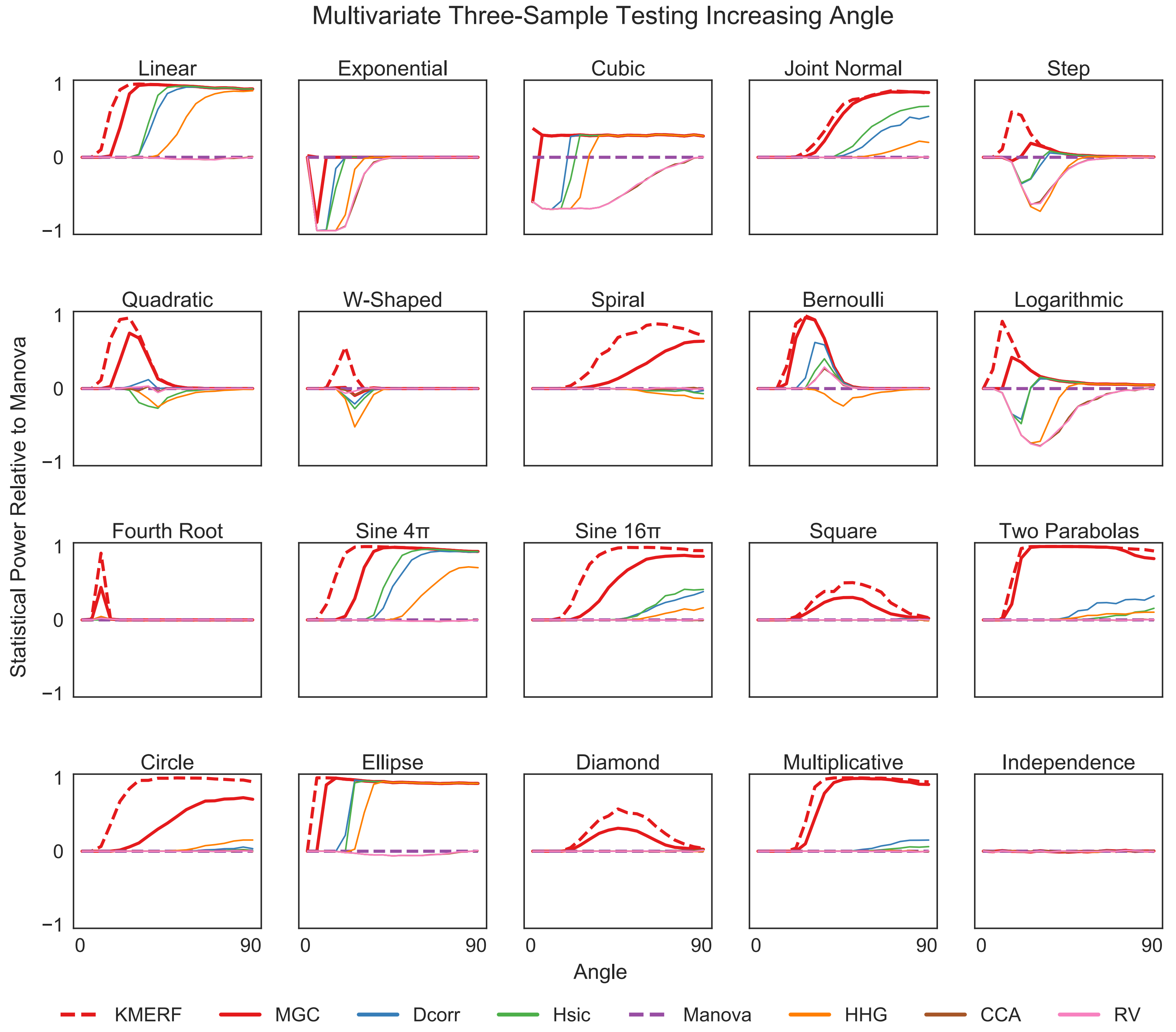
20 Different Functions (2D version)





Multivariate Three-Sample Testing Increasing Sample Size
Linear
Exponential
Cubic
Joint Normal
Step
Quadratic
W-Shaped
Spiral
Bernoulli
Logarithmic
Fourth Root
Sine 4π
Sine 16π
Square
Two Parabolas
Circle
Ellipse
Diamond
Multiplicative
Independence
Statistical Power Relative to Manova
Sample Size
1
0
−1
5
100
KMERF
MGC
Dcorr
Hsic
Manova
HHG
CCA
RV


Multivariate Three-Sample Testing Increasing Dimension
Linear
Exponential
Cubic
Joint Normal
Step
Quadratic
W-Shaped
Spiral
Bernoulli
Logarithmic
Fourth Root
Sine 4π
Sine 16π
Square
Two Parabolas
Circle
Ellipse
Diamond
Multiplicative
Independence
Statistical Power Relative to Manova
Dimension
1
0
−1
3
10
KMERF
MGC
Dcorr
Hsic
Manova
HHG
CCA
RV

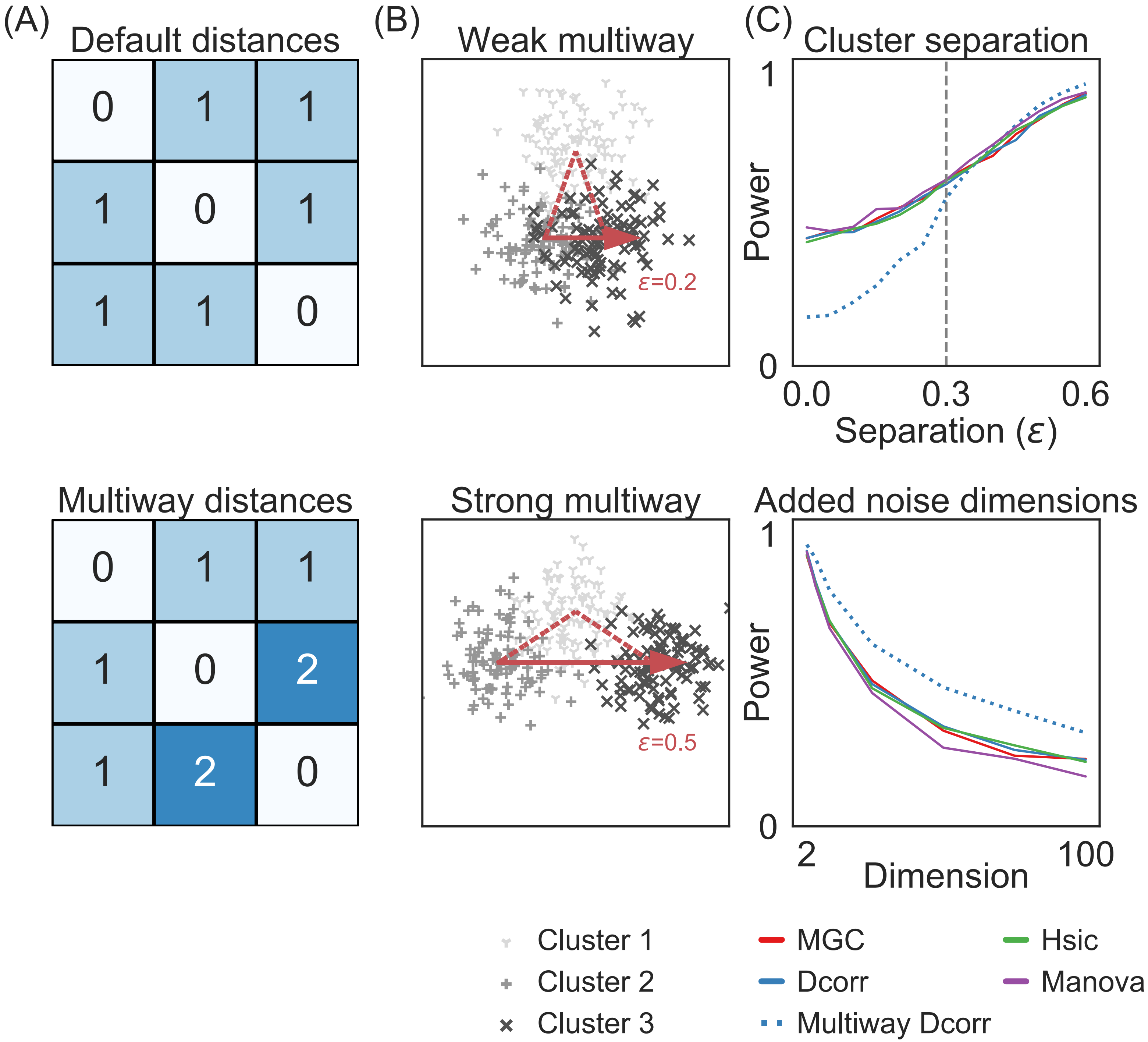


Multivariate Three-Sample Testing Increasing Angle
Linear
Exponential
Cubic
Joint Normal
Step
Quadratic
W-Shaped
Spiral
Bernoulli
Logarithmic
Fourth Root
Sine 4π
Sine 16π
Square
Two Parabolas
Circle
Ellipse
Diamond
Multiplicative
Independence
Statistical Power Relative to Manova
1
0
−1
0
90
Angle
KMERF
MGC
Dcorr
Hsic
Manova
HHG
CCA
RV

Multiway and Multilevel

Multiway Tests

- **Multiway**: More than one treatment group
- Instead of one-hot encoding, add 1’s columns of label matrix



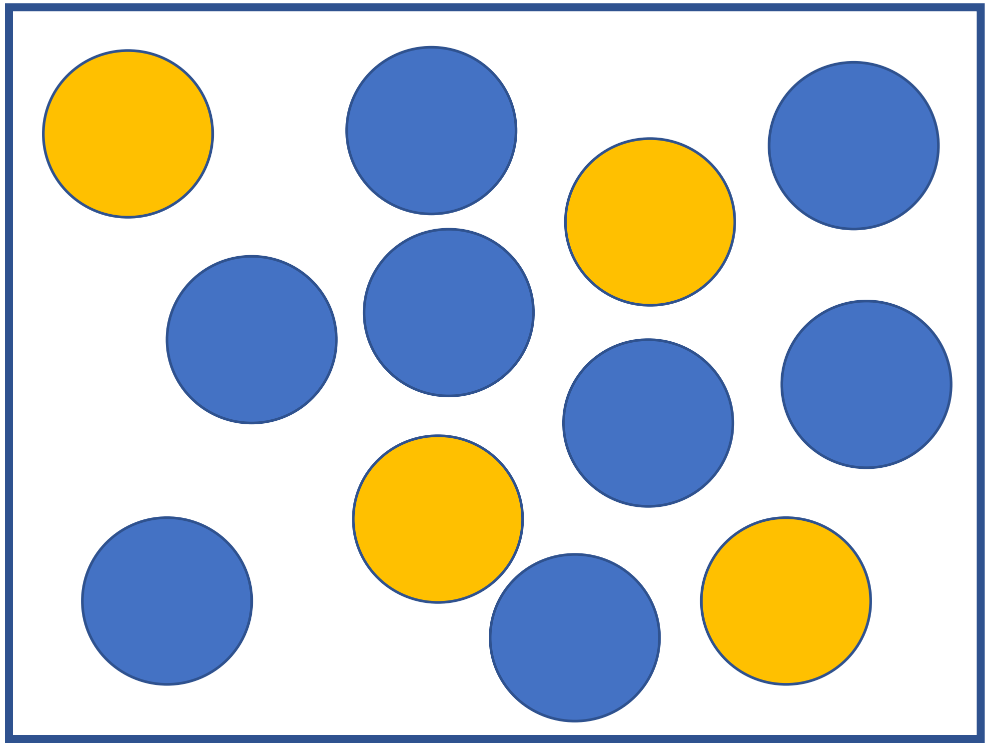
(A)
Default distances
0 1 1
1 0 1
1 1 0
Multiway distances
0 1 1
1 0 2
1 2 0
(B)
Weak multiway
ε=0.2
Strong multiway
ε=0.5
(C)
Cluster separation
Power
Separation (ε)
0.0 0.3 0.6
0 1
Added noise dimensions
Power
Dimension
2 100
0 1
Cluster 1
Cluster 2
Cluster 3
MGC
Dcorr
Multiway Dcorr
Hsic
Manova

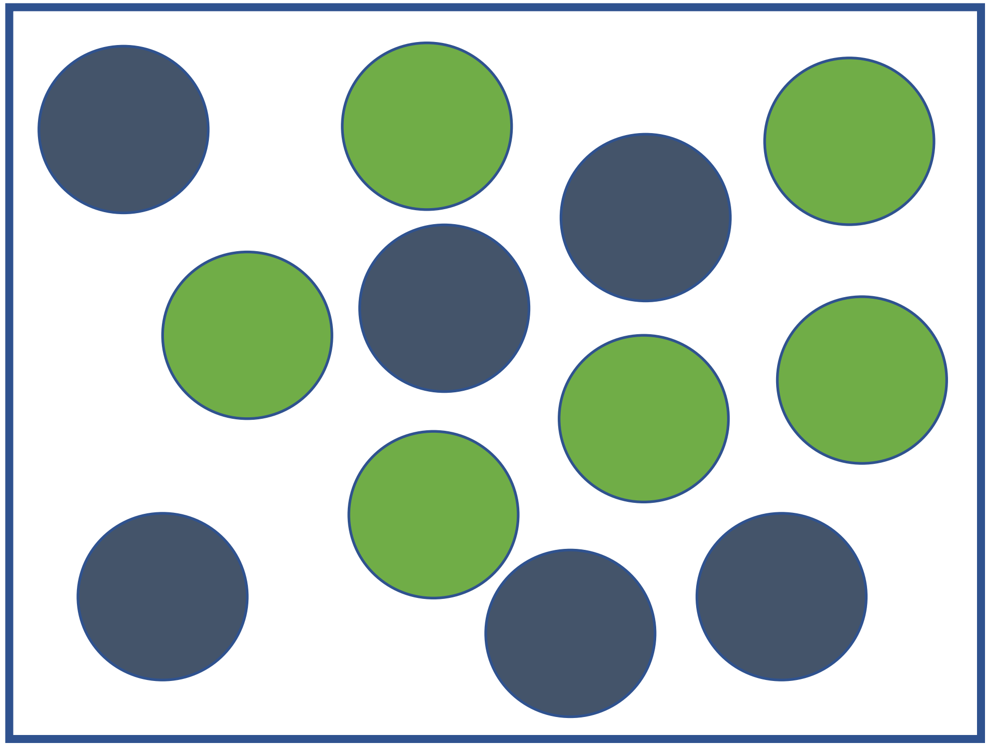
Multilevel Tests

- **Multilevel**: Samples are not always exchangeable with one another
- Need block permutation

Multilevel Tests

- **Multilevel**: Samples are not always exchangeable with one another

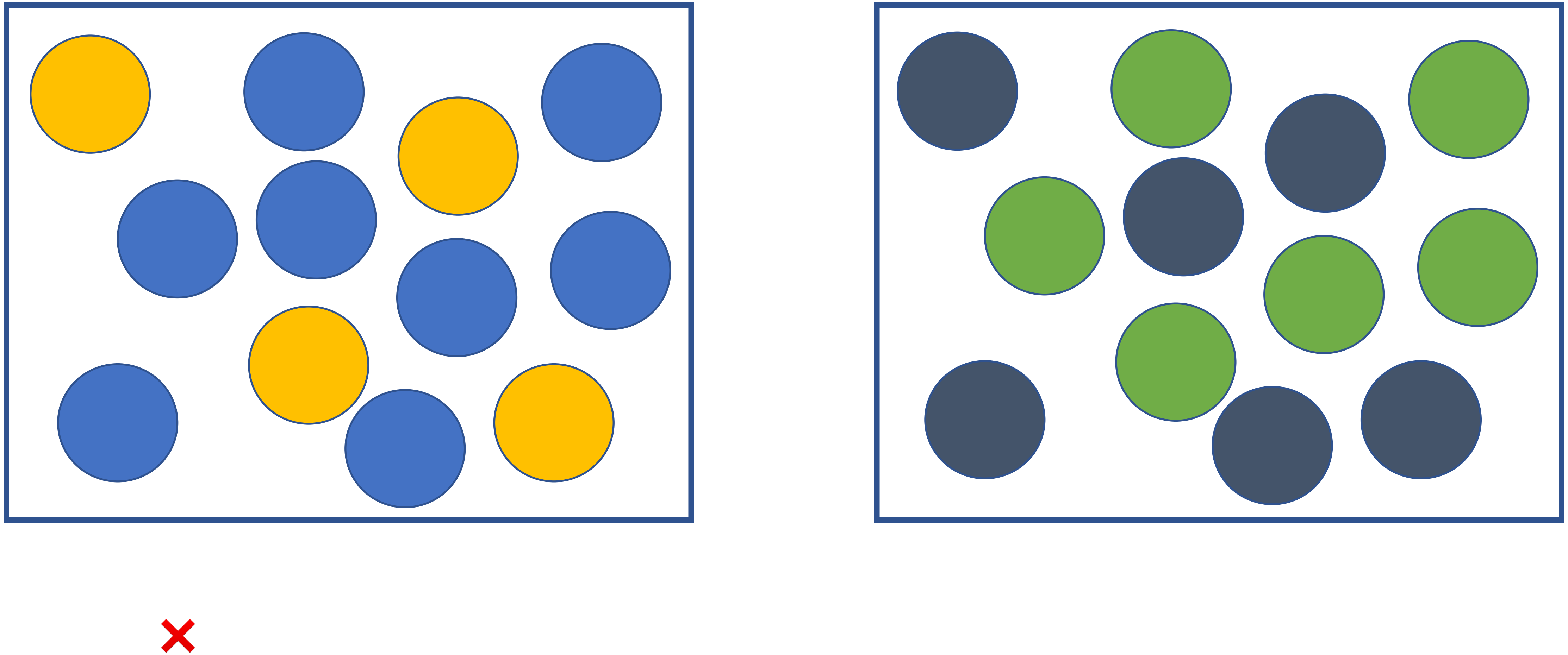




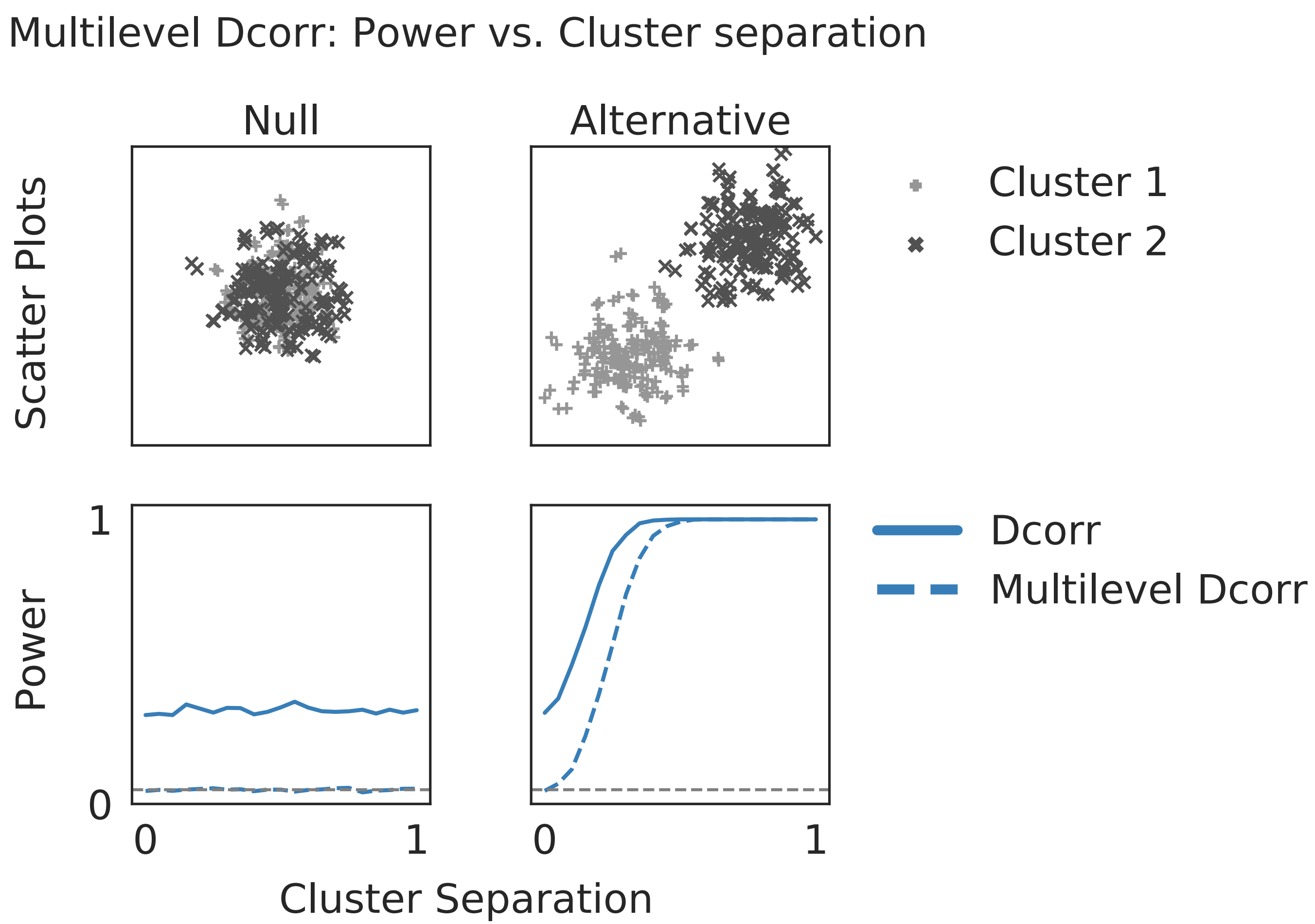
✅

Multilevel Tests

- **Multilevel**: Samples are not always exchangeable with one another



Multilevel Tests

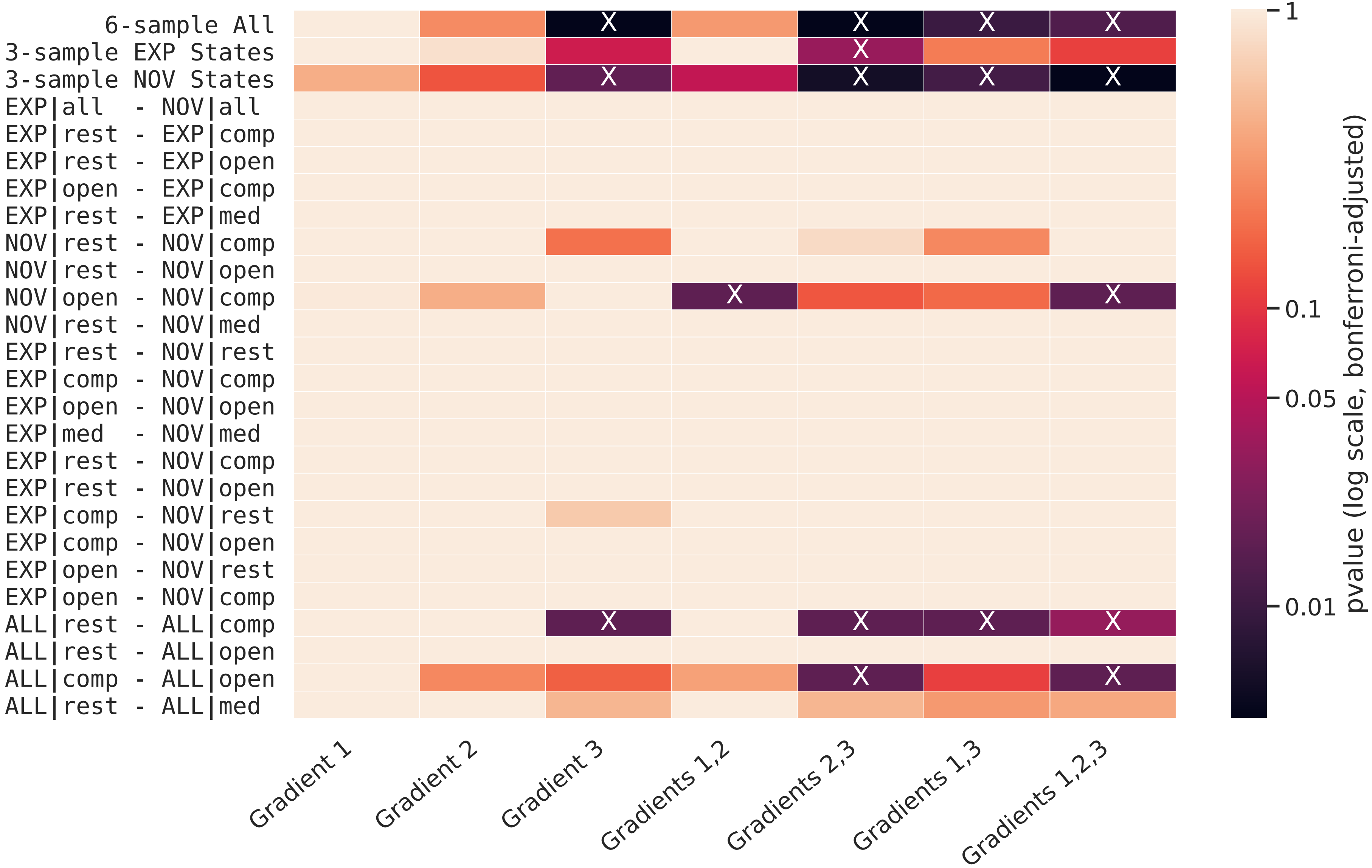


Real Data

The Procedure

- Data: 75 subjects – 28 experienced and 47 novice meditators
- 3 Recording sessions for each meditator
- Computed gradients and tested for difference between traits and novice
- This is a **multilevel and multiway test**

3^{rd} Embedding shows significance



Conclusions

- Presented several new k-sample tests using our framework
- At a simulation setting that fulfills MANOVA assumptions, our implementation performs as well or better
- Multiway tests give additional power when strong multiway effect is suspected
- Multilevel tests can now be performed

Next Steps

- All algorithms can be found in the hyppo package
    - Documentation
    - Install
    - Tutorials
- Paper

Email | Website | Twitter

Acknowledgements

- **Joshua Vogelstein, Cencheng Shen**: Theory, and paper writing
- **Ronan Perry**: Multiway, Multilevel, and Real Data
- **Jelle Zorn, Antoine Lutz**: Raw real data
- **Carey E. Priebe**: Theory
- **Russell Lyons, Minh Tang, Ronak Mehta, Eric Bridgeford**: Review
- …and the rest of the NeuroData Lab

DARPA

NIH National Institutes of Health

NSF

Microsoft Research

Questions?