8-2022

Extensions to the Syrjala Test with Eye-Tracking Data Analysis Applications in R

Eric D. McKinney
Utah State University

EXTENSIONS TO THE SYRJALA TEST WITH EYE-TRACKING DATA ANALYSIS APPLICATIONS IN R

by

Eric D. McKinney

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Mathematical Sciences

Approved:

Jürgen Symanzik, Ph.D. Major Professor	Daniel Coster, Ph.D. Committee Member
D. Richard Cutler, Ph.D. Committee Member	John R. Stevens, Ph.D. Committee Member
Breanna Studenka, Ph.D. Committee Member	Janis L. Boettinger, Ph.D. Vice Provost

UTAH STATE UNIVERSITY
Logan, Utah

2022

ABSTRACT

Extensions to the Syrjala Test with Eye-Tracking Data Analysis Applications in R

by

Eric D. McKinney, Doctor of Philosophy

Utah State University, 2022

Major Professor: Jürgen Symanzik, Ph.D.
Department: Mathematics and Statistics

This dissertation introduces a series of new two-sample tests of distributional equality. The new tests are a generalization of the Syrjala (1996) test and make use of both rotations and toroidal shifts of the data. The new tests exhibit stability across a variety of explored test statistics. The test which employs both rotations and toroidal shifts overcomes many of the limitations of the original Syrjala test, and can be used for a variety of two-sample continuous bivariate data. Furthermore, several of the new tests are shown to be competitive or better than other modern techniques via simulation experiments. One of the new tests is applied to a new study in eye-tracking and postural stability assessment, called the Utah State University (USU) Posture Study. The setup, data collection, and data preprocessing of the USU Posture Study are also provided. These new tests, called the modified Syrjala tests, are made available via the `distdiffR` package for the `R` software environment for statistical computing and graphics. A vignette and user manual for the package are also provided.

(41 pages)

PUBLIC ABSTRACT

Extensions to the Syrjala Test with Eye-Tracking Data Analysis Applications in R

Eric D. McKinney

Eye tracking is a process for measuring the movement of an individual's eye(s) when that individual is looking at something. Many eye-tracking technologies exist to aid in calculating and recording data associated with what a person focuses their visual attention on. For example, eye-tracking technology can record points on an image that a person is looking at. Often the question arises as to whether two people, or groups of people, are looking at the same thing(s). This dissertation presents a new way (or test) to quantify those differences while taking into consideration the randomness associated with such data. Hence, the test can help to determine if the differences between what the two people, or groups of people, are looking at are caused by chance or not. However, the test is also useful to many other kinds of data similar to but outside of eye-tracking research. While this test takes longer for standard household computers to run than other alternative tests currently available, it is shown to be better in many cases at correctly identifying differences when those differences were not caused by randomness. The test is also better at identifying when the differences are caused by chance, and not necessarily by the people. The test is applied to eye-tracking data from a study held at Utah State University (USU), called the USU Posture Study, where many differences are found. The test is available online, and comes with a user manual and some examples of how to use it.

This work is dedicated to my darling wife, Brianna.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Jürgen Symanzik. I am truly grateful for all of the support he has provided throughout this dissertation. I am also thankful for the countless hours of constructive feedback which has enabled me to reach higher levels of professionalism and quality in my research. I have deeply appreciated his expertise and insights.

I would also like to thank my committee members, supporting faculty, colleagues, and fellow students at USU, particularly, Dr. Breanna Studenka, Dr. Richard Cutler, Dr. Daniel Coster, Dr. John Stevens, Dr. Yan Sun, Dr. Brennan Bean, Dr. Todd Moon, Dr. Chunyang Li, Joanna Coltrin, Shannon Dixon, Chih-Ching Yeh, Kristina Casos, Melanie Athens, and Madison Hansen. Thank you for your insightful discussions and for your efforts in contributing to the USU Posture Study (Symanzik et al., 2017, 2018; Studenka et al., 2020; Coltrin et al., 2020; McKinney and Symanzik, 2019, 2021). I would like to especially thank Dr. Adele Cutler for her experienced and thoughtful contributions, particularly her guidance that led to exploring more points as origins for the bivariate empirical distribution functions.

Relatedly, I am deeply appreciative of the support provided by the Department of Mathematics and Statistics at Utah State University (`https://math.usu.edu/`). I am especially grateful to the department heads, secretaries, graduate program coordinators, and IT personnel, particularly Dr. Christopher Corcoran, Dr. James Powell, Dr. John Stevens, Nancy Smart, Gary Tanner, Kelly Seipert, Sara Poulsen, Gaby Hainsworth, Karl Dyches, and Isabel Jenson for their kindness and indispensable professional support.

This work would also not have been possible without the foundation of theoretical and applied mathematics built during my undergraduate education at Weber

State University. I am especially grateful to my undergraduate research advisor, Dr. Mahmud Akelbek, as well as the many other inspiring and supportive faculty members, including Dr. Sandra Fital-Akelbek, Dr. Julian Chan, Dr. Mihail Cocos, Dr. Matt Ondrus, Dr. Afshin Ghoreishi, Dr. George Kvernadze, Dr. Kent Kidman, Dr. Paul Talaga, and Dr. James Peters.

Additionally, the support and resources from the Center for High Performance Computing (`https://www.chpc.utah.edu/`) at the University of Utah are gratefully acknowledged. A special thank you belongs to Dr. Martin Cuma, who's high performance computing expertise made many of the simulation studies contained in this dissertation possible.

I would also like to thank my Heavenly Father, my darling wife, Brianna, and our precious children. This was indeed a family endeavor–and we did it!

Eric D. McKinney

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

This dissertation introduces a series of new two-sample tests of distributional equality. The new tests are a generalization of the Syrjala (1996) test and make use of both rotations and toroidal shifts of the data. Several of the tests are shown to be competitive or better than other modern techniques. The test which employs both rotations and toroidal shifts is applied to a new study in eye-tracking and postural stability assessment, called the Utah State University (USU) Posture Study (Symanzik et al., 2017, 2018; Studenka et al., 2020; Coltrin et al., 2020; McKinney and Symanzik, 2019, 2021). The setup, data collection, and data preprocessing of the USU Posture Study is also provided. The study results suggest that there are an abundance of detectable differences between and within the treatment and control groups captured in the subject's eye-tracking data. The new tests, called the modified Syrjala tests, are made available via the `distdiffR` package for the `R` software environment for statistical computing and graphics (R Core Team, 2019).

1.1 Motivation

While multiple hypotheses are posed within the USU Posture Study, much of the motivation is provided in the following question, "Does judging the action capabilities of another person depend on one's own experiences?" Specifically, is there a significant difference in what subjects look at when judging the stability of a posture between subjects with and without recent yoga experience? Consequently, a series of bivariate spatial eye-tracking data was collected to answer this question. Additional details are provided in Chapter 4.

Several statistical tests for comparing bivariate distributions between two samples exist in the literature (Chapter 2). Among which, it was discovered that data binning methods have been used (Chetverikov et al., 2018; McAdam et al., 2012) in order to apply the Syrjala (1996) test. This has been shown to give contradictory results depending on the granularity of the binning technique (McKinney and Symanzik, 2019). However, several proposed modifications to the Syrjala test (Chapter 5) not only make it more generally applicable, but are also shown to be more powerful and more conservative than alternative methods, including the original Syrjala test.

1.2 Outline

The structure of this dissertation proceeds as follows: In Chapter 2, an overview of two-sample tests of distributional equality is provided with an emphasis on the two-sample bivariate case. Sections 2.2.1–2.2.4 outline four two-sample tests of distributional equality which can be used in the bivariate case, namely, the Syrjala (1996), energy (Rizzo and Székely, 2016), maximum mean discrepancy (Gretton et al., 2012), and Friedman and Rafsky (1979) tests. Chapter 3 provides an overview of basic eye-tracking analyses along with definitions of commonly measured variables and statistics. The end of Chapter 3 also provides a separation between the literature review/previous research conducted and this dissertation's novel contributions, which are detailed in the remaining chapters. Chapter 4 details the setup, data collection, and data preprocessing of the USU Posture Study. The Syrjala test provides inspiration to a series of modified versions presented in Chapter 5. Multiple simulation comparisons between the modified Syrjala tests and the other tests (covered in Sections 2.2.1–2.2.4) are detailed in Chapter 6. A series of simulations which demonstrate the performance of the modified Syrjala tests on data generated to more closely resemble scenarios observed in eye-tracking studies is also provided in Chapter 6. An

application of one version of the modified Syrjala tests to the USU Posture Study is detailed in Chapter 7 followed by an introduction of the `distdiffR` `R` package for the `R` software environment for statistical computing and graphics (R Core Team, 2019) provided in Chapter 8. A vignette and user manual for the `distdiffR` package are also provided in Chapter 8. Chapter 9 provides some concluding remarks, and outlines additional areas for further study. Appendix A provides a series of mathematical proofs and additional details referred to throughout this dissertation. Appendix B gives many additional simulation results. Additionally, a collection of figures used within the USU Posture Study is provided in Appendix C.

CHAPTER 2

Two-sample Tests of Distributional Equality

This chapter provides an overview of two-sample tests of distributional equality. Section 2.1 begins with the univariate case while Section 2.2 covers the multivariate case. An emphasis is made on the bivariate case in Section 2.2, as the application to eye-tracking data (made in Chapter 7) leverages a bivariate two-sample test of distributional equality.

2.1 Tests for Univariate Data

When handling two samples of continuous bivariate data, the question arises as to whether or not the sampled data come from the same population. Well known tests for determining whether two samples come from the same population in the univariate case include the two-sample Kolmogorov-Smirnov test (Kolmogorov, 1933) and the Cramér-von Mises test (Cramér, 1928; von Mises, 1928) generalized to the two-sample case by Anderson (1962).

To define these more precisely, let $X_{1,1}, X_{1,2}, \ldots, X_{1,n_1}$ and $X_{2,1}, X_{2,2}, \ldots, X_{2,n_2}$ be two independent random samples with sample sizes n_1 and n_2, respectively. Also, let $F_1(x)$ and $F_2(x)$ be unknown cumulative distribution functions (CDFs), and $S_1(x)$ and $S_2(x)$ be empirical cumulative distribution functions (ECDFs), for each respective sample. The ECDF of a sample is defined as,

$$S_i(x) = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{1}_{X_{i,j} \leq x}, \quad -\infty < X_{i,j} < \infty$$

where $\mathbf{1}_{X_{i,j} \leq x}$ is one if $X_{i,j} \leq x$ and zero otherwise, $i = 1, 2$ is the sample index, j is

the observation index within sample i, and n_i is the sample size of the i^{th} sample. The plot of the ECDF, which is a step function, provides an exhaustive representation of the data (D'Agostino, 1986). Figure 1 shows an example of an ECDF plot.

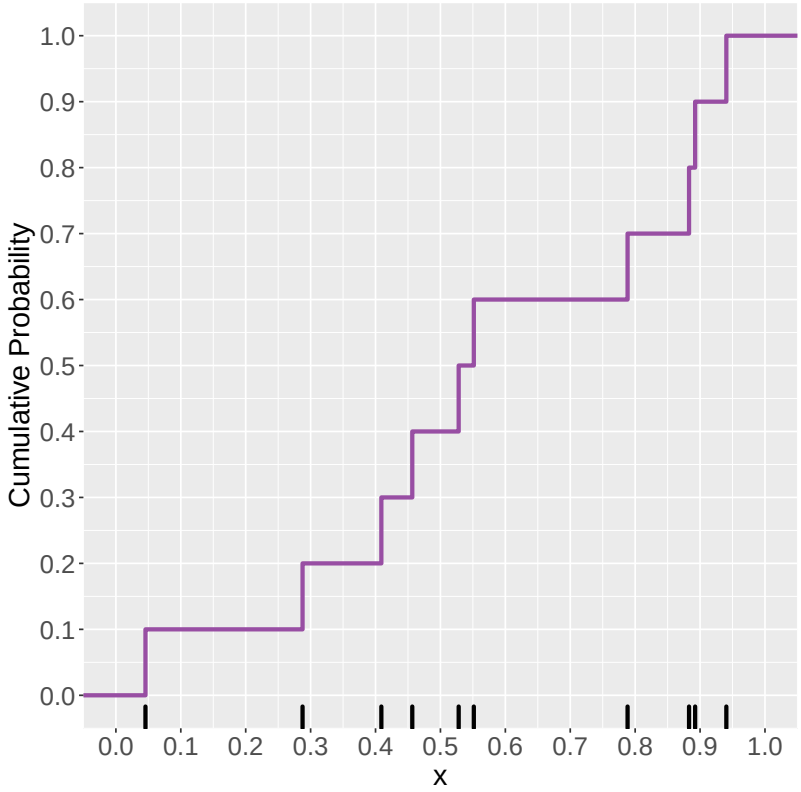


Fig. 1: A plot of a univariate ECDF for ten randomly generated uniform data values. A rug plot, displayed on the horizontal axis, shows the locations of the actual data values. Since there are ten data values, the ECDF jumps by a value of $1/10 = 0.1$ at each data value location.

Additionally, $S_i(x)$ is an unbiased and consistent estimator of $F_i(x)$. For mathematical proofs of both of these properties, see Appendix A.

Consequently, the ECDF is commonly used in test statistics which address the following hypotheses:

H_0: $F_1(x) = F_2(x) \quad \forall x$

H_a: $F_1(x) \neq F_2(x)$ for at least one value of x.

Among which, the well known two-sample Kolmogorov-Smirnov test statistic (Kolmogorov, 1933) is defined as

$$T_{KS} = \sup_x |S_1(x) - S_2(x)|,$$

where sup is the supremum function (also referred to as the least upper bound) (Rudin, 1964). Figure 2 shows an example of the two sample Kolmogorov-Smirnov test statistic calculated between two ECDFs.

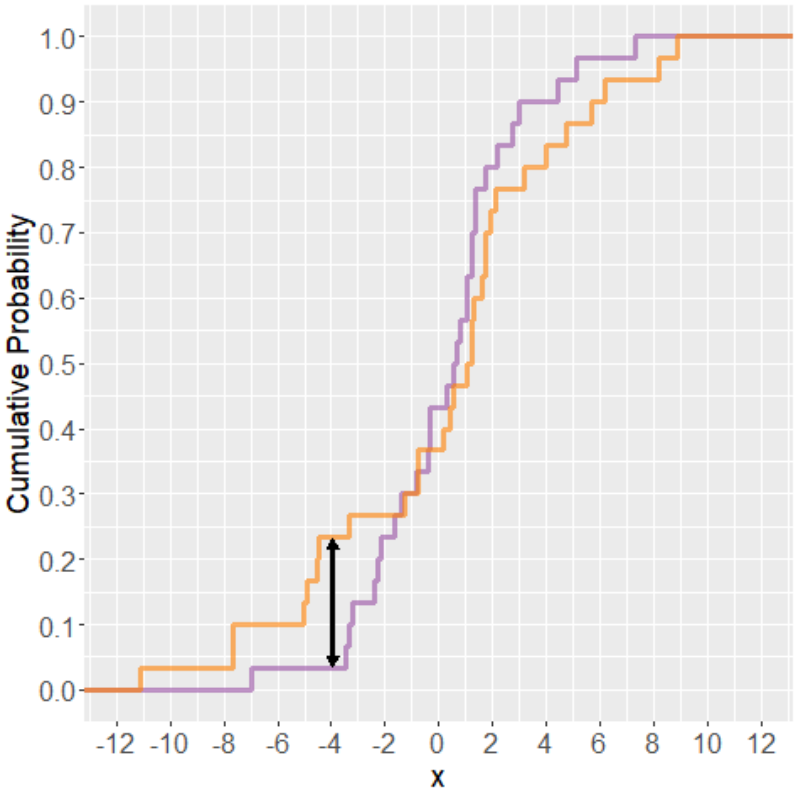


Fig. 2: An example of the two-sample Kolmogorov-Smirnov test statistic which is the maximum absolute vertical distance (represented by the black line segment) calculated between two ECDFs.

Furthermore, the two sample Cramér-von Mises test statistic (Anderson, 1962), which employs Riemann-Stieltjes integration, is

$$T_A = \frac{n_1 n_2}{n_1 + n_2} \int_{-\infty}^{\infty} [S_1(x) - S_2(x)]^2 dS^*(x),$$

which can be calculated using

$$(1) \qquad T_A = \frac{n_1 n_2}{(n_1 + n_2)^2} \left\{ \sum_{j=1}^{n_1} [S_1(x_{1,j}) - S_2(x_{1,j})]^2 + \sum_{j=1}^{n_2} [S_1(x_{2,j}) - S_2(x_{2,j})]^2 \right\}.$$

where $S^*(x)$ is the ECDF of the combined samples. Specifically,

$$S^*(x) = \frac{n_1 S_1(x) + n_2 S_2(x)}{n_1 + n_2}.$$

Additionally, if there are no duplicate observations, and both samples are sorted in increasing order such that $r_{1,1}, r_{1,2}, \ldots, r_{1,n_1}$ and $r_{2,1}, r_{2,2}, \ldots, r_{2,n_2}$ are the ranks of the observations in both samples, respectively, then Anderson (1962) showed that Equation 1 can be written as

$$T_A = \frac{U_A}{n_1 n_2 (n_1 + n_2)} - \frac{4 n_1 n_2 - 1}{6(n_1 + n_2)},$$

where

$$U_A = n_1 \sum_{j=1}^{n_1} (r_{1,j} - j)^2 + n_2 \sum_{j=1}^{n_2} (r_{2,j} - j)^2.$$

Similarly, Darling (1957) proposed a two-sample version to the Anderson-Darling statistic (Anderson and Darling, 1952, 1954), which was further detailed by Pettitt (1976). The statistic (also using Riemann-Stieltjes integration) is defined as

$$T_D = \frac{n_1 n_2}{n_1 + n_2} \int_{-\infty}^{\infty} \frac{[S_1(x) - S_2(x)]^2}{S^*(x)[1 - S^*(x)]} dS^*(x),$$

This has been further generalized to a k-sample Anderson-Darling statistic proposed by Scholz and Stephens (1987).

Since the two-sample Kolmogorov-Smirnov, Cramér-von Mises, and Anderson-Darling tests are permutation tests (also called randomization tests, re-randomization tests, or exact tests) (Kolmogorov, 1933; Berry et al., 2011), the respective test statistic T is recalculated $N = \frac{(n_1+n_2)!}{n_1!n_2!}$ times producing T_k data-permuted test statistics where $k = 1, \ldots, N$. Specifically, N is the total number of permutations of the sample labeling subscripts within $X_{1,1}, X_{1,2}, \ldots, X_{1,n_1}$ and $X_{2,1}, X_{2,2}, \ldots, X_{2,n_2}$, where each permutation will relabel n_1 points as sample one and n_2 points as sample two. The p-value is calculated as the total proportion of test statistics T_k which are greater than or equal to the statistic T computed on the non-permuted data (Davison and Hinkley, 1997), i.e.,

$$\text{p-value} = \frac{\sum_{k=1}^{N} (I_{T_k \geq T}) + 1}{N + 1}.$$

While the two-sample Kolmogorov-Smirnov test belongs to a more general class of supremum ECDF statistics, the two-sample Cramer-von Mises and Anderson-Darling tests belong to a class of quadratic ECDF statistics (D'Agostino, 1986). The new statistic proposed in Chapter 5 of this dissertation falls within this latter class, except it is designed for bivariate two-sample data. A survey of bivariate two-sample tests is covered in Section 2.2.

Outside of ECDF statistics, there exists other approaches to measuring distributional equality between two samples, such as non-parametric rank tests. A commonly used rank test is the Mann-Whitney U test (also called the Mann-Whitney-Wilcoxon,

Wilcoxon rank-sum test, or Wilcoxon-Mann-Whitney test)(Mann and Whitney, 1947). While originally designed to test for a difference in medians between the two population distributions, it can be extended as a test for any difference between two samples from an unspecified distribution (Pratt, 1964). The alternative hypothesis is the probability of an observation from the first population X_1 exceeding an observation from the second population X_2 is different than the probability of observation X_2 exceeding observation X_1, i.e., $P(X_1 > X_2) \neq P(X_2 > X_1)$. The statistic is calculated using the following steps:

1. While preserving sample labels, combine sample values into one set and sort the values in increasing order.

2. Rank all of the sample values. The smallest value will be assigned a 1 and the largest value will be assigned $n_1 + n_2$ (if there are no ties for the smallest and largest values).

3. If ties exist, replace each tied value's rank with the average value of all of the ranks for that tied value. For example, if the values are $\{3, 5, 5, 5, 5, 11\}$, then the unadjusted ranks would be $\{1, 2, 3, 4, 5, 6\}$. However, since there are four 5 values, the new rank for each 5 value would be $(2 + 3 + 4 + 5)/4 = 3.5$. Hence, the adjusted ranks would be $\{1, 3.5, 3.5, 3.5, 3.5, 6\}$.

4. Sum the ranks for each sample separately. Let each sum be R_1 and R_2, respectively. (The sum of all ranks will equal $[(n_1 + n_2)(n_1 + n_2 + 1)]/2$.)

5. Next calculate:

$$U_i = R_i - \frac{n_i(n_i + 1)}{2}, \text{ for } i = 1, 2.$$

While the same p-value can be computed from either U_i, it is common to use

$$U = \min_{i=1,2}\{U_i\}.$$

U_i will range from 0 to $n_1 \times n_2$, and it represents the number of times a value from the i^{th} sample precedes a value from the other sample in a sequence of combined sample values. If we assume that the null hypothesis is true, then each of the $\binom{n_1+n_2}{n_i}$ reassignments of the sample labels to the i^{th} sample values is equally likely. Hence, the null distribution of U can be calculated explicitly. Mann and Whitney (1947) showed that $E(U) = (n_1 n_2)/2$, where E is the expected value. Hence,

$$\text{p-value} = P\left(|U - E(U)| \geq |U^* - E(U)|\right),$$

where U^* is the computed U value using the original sample labels. Additionally, Mann and Whitney (1947) showed that the distribution of U approaches a normal distribution as n_1 and n_2 jointly approach infinity.

The Mann-Whitney U test was also extended to k-samples by Kruskal and Wallis (1952) for use in comparing the medians in the k-samples. Steel's test (Steel, 1960, 1961) conducts pair-wise Mann-Whitney U tests for k-samples. Additionally, the Jonckheere-Terpstra test (Jonckheere, 1954; Terpstra, 1952) has been shown to be more powerful than the Kruskall-Wallis test when there exists an apriori ordering in the populations.

Wald and Wolfowitz (1944) also developed a test for measuring whether two binary values which occur in a sequence are drawn independently from the same distribution. The test measures the number of "runs" in the sequence, where "runs" are defined as non-empty segments of the sequence which consist of subsequent elements of the same binary class.

For example, a coin that is flipped 12 times produces the sequence

$$\{H,\ H,\ H,\ H,\ H,\ H,\ T,\ T,\ T,\ H,\ T,\ T\},$$

where H represents a heads, and T represents a tails. This sequence consists of four runs, namely $\{H,\ H,\ H,\ H,\ H,\ H\}, \{T,\ T,\ T\}, \{H\}$, and $\{T,\ T\}$.

Assuming that the heads and tails are being drawn from the same distribution, let the number of runs in a sequence of length $N_{HT} = N_H + N_T$, where N_H is the number of heads flipped, and N_T is the number of tails flipped. Wald and Wolfowitz (1944) showed that the number of runs, N_R, follows a normal distribution such that

$$N_R \sim Normal\left(\frac{2N_H N_T}{N_{HT}} + 1, \frac{2N_H N_T(2N_H N_T - N_{HT})}{N_{HT}^2(N_{HT} - 1)}\right).$$

However, the Wald-Wolfowitz runs test is well known to be generally one of the less powerful non-parametric tests (Friedman and Rafsky, 1979).

2.2 Tests for Bivariate Data

This section provides a review of the literature on two-sample tests of distributional equality for bivariate data, and builds upon the univariate two-sample tests overviewed in Section 2.1. Some attempt has been made to develop methods for applying univariate tests (see Section 2.1) to higher dimensional data, such as using statistically equivalent blocks (Wilks, 1941; Fraser, 1951). However, many of the univariate two-sample tests have also been generalized to bivariate settings.

For example, Friedman and Rafsky (1979) generalized the two-sample Kolmogorov-Smirnov test to the bivariate case, and Bickel (1969) developed a p-variate version. In the same paper, Friedman and Rafsky (1979) also proposed a bivariate version of the Wald-Wolfowitz runs test (Wald and Wolfowitz, 1944) which is based on a minimal

spanning tree on the pooled sample. Both the Wald-Wolfowitz and Kolmogorov-Smirnov generalizations of the Friedman and Rafsky (1979) test are discussed in greater detail in Section 2.2.4. The Friedman and Rafsky's Kolmogorov-Smirnov generalization is included in the simulation comparisons detailed in Chapter 6. The Friedman and Rafsky's Wald-Wolfowitz generalization is not included due to its low exhibited power as noted by Friedman and Rafsky (1979), which was also exhibited in preliminary results of the simulations included in Chapter 6.

Furthermore, Choi and Marden (1997) proposed multivariate tests which mimic the two-sample Mann-Whitney U test (Mann and Whitney, 1947), the Jonckheere-Terpstra test for trend (Jonckheere, 1954), and the Kruskal-Wallis one-way analysis of variance test (Kruskal and Wallis, 1952). Anderson et al. (1994) extended a goodness-of-fit test by Hall (1984) to a two-sample test for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. Additionally, tests involving the nearest neighbors algorithm have been developed for the multivariate two-sample setting (Schilling, 1986; Henze, 1988; Mondal et al., 2015).

Similar to nearest neighbor techniques, there exists a variety of published methods which employ pairwise distances between observations (Hall and Tajvidi, 2002; Baringhaus and Franz, 2004). Within this group, several publications have centered on a concept of "energy" which can be measured on a data set (Zech and Aslan, 2003; Székely and Rizzo, 2004). In addition, the Székely and Rizzo (2004) energy test, discussed in more detail in Chapter 2.2.2, is rotational-invariant and can be applied in the univariate or multivariate setting. It is also included in the simulation comparisons detailed in Chapter 6.

Additionally, Gretton et al. (2012) proposed a statistic called the maximum mean discrepancy, which measures the largest difference in expectations over functions in

the unit ball of a reproducing kernel Hilbert space. This test is presented in greater detail in Chapter 2.2.3, and is also included in the simulation comparisons detailed in Chapter 6.

Syrjala (1996) proposed a generalization of the Cramér-von Mises test to the bivariate case. The Syrjala test, discussed in more detail in Chapter 2.2.1, has been applied in many cases in the literature including tests of differences in the spatial distributions of adult vs. juvenile Pacific cod off the coast of Alaska (Syrjala, 1996), tests of differences in the distribution of the same bird species over three consecutive years in Central Spain (Benayas et al., 2010), and tests of differences in the distribution of two different respiratory infections affecting turtles in the Mojave Desert (Berry et al., 2015). In some cases, due to restrictions on the sampling locations being identical within the Syrjala test, preliminary data binning has been carried out (Chetverikov et al., 2018; McAdam et al., 2012). Unfortunately, depending on the granularity in the binning of the data, the simulations detailed in Chapter 6 show that the results can be contradictory. Chiu and Liu (2009) also summarized several other proposed generalizations of the Cramér-von Mises test.

The Syrjala test, while limited to two-sample tests with identical sampling locations, provides some of the inspiration to multiple proposed generalizations and modifications, called the modified Syrjala tests, which are discussed in Chapter 5. These proposed modifications to the Syrjala test not only make the original test more generally applicable, but are also shown to be more powerful and more conservative than alternative methods, including the original Syrjala test (see Chapter 6).

2.2.1 The Syrjala Test

Stephen E. Syrjala proposed a test for differences in the spatial distribution of two groups as a generalization of the univariate Cramér-von Mises test to the bi-

variate setting (Syrjala, 1996). The test is designed for determining whether there is a difference between the locations of two populations at a single point in time, or whether there is a difference between the locations of the same population at two points in time. However, the test requires the two samples both occur at an identical set of predefined locations. Furthermore, "The random variable in this case is the observed density at the sampling location, not the location itself." (Syrjala, 1996). Hence, the hypotheses under consideration for the Syrjala test can be stated as follows:

H_0: The normalized distributions of the populations are equal across the study area.

H_a: There is some unspecified difference in the normalized population distributions.

Adjusting the notation from Syrjala (1996), let $d_i(x_j, y_j)$ denote the density of observations for sample i; $i = 1, 2$, at sample locations (x_j, y_j); $j = 1, \ldots, n$, relative to their position on a bounding rectangle $\mathcal{A}$, where n is the total number of sampling locations. Then, $D_i = \sum_{j=1}^{n} d_i(x_j, y_j)$ is the sum of all densities across $\mathcal{A}$, and $\gamma_i(x_j, y_j) = \frac{d_i(x_j, y_j)}{D_i}$ are the normalized densities. With these, we can construct $\Gamma_i(x, y) = \sum_{x_j \leq x, y_j \leq y} \gamma_i(x_j, y_j)$ which is analogous to the bivariate empirical distribution function. Figure 3 shows a visualization of densities of observations at each of the identical sampling locations for two separate samples.

Similar to the Cramér-von Mises permutation test, the Syrjala first calculates a statistic (Ψ) on the non-permuted samples.

$$\Psi = \sum_{j=1}^{n} \left[\Gamma_1(x_j, y_j) - \Gamma_2(x_j, y_j)\right]^2 .$$

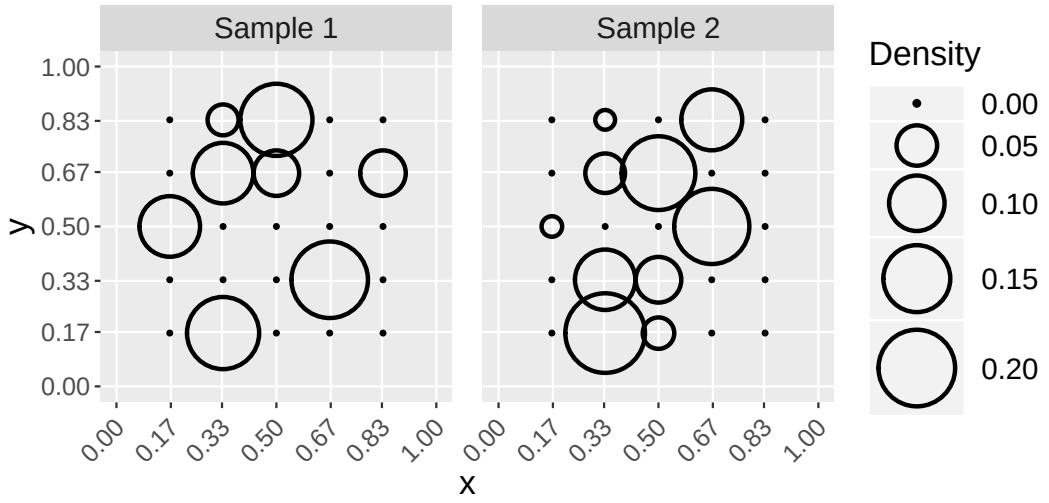


Fig. 3: An example of identical sampling locations with differing location densities from two spatial distributions inspired by Benayas et al. (2010), page 312. Larger radii of the circles represent greater densities at each of the sampling locations.

Because this construction uses points near the origin more often than points in the center, a common adjustment to Syrjala's test is

$$\widetilde{\Psi} = \frac{1}{4}\sum_{c=1}^{4}\Psi_c, \text{where } \Psi_c = \sum_{j=1}^{n}\left[\Gamma_1(x_{c,j}, y_{c,j}) - \Gamma_2(x_{c,j}, y_{c,j})\right]^2,$$

and $(x_{c,j}, y_{c,j})$ are positive coordinates defined relative to each of the four corners of $\mathcal{A}$ (Syrjala, 1996).

Next, permutations of the data are made by choosing to swap or leave $\gamma_1(x_j, y_j)$ and $\gamma_2(x_j, y_j)$ at each (x_j, y_j) locations. This results in $M = 2^n$ possible permutations of the data. Test statistics $\widetilde{\Psi}_k; k = 1, \ldots, M$, are calculated for each of the permutations of the data. The $\widetilde{\Psi}_k$ calculations are identical to that of $\widetilde{\Psi}$ (including the four rotations) except that they are computed from the permuted data. The p-value is calculated as the proportion of permuted statistics $\widetilde{\Psi}_k$ which are greater than or equal to the original statistic $\widetilde{\Psi}$ that is computed from the non-permuted data, i.e.,

$$\text{p-value} = \frac{\sum_{k=1}^{M} \left(I_{\widetilde{\Psi}_k \geq \widetilde{\Psi}} \right) + 1}{M + 1}.$$

It should be noted that in practice only a subset of $M' \ll M$ (typically $M' \approx 999$) possible permutations are used to calculate the level of significance due to computational limitations.

The Syrjala test has been implemented in the `ecespa` `R` package (de la Cruz Rot et al., 2008) for the `R` software environment for statistical computing and graphics (R Core Team, 2019).

However, depending on the data aggregation steps (such as binning) the results of the Syrjala test can be contradictory (see Chapter 6). Proposed modifications to the Syrjala test in Chapter 5 eliminate the restriction for the sampling locations to be identical and the need to bin the data. A simulation study in Chapter 6 suggests that our modified versions of the Syrjala test are in general more powerful and more conservative as compared to the original Syrjala test.

Furthermore, McKinney and Symanzik (2019) applied both the Syrjala test and one of the modified Syrjala tests to a study involving eye tracking and posture perception of individuals at Utah State University (USU) (Symanzik et al., 2017, 2018; Studenka et al., 2020; Coltrin et al., 2020; McKinney and Symanzik, 2019, 2021). The modified Syrjala test showed stable results as compared to differing results from the original Syrjala test depending on the data binning technique employed.

2.2.2 The E-Statistics (Energy) Test

There exists a variety of published methods which employ pairwise distances between observations (Hall and Tajvidi, 2002; Baringhaus and Franz, 2004). Within this group, several publications have centered on a concept of "energy" which can be

measured on a data set (Zech and Aslan, 2003; Székely and Rizzo, 2004; Rizzo and Székely, 2016). In addition, the Székely and Rizzo (2004) energy test is rotational-invariant and can be applied in the univariate or multivariate setting.

Inspired by Newton's gravitational potential energy, Rizzo and Székely (2016) proposed a class of "energy" statistics which measure the distance between distributions of random vectors. Consequently, not only can the statistic be applied in the two-sample case, but it can also be used to measure the difference between sampled and hypothesized distributions. Furthermore, applications of the statistic have been found within testing for independence by distance covariance, extensions to analysis of variance, generalizations of clustering algorithms, change point analysis, and feature selection (Rizzo and Székely, 2016) to name a few.

If we let $\vec{X} = \{\widetilde{X}_1, \widetilde{X}_2, \ldots, \widetilde{X}_d\}$ and $\vec{Y} = \{\widetilde{Y}_1, \widetilde{Y}_2, \ldots, \widetilde{Y}_d\}$ be independent random vectors which belong to $\mathbb{R}^d$ with CDFs F_X and F_Y, respectively, then the energy distance between F_X and F_Y is defined as

$$D(F_X, F_Y) = \sqrt{2E||\vec{X} - \vec{Y}|| - E||\vec{X} - \vec{X}'|| - E||\vec{Y} - \vec{Y}'||} \geq 0,$$

where E is the expected value, $||\cdot||$ is the Euclidean norm, and $\vec{X}'$ and $\vec{Y}'$ are independent and identically distributed copies of $\vec{X}$ and $\vec{Y}$, respectively. In practice, $\vec{X}'$ and $\vec{Y}'$ are obtained by the bootstrap (Davison and Hinkley, 1997) re-sampling method.

Similar to potential energy, which is zero if and only if the gravitational center of two objects coincide, $D(F_1, F_2) = 0$ if and only if $F_1 = F_2$. Additionally, just as potential energy increases between two objects as the distance between their gravitational centers increase, $D(F_1, F_2)$ will increase as F_1 and F_2 move apart from each other. Hence, as a well defined metric the energy distance can characterize the

equality of distributions.

Therefore, for the multivariate case, let $\vec{X}_1, \vec{X}_2, \ldots, \vec{X}_{n_1}$ and $\vec{Y}_1, \vec{Y}_2, \ldots, \vec{Y}_{n_2}$ be random samples of random vectors in $\mathbb{R}^d$, $(d > 1)$. Rizzo and Székely define the energy statistic or "E-statistic" $(\mathcal{E}_{n_1,n_2})$ as

$$\begin{aligned}\mathcal{E}_{n_1,n_2} =& \frac{2}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} ||\vec{X}_i - \vec{Y}_j|| \\ &- \frac{1}{n_1^2} \sum_{i=1}^{n_1} \sum_{k=1}^{n_1} ||\vec{X}_i - \vec{X}_k|| \\ &- \frac{1}{n_2^2} \sum_{j=1}^{n_2} \sum_{l=1}^{n_2} ||\vec{Y}_j - \vec{Y}_l||,\end{aligned}$$

and show that $\mathcal{E}_{n_1,n_2} = D^2(F_1, F_2) = 0$ if and only if F_1 and F_2 have the same distribution (i.e., $F_1 = F_2$).

Furthermore, the test statistic

$$T_{\mathcal{E}} = \frac{n_1 n_2}{n_1 + n_2} \mathcal{E}_{n_1,n_2}$$

converges in distribution to a quadratic form of independent standard normal random variables (Rizzo and Székely, 2016). Since the null distribution of $T_{\mathcal{E}}$ is dependent on the distributions of X_1 and X_2, the test (similar to those described in Chapter 2) is implemented as a nonparametric permutation test in the `energy` `R` package (Rizzo and Szekely, 2019). Rizzo and Szèkely's energy test is included in the simulation comparisons detailed in Chapter 6.

2.2.3 The Kernel Maximum Mean Discrepancy Test

Gretton et al. (2012) proposed a statistic called the maximum mean discrepancy (MMD), which measures the largest difference in expectations over functions in the

unit ball of a reproducing kernel Hilbert space (RKHS). Their two-sample test leverages a technique where the distributions of the two samples are embedded in a RKHS (Song, 2008) by use of a characteristic kernel function.

Kernel Functions

Given $n \in \mathbb{N}$ and $c_1, \ldots, c_n \in \mathbb{R}$, any symmetric function $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel provided that for all $x_1, \ldots, x_n \in \mathcal{X}$ the following holds:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j f(x_i, x_j) \geq 0.$$

A kernel function (or simply kernel) ($\mathcal{K}(\cdot, \cdot)$) can be viewed as a special kind of similarity measure that takes as input two elements from the same space and outputs a real number:

$$\mathcal{K}(x, x') : \mathcal{X} \times \mathcal{X} \to \mathbb{R},$$

where x and x' can be considered here as two realizations of a random variable X. Kernels possess desirable mathematical properties (Song, 2008), including the ability to be decomposed into the inner product of a feature space ($\mathcal{H}$) mapping (ϕ) between two elements in $\mathcal{X}$:

$$\mathcal{K}(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}.$$

Kernels can be especially useful (e.g., their use in support vector machines) since they allow linear classification models to be fit to the feature space representation of the data without the necessity of actually computing $\phi(x)$, which can be computationally expensive. Additionally, the Moore-Aronszajn theorem (Aronszajn, 1950) guarantees the existence of a unique RKHS provided a given kernel. If the kernel used is "characteristic", meaning if the mapping of the family of distributions over

the domain of the random variable onto the feature space is injective, then each distribution can be uniquely represented in the RKHS. This will preserve all statistical features of the distributions within $\mathcal{H}$.

Computation of the MMD

Provided two samples of size n_1 and n_2, and a given kernel ($\mathcal{K}$) the MMD is estimated by

$$\begin{aligned}\widehat{MMD}(F_1, F_2) =& \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{k=1}^{n_1}\mathcal{K}(x_{1,i}, x_{1,k}) \\ &+ \frac{1}{n_2^2}\sum_{j=1}^{n_2}\sum_{l=1}^{n_2}\mathcal{K}(x_{2,j}, x_{2,l}) \\ &- \frac{2}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\mathcal{K}(x_{1,i}, x_{2,j}).\end{aligned}$$

The test is conducted as a nonparametric permutation test using the `kmmd` function within the `kernlab` `R` package (Karatzoglou et al., 2019). The `kmmd` function uses the radial basis function or Gaussian kernel, $\mathcal{K}(x, x') = \exp\left(-\frac{||x-x'||^2}{2\sigma^2}\right)$, as the default kernel (where σ is a free parameter). Gretton's kernel maximum mean discrepancy test is also included in the simulation comparisons detailed in Chapter 6.

2.2.4 The Friedman-Rafsky Test

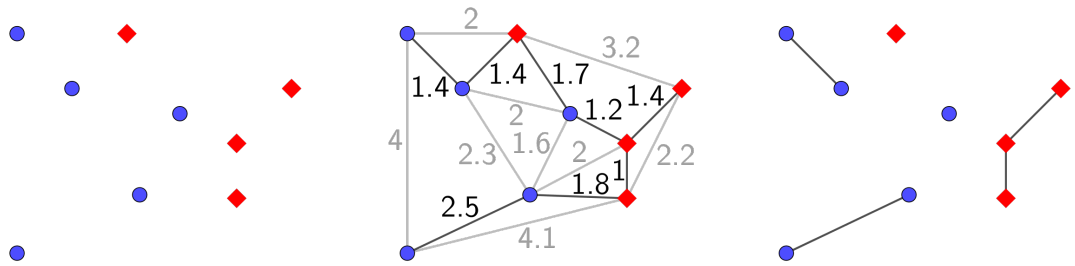
In an effort to generalize the idea of ordering to mulitvariate data, Friedman and Rafsky (1979) proposed a graph based generalization to the Wald and Wolfowitz (1944) test. While the Wald and Wolfowitz test (see Chapter 2 for more detail) considers adjacent observations of the same class as a "run" in its test statistic, the Friedman-Rafsky test generalizes a "run" to the multivariate setting by use of a minimum spanning tree (MST). Friedman and Rafsky also use MSTs to provide a way to apply the Kolmogorov-Smirnov test (Kolmogorov, 1933) to the multivariate

setting.

In graph theory, an MST (also called a minimum weight spanning tree) is defined as subset of the edges of a connected, edge-weighted undirected graph that connects all of the vertices together, with the minimum possible sum of all edge weights, and without any cycles (Wilson, 1996). An MST is obtained by use of the greedy algorithm, and is unique if there are no ties among all of the edge weights (Wilson, 1996). Figure 4(a) shows how the bivariate data values from two samples (depicted with blue circles and red squares) become a connected graph (in Figure 4(b)) by use of some difference measure (e.g., Euclidean distance). The MST highlighted within Figure 4(b) is obtained by use of the greedy algorithm. The Wald-Wolfowitz version of the Friedman-Rafsky statistic is then computed by removing all edges which connect points from separate samples and summing the total number of remaining trees (including individual data values with no edges). This is always one more than the number of edges removed (Friedman and Rafsky, 1979). Small test statistics provide evidence against the null-hypothesis that the two samples share the same distribution. In Figure 4(c), the computed Wald-Wolfowitz version of the Friedman-Rafsky statistic is five. Additionally, the statistic is usually standardized before being used within a nonparametric permutation test such as those described in Sections 2.1 and 2.2.

Furthermore, with the construction of an MST on the pooled samples, Friedman and Rafsky extend the Kolmogorov-Smirnov test (Kolmogorov, 1933) by providing a way to assign univariate ranks to each sample point by use of "rooted" trees and computing the "depth" of each node within the rooted tree (Friedman and Rafsky, 1979). Once the ranks are obtained the maximum difference in the univariate ECDFs of the ranks between the samples are computed (see Section 2.1) for more details).

Both the Wald-Wolfowitz and Kolmogorov-Smirnov versions of the Friedman-Rafsky test can be used within the `GSAR R` package (Rahmatallah et al., 2017)



(a) (b) (c)

Fig. 4: A visualization of the Wald-Wolfowitz version of the Friedman-Rafsky statistic. Subplot (a) shows bivariate data values from two samples (depicted with blue circles and red diamonds), which become a connected graph in subplot (b) by use of some difference measure (e.g., Euclidean distance). Subplot (b) also highlights the MST obtained by use of the greedy algorithm. Subplot (c) shows all of the remaining sub-graphs (also trees) obtained by removing all edges which connect points from separate samples. This provides us with a means to compute the Friedman-Rafsky statistic by summing the total number of remaining trees (including individual data values with no edges). The Wald-Wolfowitz version of the Friedman-Rafsky statistic computed in (c) is five.

available through the Bioconductor distribution (`https://git.bioconductor.org/packages/GSAR`). The Friedman and Rafsky's Kolmogorov-Smirnov generalization is included in the simulation comparisons detailed in Chapter 6.

CHAPTER 3

Overview of Eye Tracking Analyses

An overview of eye tracking analyses is provided in this chapter, including a treatment of fixation points (Section 3.2), areas of interest (Section 3.3), gaze transitions (Section 3.4), as well as a general analysis and overview of commonly used visualizations (Section 3.5).

3.1 Overview

The overall goal of eye-tracking analyses is to gain insight into what subjects give their attention to. Due to the rich spatial and temporal information that can be collected through eye-tracking analyses, many measures have been created to address a wide variety of hypotheses. When a subject focuses their visual attention on a single point, that point is called a "true" fixation point. To estimate these "true" fixation points, eye-tracking devices record points called gaze points (among other recorded measures). If an eye-tracking device records gaze points at a frequency of 30 Hz, the output will produce approximately (depending on the device) 30 gaze points per second in the data. Aggregations of gaze points produce fixation point estimates. Another commonly recorded measure is the saccade. A saccade occurs when a subject transitions from one fixation point to another (Holmqvist et al., 2011).

While not comprehensive, Table 1 shows a list of other commonly measured variables and statistics in eye-tracking analyses along with brief definitions. However, keep in mind that these definitions are not consistent across the literature. For example, fixation point estimates are referred to by a variety of names including fixation points (Duchowski, 2007; Kumar et al., 2018; Matsuda and Takeuchi,

2012), fixation locations (Deniz, 2016; Duchowski, 2007; Hessels et al., 2016), fixation positions (Goldberg and Helfman, 2010; Hessels et al., 2016), or simply fixations (Blascheck et al., 2017; Deniz, 2016; Duchowski, 2007; Hessels et al., 2016; Holmqvist et al., 2011; Matsuda and Takeuchi, 2012).

Table 1: Brief definitions of commonly measured variables/statistics in eye-tracking analyses taken from Holmqvist et al. (2011), Duchowski (2007), and Blascheck et al. (2017).

Common Variables/Statistics	Definition
Gaze point	Coordinates measured by the eye-tracking device indicating what the subject is looking at in the viewing region.
Fixation point estimate (or fixation point)	An aggregation of gaze points estimating where the subject is staring at (or fixating on) in the viewing region.
Area of interest (or AOI)	Defined region (or object) in the observation area of a subject that the researcher is interested in.
Number of fixation points (per subject or across subjects)	The number of fixation points on an object.
Fixation rate (per subject or across subjects)	The proportion of fixation points on an object.
Fixation duration	The length of time a subject fixated or stared at an object.
Fixation duration mean (per subject or across subjects)	The average fixation duration for an object.
Time to first fixation point	The length of time before a subject first fixates on a particular object.
Saccade	A transition from one fixation point to another.
Saccade velocity	The speed with which the eyes transition from one fixation point to another.
Number of revisits	The number of fixation transitions (saccades) to a previously fixated object.
Scanpath	A sequence of fixation points.

3.2 Fixation Points

However, raw data (e.g., the gaze points) collected with eye tracking devices contains noise caused by two sources: the eye tracking device and/or the subject. For example, variation in eye tracking algorithm calculations cause a random scattering of gaze points around the true fixation point of the eye. Additionally, major perturbations within the data can arise in the event that the subject blinks. Consequently, aggregating gaze points into fixation point estimates is a common technique for eliminating noise in the data. Two main classes of techniques used in accomplishing this involve either (1) averaging gaze points into a fixation point estimate (also called position-variance techniques by Duchowski (2007) and dispersion-based fixations by Holmqvist et al. (2011)), or (2) separating gaze points using thresholds on saccade velocities (also called saccade velocity techniques) (Duchowski, 2007; Holmqvist et al., 2011).

Yarbus (2013) modeled a formula for the angular velocity of the eye movement during a saccade. An angular velocity threshold given as an example by Duchowski (2007) is 30 degrees per second. However, Holmqvist et al. (2011) mentioned that typical saccade velocity thresholds range from 20 to 130 degrees per second in the literature. If time thresholds are specified in position-variance fixation points, then researchers should take into consideration that Irwin (1992) established the minimum theoretical fixation duration for the human eye as 150 milliseconds. However, according to Duchowski (2007), "The position-variance and velocity-based algorithms give similar results, and both methods can be combined to bolster the analysis by checking for agreement."

3.3 Areas of Interest

Next, Areas Of Interest (AOIs), or subregions of the observation area, are often defined to address more detailed hypotheses. A variety of techniques are being used to create AOIs in the current literature. Some researchers define AOIs using regular grids across the observation area, e.g., see Figure 5 inspired by Matsuda and Takeuchi (2012). This is useful since these AOI are content independent, and consequently, easily generated. However, inferential statistics have shown to be dependent on the granularity of the grid (Duchowski, 2007; McKinney and Symanzik, 2019). In contrast, other researchers have elected to define AOIs more subjectively, such as hand drawn areas over facial features found in Cantoni et al. (2012).

Hessels et al. (2016) compared these and several other different AOI constructions and noted their respective pros and cons. They concluded that the most objectively defined AOIs are constructed using the Voronoi (1908) method. However, Goldberg and Helfman (2010) suggested that the size of the AOI of an image object should depend on three factors: "(1) the importance of capturing every fixation [point] on that object, (2) the amount of white space surrounding the object, and (3) expected variance in fixation positions [(points)] across participants." Holmqvist et al. (2011) also pointed out that the minimal AOI size is limited by the precision and accuracy of the recorded data, which the Voronoi method does not take into consideration if tessellation centroids are defined close to each other.

3.4 Gaze Transitions

If researchers consider the order in which subjects viewed the AOIs as important to their analysis, then transition matrices can be calculated to depict the observed probabilites of transitions from one AOI to another (Holmqvist et al., 2011). Krejtz et al. (2015) detailed the computational steps for constructing transition matrices.

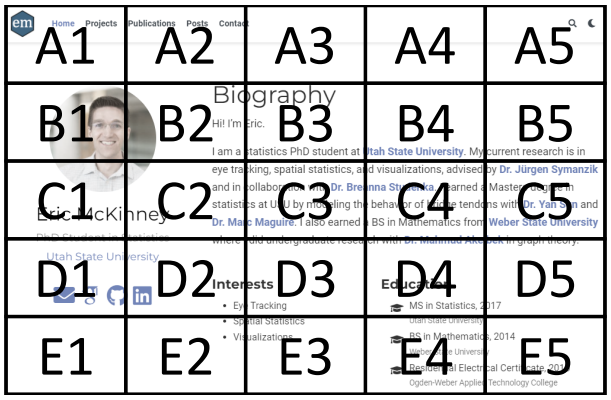


Fig. 5: An overlaid 5×5 grid on a viewing region (a website `https://ericmckinney.net/`) inspired by Matsuda and Takeuchi (2012), page 111.

Alternatively, Markov models (Rabiner and Juang, 1986) can be constructed to estimate the probabilities of transitioning from one AOI to another. Harris (1993) showed that AOI transition data is readily modeled by a so-called stationary, reversible first-order Markov model. This result, replicated by Gordon and Moser (2007), Epelboim and Suppes (2001), and Pieters et al. (1999), can be interpreted as showing that the probability of fixating on an object depends considerably on the object of the immediately preceding fixation point, but not on the objects fixated further back in the fixation sequence (Holmqvist et al., 2011).

Similarly, transition entropy can be analyzed to statistically compare fixation point transitions (Krejtz et al., 2015; Holmqvist et al., 2011). Transition entropy requires construction of the transition matrices in addition to their transformation into conditional probability matrices for which conditional transition entropy H_t is calculated,

$$H_t = -\sum_{i=1}^{n_{AOI}} \tilde{p}_i \sum_{j=1}^{n_{AOI}} \tilde{p}_{ij} \log_2 \tilde{p}_{ij},$$

where $\tilde{p}_i$ is the observed probability of viewing the ith AOI, $\tilde{p}_{ij}$ is the conditional probability of viewing the jth AOI given the previous viewing of the ith AOI, and n_{AOI} is the number of AOIs (Krejtz et al., 2015).

Hence, entropy H_t provides a measure of statistical dependency in the spatial pattern of fixation points represented by the transition matrix, and may be used to compare one matrix to another. (Note that a uniform grid for defining AOI is not necessary in order to compute the transition entropy.) Weiss et al. (1989) noted that a small H_t suggests dependencies between the fixation points, whereas a large H_t suggests a random scanning pattern.

3.5 General Analysis and Visualizations

Another employed measure related to spatial distribution of fixation points is the Nearest Neighbor Index (NNI), as described by Clark and Evans (1954). The NNI is based on the "distance from an individual to its nearest neighbor, irrespective of direction." The NNI ($\mathcal{N}$) describes the spatial distribution of points, e.g., fixation points, as either ordered ($\mathcal{N} > 1$), random ($\mathcal{N} = 1$), clustered ($\mathcal{N} < 1$), or maximally aggregated, i.e., singular ($\mathcal{N} = 0$). For n points, the NNI is calculated as

$$\mathcal{N} = \frac{2\sqrt{\rho}}{n} \sum_{i=1}^{n} r_i$$

where r_i is the distance from the ith (fixation) point to its nearest neighbor, and ρ is the density of the observed distribution, i.e., $\rho = n/A$ where A is the observation area. Holmqvist et al. (2011) discussed a variety of other measures including the convex hull area, the Mannan similarity index (Ruddock et al., 1995), the attention map difference, and the Kullback-Leibler distance (Kullback and Leibler, 1951).

Duchowski (2007) pointed out that statistical analyses generally consist of an analysis of variance (ANOVA) on the dependent variables collected during the eye tracking experiment, e.g., fixation durations, number of fixation points, etc., depending of course on the experimental design and its hypotheses.

In a comparison of a variety of eye-tracking analysis visualizations (categorized

as AOI based, point based, or both), Blascheck et al. (2017) indicated that variations of heatmaps and scanpath graphs are among the most commonly used.

CHAPTER 4

The Utah State University Posture Study

The Utah State University (USU) Posture Study aims at answering, among other hypotheses, the following question, "Does judging the action capabilities of another person depend on one's own experiences?" Specifically, is there a significant difference between subjects with and without recent yoga experience expressed in what those subjects look at when judging the stability of a posture?

While this chapter outlines the setup, data collection process, and data preprocessing conducted during the study (see Section 4.1), the USU Posture Study has already been the subject of several publications (Symanzik et al., 2017, 2018; Studenka et al., 2020; Coltrin et al., 2020; McKinney and Symanzik, 2019, 2021). These other publications provide preliminary results of the study along with additional details. Further analyses based on the new Syrjala tests conducted on the collected eye-tracking data are provided in Chapter 7.

4.1 USU Posture Study Details

A participant in the study was considered a "treatment" subject if they indicated that they have been practicing yoga for at least two hours a week on average for the past three months. A participant was a "control" subject if they indicated otherwise. In total, the data from 20 treatment and 20 control subjects was successfully collected in the study.

After being fitted with an eye-tracking device (as seen in Figure 6) from ETMOBILE (`http://www.argusscience.com/ETMobile.html`) and successful calibrated, each subject was randomly shown a series of 22 postures being held by an actor and

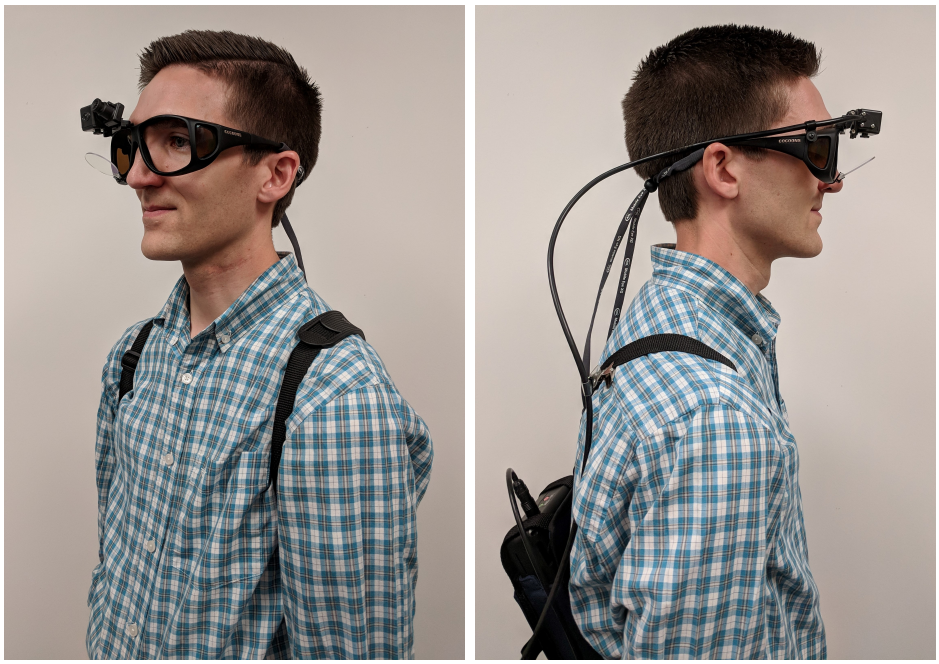


Fig. 6: An example of a subject wearing the ETMOBILE eye-tracking device. The eye tracker has one forward facing camera and one infrared camera that tracks the eye's movement by use of a transparent mirror.

asked, "How long do you think this person could hold this posture?" Appendix C includes figures of each of the postures shown to subjects (including the calibration images). Figure 7 demonstrates how the data was recorded. After the 22nd posture, the subject's eye-tracking calibration was reassessed to ensure valid data was collected during the entire trial. Figure 8 shows the lineups of the 22 postures shown to the 40 subjects. Within the figure, each row represents a subject, indicated as either a treatment (T) or a control (C) followed by an index number. The columns assign a "View Number" (V followed by an index) representing the order in which the subjects viewed the postures. Views one and 24 (V1 and V24) were identical calibration images (Figures 90 and 113 in Appendix C) for all of the subjects. Notice that each treatment subject had a corresponding control subject who was shown the 22 postures in the same order. However, the orders were randomized across subject-pairs within each group.

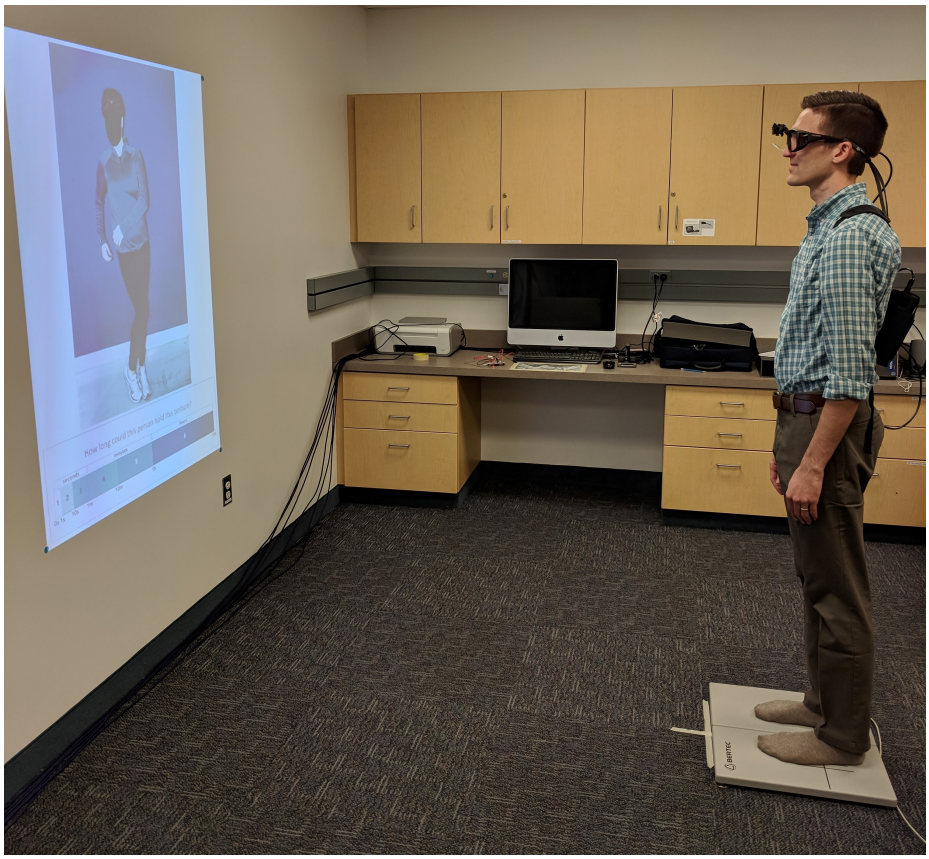


Fig. 7: A demonstration of how a subject's eye-tracking data is being recorded while the subject determines how long an actor could stay balanced in the displayed posture.

Among other variables, the eye tracker recorded the x and y coordinates of the subject's gaze points in 30 Hz video output. Once extracted from the individual video frames, the gaze points for each of the 22 postures were mapped to master images using an algorithm and software developed by Li (2017). The subjects also stood on a force plate which measured postural sway throughout each subject's recording.

During the data collection process, it was necessary to replace data from several participants with data from new participants for a variety of reasons. Some of the reasons were more technological and others were due to confounding factors arising during the recording of the data. Overall, eight treatment and two control participants were replaced until a successful recording was achieved. In total, 57 subjects

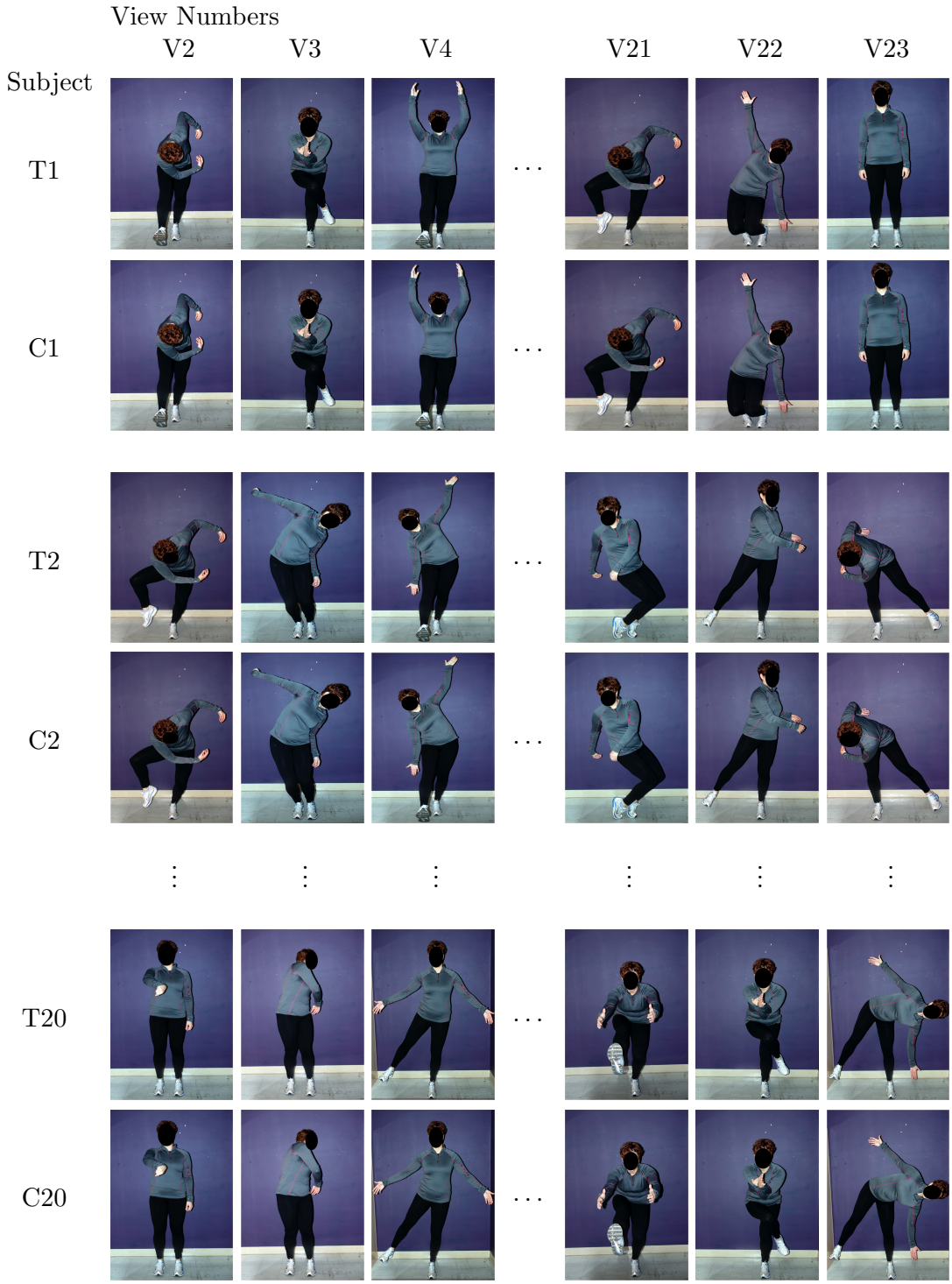


Fig. 8: A visualization of the posture lineups. The rows represent each subject indicated as either a treatment (T) or a control (C) followed by an index number. The columns assign a "View Number" (V followed by an index) representing the order in which subjects viewed the postures. Views one and 24 (V1 and V24) were identical calibration images (Figures 90 and 113 in Appendix C) for all of the subjects. Notice that each treatment subject had a corresponding control subject who was shown the 22 postures in the same order. However, the orders were randomized across subject-pairs within each group.

participated 17 of which were deemed unfit for subsequent data processing. Most often, if a participant's initial or final calibration data was not properly recorded, a new participant was recruited and provided with the same slide ordering (and Subject ID) as the previous participant. Two treatment participants did not complete the experiment due to feeling dizzy. Other reasons for replacing participant's data include the following: participants unconsciously manipulating the eye-tracking equipment during the experiment, data corruption during removal of the eye-tracking device, and miscommunication between participants and the data recorder. In the end, only participants who followed the experiment instructions, provided congruent initial and final calibration data, and who's eye-tracker recordings demonstrated a clean capture of the participant's eye movements throughout the experiment were used in subsequent analyses. Otherwise, their data was replaced by data from a new participant. Figure 9 shows a side-by-side comparison of the gaze point scatterplots. The comparisons are made between gaze scatterplots for posture IDs 2 (top row), 20 (middle row), and 19 (bottom row), which are analyzed further in Sections 7.1.2 and 7.1.4. While the top row compares all of the points between the treatment (left) and control (right) groups, the middle and bottom rows compare gaze scatterplots between two individual subjects.

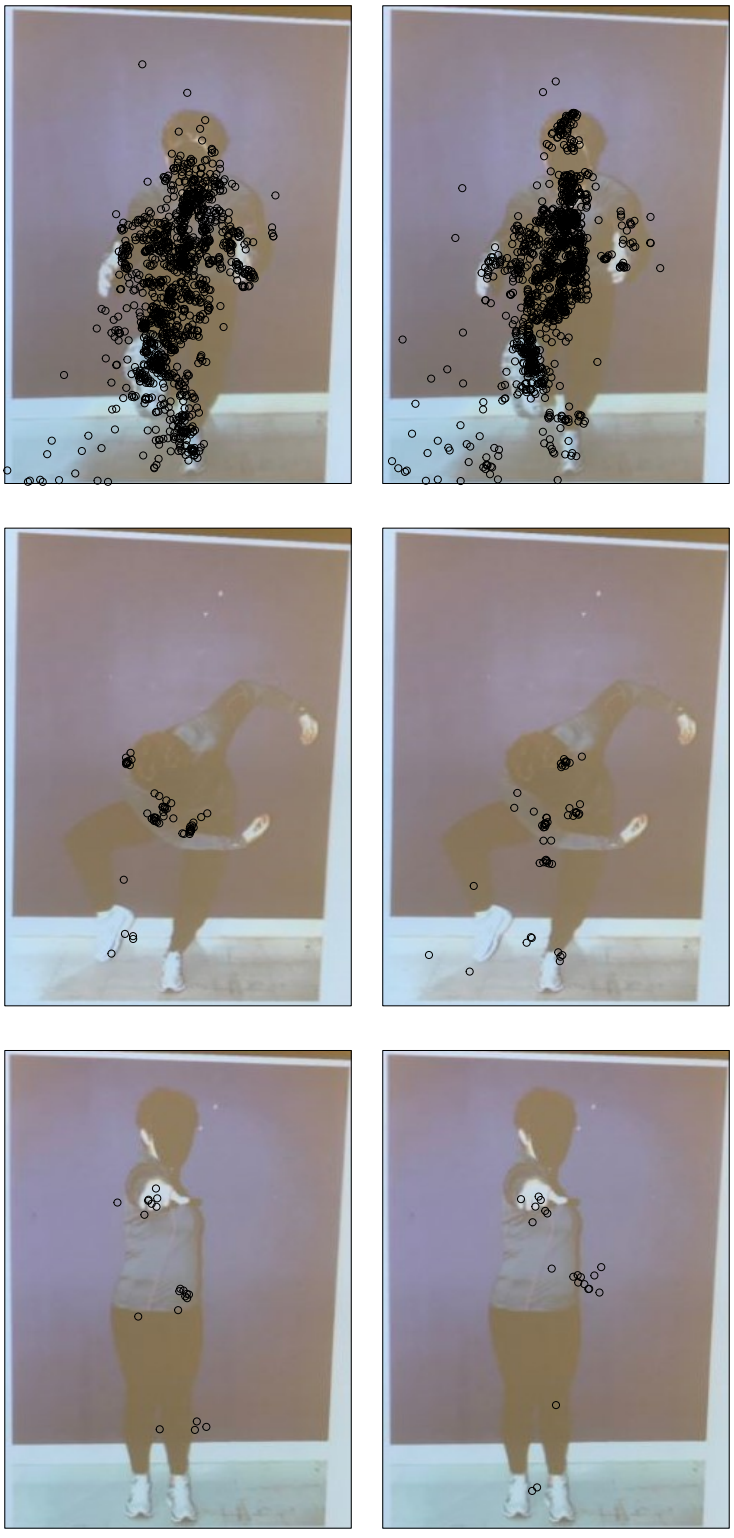


Fig. 9: Comparisons of the gaze scatterplots for posture IDs 2 (top row), 20 (middle row), and 19 (bottom row), which are analyzed further in Sections 7.1.2 and 7.1.4. While the top row compares all of the points between the treatment (left) and control (right) groups, the middle and bottom rows compare gaze scatterplots between two individual subjects.

CHAPTER 5

Modifications to the Syrjala Test

This chapter details several proposed modifications to the Syrjala test, which is introduced in Section 2.2.1. The modifications not only make the new tests more generally applicable, but several are also shown to be more powerful and more conservative than alternative methods, including the original Syrjala test. The modifications include lifting the restriction for identical sampling locations between the two samples (Section 5.1), exploration of six test statistics (also in Section 5.1), a generalization to the rotations within the original Syrjala test (Section 5.2), the introduction of toroidal shifts within the test (Section 5.3), and a combination of both rotational and toroidal shifts (Section 5.4). Section 5.5 details how these modifications fit within the context of the test being a permutation test.

5.1 Motivation and Details

The Syrjala (1996) test checks for equality between normalized distributions from bivariate two-sample data. Furthermore, "The random variable in this case is the observed density at the sampling location, not the location itself." (Syrjala, 1996). Consequently, it requires the two samples both occur at an identical set of predefined locations. The test also suffers from being overly conservative (Fuller et al., 2006).

While useful in its own right, researchers have attempted to apply the Syrjala test to other scenarios by use of data aggregation steps (Chetverikov et al., 2018; McAdam et al., 2012). However, the Syrjala test has been shown to depend on the data aggregation steps such as binning (McKinney and Symanzik, 2019).

Four modifications are proposed for the Syrjala test: (1) removing the restriction

of identical sampling locations between the two samples, (2) exploring three different weights for both the squared and absolute differences in the empirical cumulative distribution functions (i.e., six total combinations of weights and differences), (3) extending the rotational component of the original Syrjala test to higher than four rotations (discussed in Section 5.2), and (4) implementing the use of toroidal shifts of the data within the test (discussed in Section 5.3). A combination of both the rotational and toroidal shift modifications is detailed in Section 5.4. These modifications differ from those proposed by McAdam et al. (2012) who extended the Syrjala test to telemetry data. No other proposed modifications to the Syrjala test were found in the literature.

Extending the notation introduced in Chapter 2, let $(X_{1,1}, Y_{1,1}), (X_{1,2}, Y_{1,2}), \ldots, (X_{1,n_1}, Y_{1,n_1})$ and $(X_{2,1}, Y_{2,1}), (X_{2,2}, Y_{2,2}), \ldots, (X_{2,n_2}, Y_{2,n_2})$ be two independent random samples with unknown distribution functions $F_1(x, y)$ and $F_2(x, y)$ and bivariate empirical cumulative distribution functions (ECDFs) $\Gamma_1^*(x, y)$ and $\Gamma_2^*(x, y)$, respectively. Then the hypotheses under consideration are as follows:

H_0: $F_1(x, y) = F_2(x, y) \quad \forall (x, y)$

H_a: $F_1(x, y) \neq F_2(x, y)$ for at least one coordinate pair (x, y).

In contrast to the Syrjala test, $\Gamma_1^*(x, y)$ and $\Gamma_2^*(x, y)$ in this test are evaluate at each sampling location within their respective samples instead of at identical sampling locations from the two samples. Also let, $n_T = n_1 + n_2$ and $D_{g,k} = \Gamma_1^*(x_{g,k}, y_{g,k}) - \Gamma_2^*(x_{g,k}, y_{g,k}); g = 1, 2$ and k be the observation index. From here, a series of six statistics are proposed as follows:

$$\xi^{DWS} = \frac{n_1}{n_T} \sum_{i=1}^{n_1} [D_{1,i}]^2 + \frac{n_2}{n_T} \sum_{j=1}^{n_2} [D_{2,j}]^2 \tag{2}$$

$$\xi^{UWS} = \sum_{i=1}^{n_1} [D_{1,i}]^2 + \sum_{j=1}^{n_2} [D_{2,j}]^2 \tag{3}$$

$$\xi^{CWS} = \frac{n_2}{n_T} \sum_{i=1}^{n_1} [D_{1,i}]^2 + \frac{n_1}{n_T} \sum_{j=1}^{n_2} [D_{2,j}]^2 \tag{4}$$

$$\xi^{DWA} = \frac{n_1}{n_T} \sum_{i=1}^{n_1} |D_{1,i}| + \frac{n_2}{n_T} \sum_{j=1}^{n_2} |D_{2,j}| \tag{5}$$

$$\xi^{UWA} = \sum_{i=1}^{n_1} |D_{1,i}| + \sum_{j=1}^{n_2} |D_{2,j}| \tag{6}$$

$$\xi^{CWA} = \frac{n_2}{n_T} \sum_{i=1}^{n_1} |D_{1,i}| + \frac{n_1}{n_T} \sum_{j=1}^{n_2} |D_{2,j}| \tag{7}$$

These statistics explore the use of three different types of ECDF weightings, namely, double (DW) (Equations 2 and 5), uniform (UW) (Equations 3 and 6), and complementary (CW) (Equations 4 and 7) weightings along with squared (S) (Equations 2–4) vs. absolute (A) (Equations 5–7) differences between the ECDFs. The six statistics are chosen to further explore the effects of squared vs. absolute ECDF differences along with a series of weightings of the respective differences. The respective statistics ($\xi^{DWS}, \xi^{UWS}, \xi^{CWS}, \xi^{DWA}, \xi^{UWA},$ and ξ^{CWA}) are referred to generally as ξ^* statistics. Note that the CWS and CWA statistics were previously called RWS and RWA in McKinney and Symanzik (2021).

Justification for the names given to each of the weightings can be found in the context of multi-criteria decision making. If each of the summations across the respec-

tive sample sizes are considered as separate entities which contribute to the overall decision of statistical significance, then the summation coefficient weightings in Equations 2–7 can be considered as a ratio weighting method (Edwards, 1977; Zardari et al., 2015). Hence, double weightings refers to the scaling ratio $\frac{n_1}{n_T}$ being multiplied to the sum across the first sample index (i), and $\frac{n_2}{n_T}$ being multiplied to the sum across the second sample index (j). Furthermore, these sums are considered "double" weighted since any difference in the sample sizes, which would result in a differing number of terms between the sums, would be exaggerated by the ratio of the sample size by the pooled sample size. Since equal weightings would result in an identical coefficient which could be factored from both summations resulting in a unnecessary scalar to the test statistic, the uniform weightings, also called mean weightings (Zardari et al., 2015), simply omit these scaling ratios. Additionally, complementary weightings apply the scaling ratios to the opposite sums as the double weightings since $\frac{n_2}{n_T}$ is the compliment of $\frac{n_1}{n_T}$, i.e., $\frac{n_2}{n_T} = 1 - \frac{n_1}{n_T}$, and vice versa. Complementary weightings have also been used in other contexts (Rey, 1986; Lai et al., 2005).

Similarities can be seen between Equations 2–7 (especially Equation 3) and the univariate two-sample Cramer-von Mises test statistic (Chapter 2 Equation 1) from which the Syrjala test is also an extension of (Syrjala, 1996). However, the assumption for identical sampling locations found within the Syrjala test has been lifted.

Two additional modifications are studied as well, namely an extension of the original four rotations (detailed in Section 5.2), and the introduction of a toroidal shift modification (proposed in Section 5.3). While much more computationally intensive, a combination of both rotational and toroidal shift modifications is also explored in Section 5.4.

5.2 Rotational Modification

Due to the nature of bivariate data, the origin of the bivariate ECDF is defined as the data value which falls below and furthest to the left of all of the sampled data in the Cartesian plane. Consequently, the original Syrjala test was rotated four times as an attempt to remove a dependency of the test on data which lies closer to this origin. However, the extent to which these four rotations corrected this issue has not been explored. Therefore, we propose a more generalized statistic which rotates the sampled data R times (instead of four times). Hence, the test statistic can be written as

$$\Psi^R = \frac{1}{R}\sum_{r=1}^{R} \xi_r^*, \tag{8}$$

where ξ_r^* is a redefined rotational version of one of the six ξ^* statistics defined by Equations 2–7, and R is a discrete number of rotations within 360°. In doing so, a new ECDF will be generated for each of the r rotations within ξ_r^*. Hence, this modification replaces each Γ_1^* and Γ_2^* within Equations 2–7 with $\Gamma_{1,r}^*$ and $\Gamma_{2,r}^*$, respectively. When $r = 1$ the original orientation of the data is used. Hence, $R = 1$ is the case when no rotation is applied to the pooled samples, and $R = 4$ is the case which the original Syrjala test employed (four 90° rotations).

Our modified test statistic (Ψ^R) first computes the squared or absolute difference between the bivariate ECDFs evaluated at all of the data from the first sample. This sum of squared or absolute differences is weighted depending on the . This process is repeated for the second sample, and the two weighted sums of squared or absolute differences are then combined. Next, the data are rotated $360/R$ degrees, and another weighted average of the sums of squared or absolute differences of the ECDF values is computed. This computation is repeated for a total of R rotations (using $\Gamma_{1,r}^*$ and

$\Gamma^*_{2,r}$ for each of the r rotations), where each rotation is weighted by $1/R$. This is the computation of the test statistic on the original data.

A visualization of the calculation of our modified Syrjala test can be seen in Figure 10. The top left graph in Figure 10 highlights three points from the two samples. The highlighted vertical bars seen between the two ECDFs in the bottom left graph represent the differences between the ECDFs evaluated at the respective highlighted points. The remaining two columns in Figure 10 suggest similarly made calculations (on the same highlighted points), but for rotated versions of the data. In this case, the data are being rotated every $40°$ for a total of $R = 360/40 = 9$ rotations. However, only the first two rotations are shown in Figure 10.

It should be noted that the bottom row of graphs in Figure 10 displays only the marginal ECDFs for each sample (and not the bivariate ECDFs). However, the difference between overlapping bivariate ECDFs is difficult to represent visually. Hence, the marginal ECDFs are shown for visualization purposes only. Figure 11 compares the two bivariate ECDFs for the same data used in Figure 10.

Figure 12 outlines the process in which the rotational modification is integrated into the modified Syrjala tests.

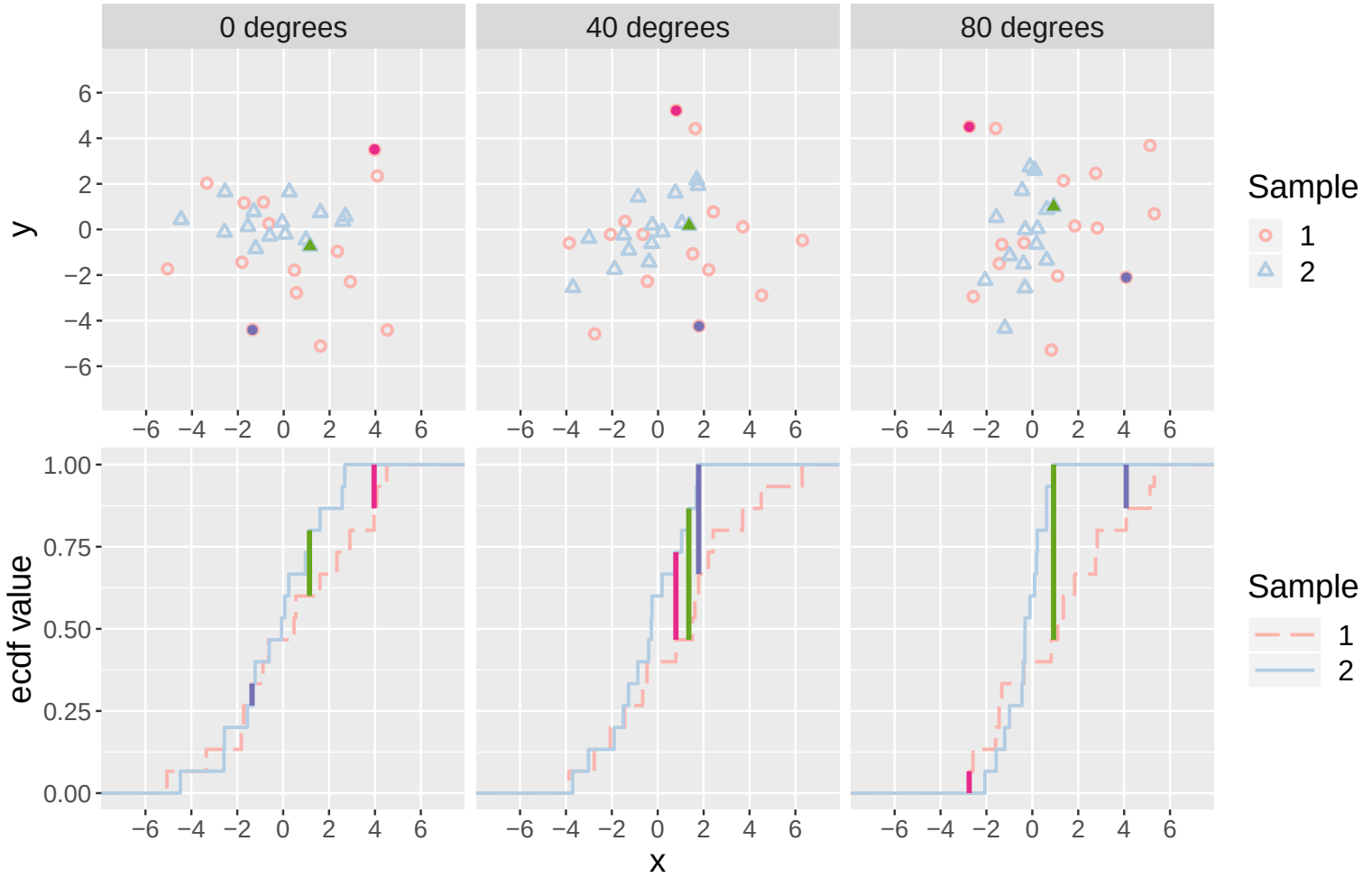


Fig. 10: A visualization of calculations within the statistics of the modified Syrjala tests. The same three demonstrative colored points (two from sample one, and one from sample two) are highlighted in the scatter plots (top row) across three different rotations of the data. The bottom row of graphs highlights three differences (vertical colored bars) between the ECDFs. Each ECDF difference (below) corresponds to a highlighted scatter plot point (above). While only three points and differences are highlighted, the calculation involves squared differences between ECDFs across all of the points from both samples. The bottom row shows differences between the marginal (and not bivariate) ECDFs. This is due to the difficult nature of visualizing differences in overlapping bivariate ECDFs. Hence, the marginal ECDFs are displayed for visualization purposes only. A comparison of the bivariate ECDFs is shown in Figure 11.

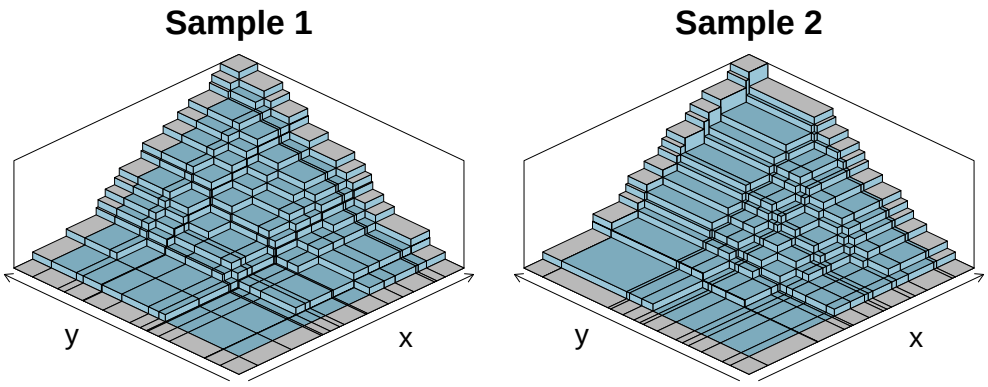


Fig. 11: A visualization of the two bivariate ECDFs for the non-rotated samples shown in Figure 10.

5.3 Toroidal Shift Modification

McAdam et al. (2012) noticed a reduced emphasis that the Syrjala test places on observed differences located near the center of the bounding region. This is confirmed in Chapter 6. Hence, in addition to removing the necessity for common sampling locations, an additional modification is employed which using toroidal shifts. This modification also addresses the ECDF origin issue (see Section 5.2). Hence, the toroidal shift modification is first considered here without the rotational modification. In the next section, a combination of both the rotational and toroidal modifications is presented.

The toroidal shift is a well established technique in the spatial statistics literature (Diggle and Milne, 1983; Upton et al., 1985; Berman, 1986; Díaz et al., 2008; Dixon, 2014; Moreno-Fernández et al., 2020), and was first suggested by Lotwick and Silverman (1982). Upton et al. (1985) defined a toroidal shift, along with its use in non-parametric statistics. A toroidal shift has been used by Díaz et al. (2008) in their generalization of the Ripley (1976) K-function to a spatio-temporal dispersion measure.

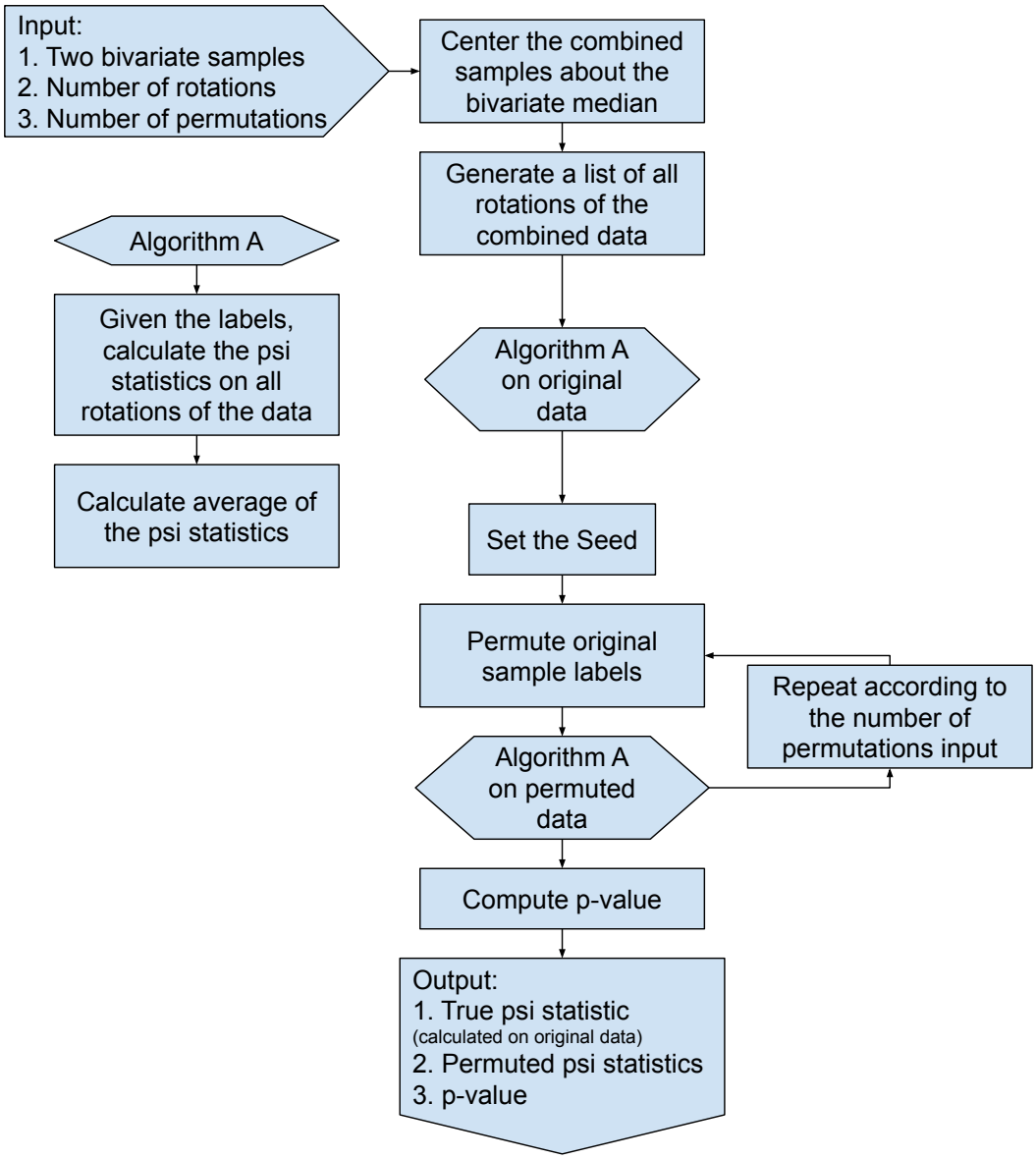


Fig. 12: A flowchart which displays the process in which the rotational modification is integrated into the modified Syrjala tests. The psi statistics computed on the original data and permuted data referred to in the figure are the Ψ^R and Ψ_l^R discussed in Sections 5.2 and 5.5, respectively.

A toroidal shift is accomplished by first treating the bounding rectangle of the data as a torus. This is equivalent to forming a bounding rectangle around the data and wrapping it such that the left and right edges join, and the top and bottom edges join. This wrapping of the horizontal and vertical axes will form a donut-shaped torus. Hence, the data on the left edge will now be considered as "close" to the data on the right edge of the bounding rectangle. This affect is applied similarly to data near the top or bottom edges of the bounding rectangle.

A common approach (Upton et al., 1985) for achieving the shift is accomplished by randomly sampling a $\Delta x \sim \text{Uniform}(0, \max(x) - \min(x))$ and $\Delta y \sim \text{Uniform}(0, \max(y) - \min(y))$, and adding Δx and Δy to every data value's x and y coordinate, respectively. If the shift moves a data value outside of the bounding rectangle, then the data value will be replaced on the opposite side of the bounding rectangle.

However, in this dissertation the toroidal shift is applied in a slightly different manner than Upton et al. (1985) in order to ensure that every ECDF of the shifted data is equally likely. Instead of randomly sampling a Δx and Δy, a random data value is selected as the origin of the toroidal shift. All data values to the left of the selected data value are shifted horizontally by adding a distance equal to the width of the bounding rectangle $(\max(x) - \min(x))$ to their respective x coordinates. A similar shift is applied to the data values below the selected data value except the height of the bounding rectangle $(\max(y) - \min(y))$ is added to the y coordinates. While this results not only in shifted data, but also in a shifted bounded rectangle, the subsequent ECDF calculations do not depend on the relative position of the data.

Figure 13 shows two subject's data which undergo the latter described random toroidal shift. The center plot in Figure 13 shows a random point which is chosen as the origin of the toroidal transformation. Any point which lies below this toroidal origin is shifted up equal the height of the bounding rectangle of the data. Similarly,

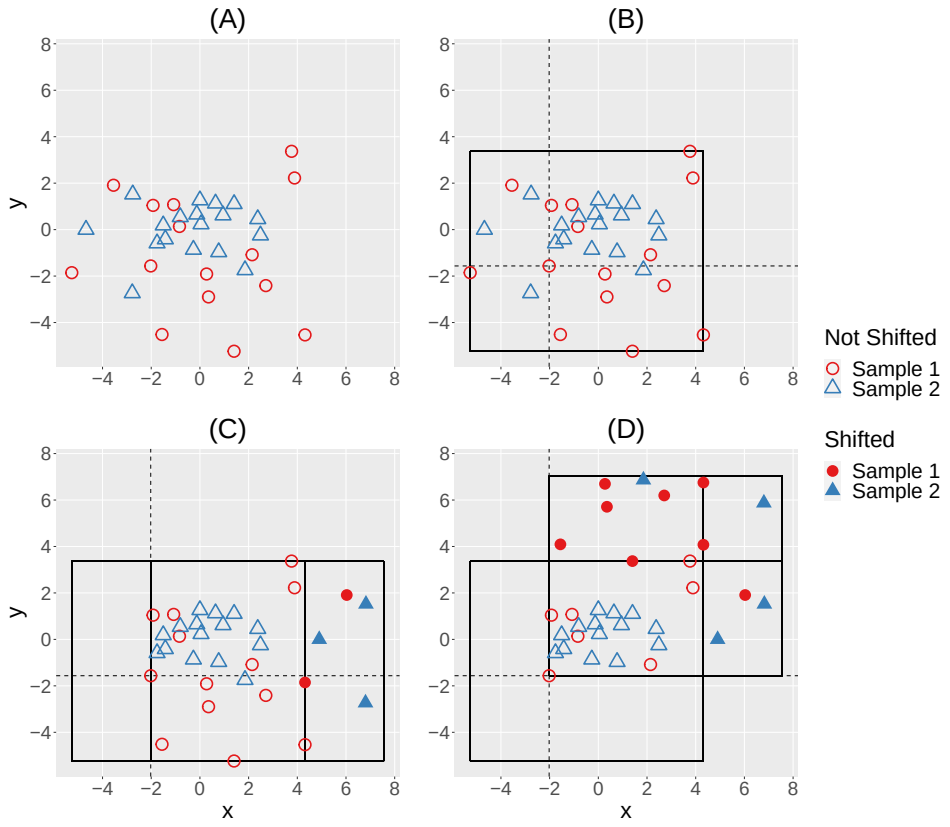


Fig. 13: A visualization of two-sample data before (A) and after (D) a toroidal shift transformation. Plot (B) shows a bounding rectangle around combined samples along with the randomly selected data value which serves as the origin of the toroidal shift. Plot (C) shows an intermediate step within the toroidal shift where only the horizontal shift has occurred. Plot (D) completes the toroidal shift with a subsequent vertical shift. The data values unaffected by the toroidal shift are indicated by hollow circles for sample 1 and hollow triangles for sample 2, whereas those affected by the toroidal transformation are indicated by their respective filled-in shapes.

points which lie to the left of the origin are shifted to the right a distance equal the width of the bounding rectangle of the data. Consequently, data values which lie both to the left and below the toroidal origin will experience both of the previously mentioned shifts.

These new toroidal shifted data provides the basis for an additional modification to the Syrjala test. After the data are transformed using the toroidal shift, the test statistic of choice ξ_t^* is computed, where ξ_t^* is one of the six statistics defined by

Equations 2–7. As in the rotational modification detailed in Section 5.2, a new ECDF will be generated for each of the toroidal shifts within ξ_t^*. Hence, this modification replaces each Γ_1^* and Γ_2^* within Equations 2–7 with $\Gamma_{1,t}^*$ and $\Gamma_{2,t}^*$, respectively.

This calculation can be applied across all possible toroidal shifts, or a large random subset if all possible shifts are computationally infeasible. This is similar to the modification shown in Equation 8, except the individual computations are weighted according to the number of toroidal shifts R_T, as seen in the following equation:

$$\Psi^T = \frac{1}{R_T} \sum_{t=1}^{R_T} \xi_t^*. \tag{9}$$

Figure 14 outlines the process in which a toroidal shift modification is integrated into the modified Syrjala tests.

A random toroidal shift has been implemented in the `spatstat` R package (Baddeley and Turner, 2005).

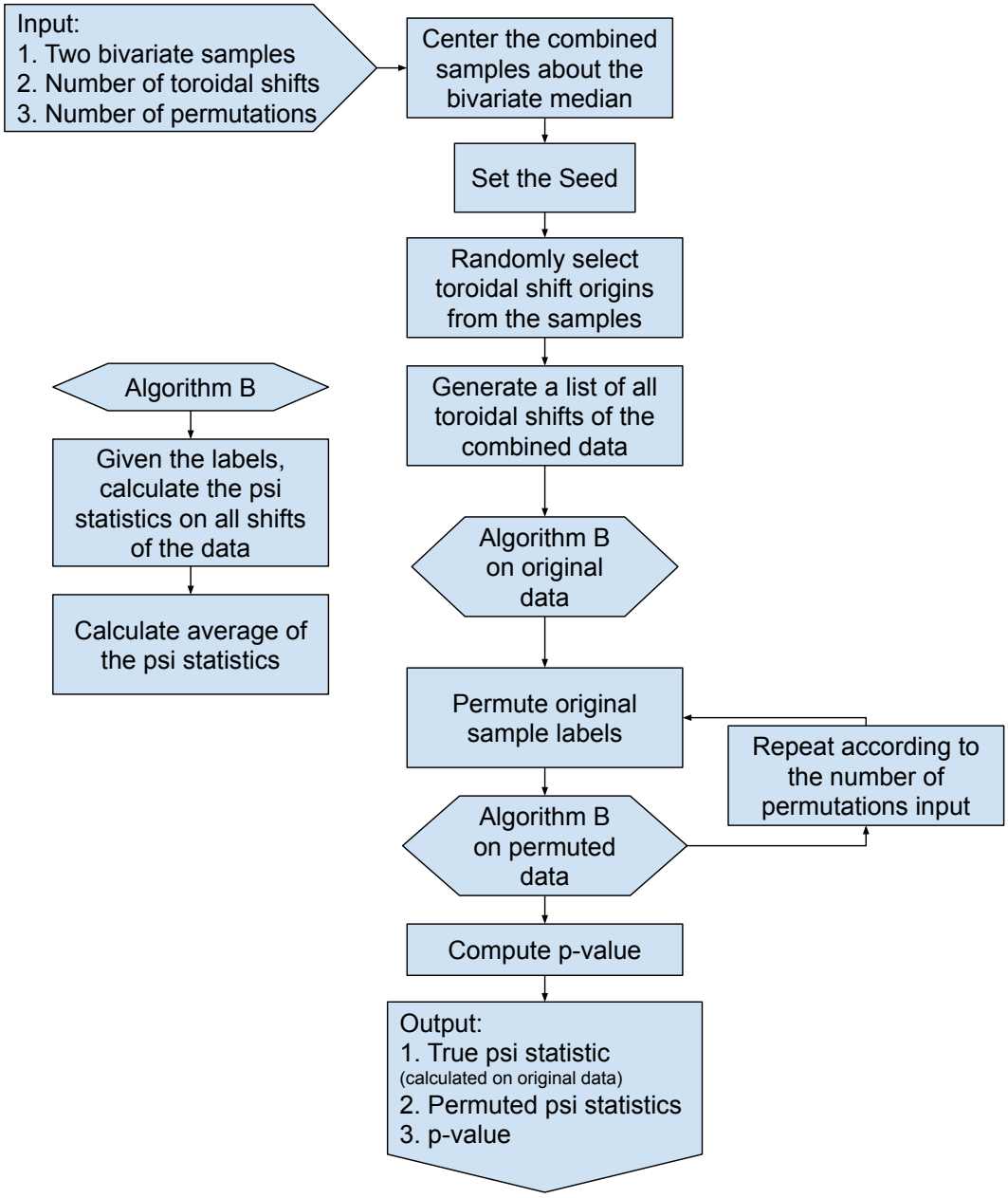


Fig. 14: A flowchart which displays the process in which a toroidal shift modification is integrated into the modified Syrjala tests. The psi statistics computed on the original data and permuted data referred to in the figure are the Ψ^T and Ψ_l^T discussed in Sections 5.3 and 5.5, respectively.

5.4 Combining Modifications

In an effort to remove both the dependency of the Syrjala test on the ECDF's origin (see Section 5.2 for more details) while also alleviating the reduced emphasis the Syrjala test places on the observed differences near the center of the bounding rectangle (McAdam et al., 2012), an additional modification is considered which combines both the rotational (see Section 5.2) and toroidal shift (see Section 5.3) modifications.

Due to computational efficiency, the rotational modification is applied first within the test. This also removes the need to recenter the data around the bivariate median since differences in the bivariate ECDFs computed after the toroidal shift will be the same regardless of relative position to the origin. Hence, for every rotation of the combined data, the toroidal modification is applied separately. Consequently, combining the modifications in Equations 8 and 9 gives us the following equation:

$$\Psi^{RT} = \frac{1}{R \cdot R_T} \sum_{r=1}^{R} \sum_{t=1}^{R_T} \xi^*_{r,t},$$

where $\xi^*_{r,t}$ is one of the six statistics defined by Equations 2–7. Similar to Equations 8 and 9, this combined modification redefines Γ^*_1 and Γ^*_2 in Equations 2–7 as $\Gamma^*_{1,r,t}$ and $\Gamma^*_{2,r,t}$, respectively, for each of the r rotations and t toroidal shifts.

Figure 15 outlines the process in which a toroidal shift modification is integrated into the modified Syrjala tests.

5.5 Permutation Test Computations

Let Ψ^* be one of the previously discussed test statistics Ψ^R, Ψ^T, or Ψ^{RT} (see Sections 5.2–5.4). As a permutation test, the test statistic $\Psi^*_l; l = 1, \ldots, N_{\max}$, is recalculated $N_{\max} = \frac{n_T!}{n_1! n_2!}$ times where n_1 and n_2 are the respective sample sizes, $n_T = n_1 + n_2$, and $N_{\max}$ is the total number of permutations of the sample labeling

subscripts. However, in practice, computing Ψ_l^* for all $l = 1, \ldots, N_{\max}$ is computationally infeasible, and a sufficient $N \ll N_{\max}$ (e.g., $N \approx 999$) are computed instead.

The p-value is calculated as the total proportion of test statistics Ψ_l^* which are greater than or equal to the statistic Ψ^* computed from the non-permuted data, i.e.,

$$p - value = \frac{\sum_{l=1}^{N} \left(I_{\Psi_l^* \geq \Psi^*} \right) + 1}{N + 1}.$$

where $I_{\Psi_l^* \geq \Psi^*}$ is one if $\Psi_l^* \geq \Psi^*$ and zero otherwise.

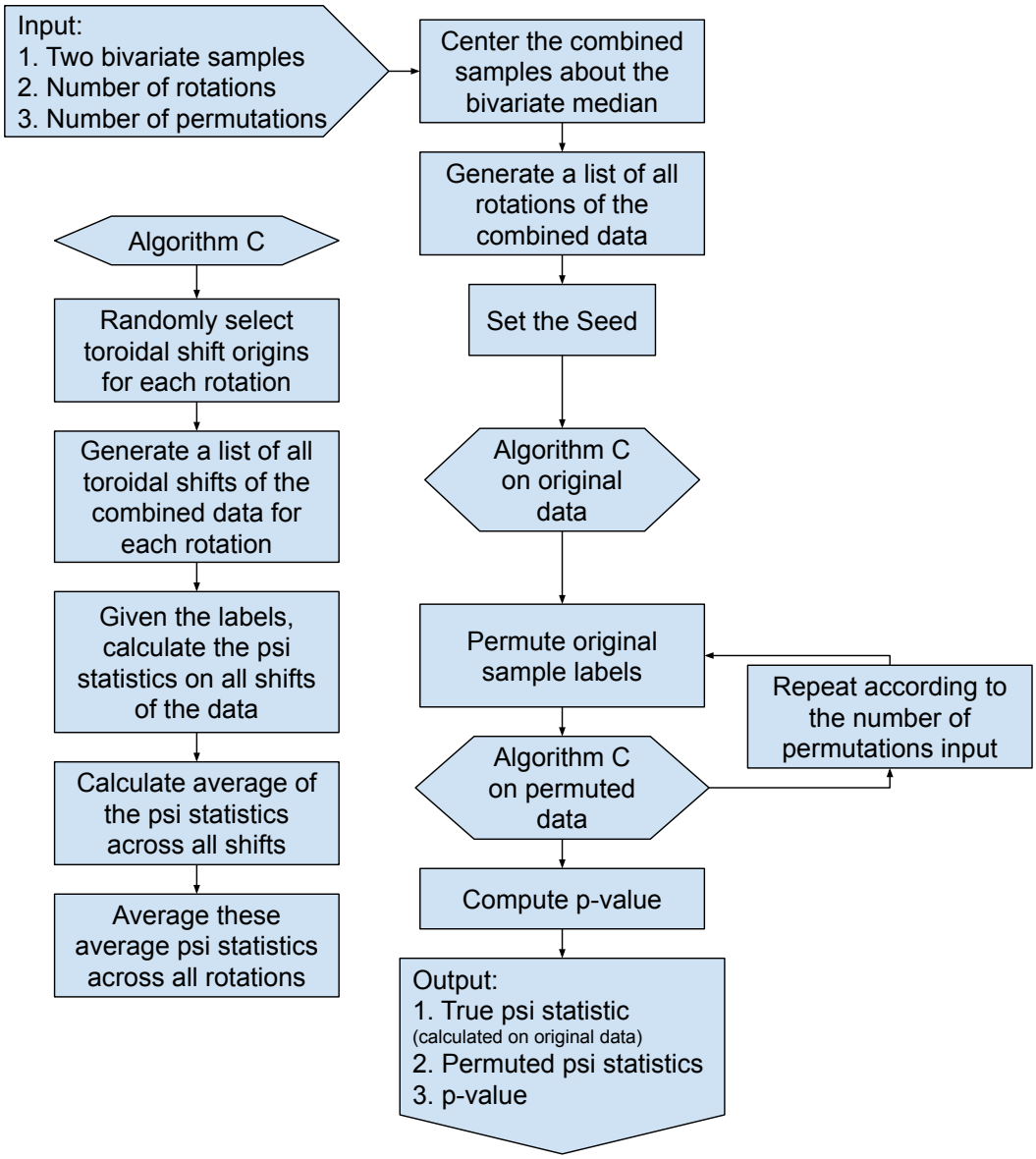


Fig. 15: A flowchart which displays the process in which a combination of both the rotational and toroidal shift modifications are integrated into the modified Syrjala tests. The psi statistics computed on the original data and permuted data referred to in the figure are the Ψ^{RT} and Ψ_l^{RT} discussed in Sections 5.4 and 5.5, respectively.

CHAPTER 6

Simulation Studies

This chapter outlines four simulation studies. The first (in Section 6.2) investigates the effects of two data binning techniques on the outcome of the Syrjala test (see Section 2.2.1 for more details on the Syrjala test itself), and is an extension of the research conducted by McKinney and Symanzik (2019). The second (in Section 6.3) explores the performance of three proposed modifications of the Syrjala test (outlined in Chapter 5). The third (in Section 6.4) compares the performances of the modified Syrjala test which employs both the rotational and toroidal shift modifications to four other multivariate two-sample tests (discussed in Sections 2.2.1–2.2.4) including the Syrjala test. A fourth simulation (in Section 6.5), which employs multimodal bivariate mixture distributions, demonstrates the appropriateness of applying the modified Syrjala tests to eye-tracking data. Before each study is discussed in detail, an overview of the simulation design including generated data structure, reproducibility, and the use of common random numbers is discussed in Section 6.1.

6.1 Simulation Design

The following subsections outline the setup of the simulation studies including the structure of the generated data both for when the null hypotheses are true and otherwise. Additional details are provided on the reproducibility of both the generated data and simulation results, as well as the use of common random numbers in order to reduce the overall variability in the simulation results.

Additionally, as the power, false positive rate, conservative or anti-conservative nature of a test are referred to in this chapter, those definitions are provided here:

- The power of a test is the probability of rejecting the null hypothesis when the null is indeed false (Rice, 2006).
- The false positive rate (or type-1 error rate) of a test is the probability of rejecting the null when the null hypothesis is, in fact, true (Rice, 2006).
- A conservative test is one that rejects the null at a lower rate than the significance level (Rice, 2006) when the null hypothesis is true. Hence, an anti-conservative rate rejects the null at a higher rate than the significance level when the null hypothesis is true.

6.1.1 Generated Data Structure

For the simulations discussed in Sections 6.2 and 6.3, two realizations of independent, uniformly distributed, i.e., completely spatially random (CSR), data were simulated on $[0,1] \times [0,1]$ square regions to assess the tests when the null hypothesis is true. A bivariate CSR datum of this form is generated by assigning two univariate uniform random values (defined on $[0,1]$) to their Cartesian coordinates, respectively.

To assess the tests when the null hypothesis is false, four other bivariate distributions were employed, each of which was compared to CSR. The four departures from CSR (also simulated on the $[0,1] \times [0,1]$ square) were constructed using the following intensity functions for the heterogeneous Poisson process where the values a_1, a_2, a_3, and a_4 are height parameters.

$$f_1(x,y) = a_1 \cdot \exp\Big\{ -20 \cdot \big[(x-0.5)^2 + (y-0.5)^2\big] \Big\} \qquad \text{(Center)}$$

$$f_2(x,y) = a_2 \cdot \Big(1 - \exp\Big\{ -80 \cdot \big[(x-0.5)^4 + (y-0.5)^4\big] \Big\}\Big) \qquad \text{(Repel)}$$

$$f_3(x,y) = a_3 \cdot \exp\Big\{ -5 \cdot \big[(x-1)^2 + (y-1)^2\big] \Big\} \qquad \text{(Corner)}$$

$$f_4(x,y) = a_4 \cdot \exp\Big\{ -5 \cdot (x-1)^2 \Big\} \qquad \text{(Right)}$$

Let μ be the average number of points within the unit square for the heterogeneous Poisson process. For reproducibility, Table 2 shows the values for the height parameters a_1, a_2, a_3, and a_4 that achieve a specified intensity μ for each departure from CSR. The coefficients within the exponents of each of the intensity functions were chosen to ensure a sufficient departure from CSR was simulated. These coefficients also guarantee at least 97% of the volume under each bivariate intensity function lie within the unit square. For each of the five comparisons (CSR compared with CSR, Center, Repel, Corner, Right), CSR realizations of 50, 100, 250, and 500 points were compared to each of four different sample sizes (also 50, 100, 250, and 500) for each of the comparison distributions. Additionally, ten realizations were generated for each comparison in the simulation data. Visualizations of the CSR and heterogeneous Poisson process realizations (with $\mu = 500$ points) using each one of the intensity functions (referred to as CSR, Center, Repel, Corner, and Right, respectively) can be seen as a column of graphs in the far right of the figures found in Section 6.2 and 6.3.

However, to pattern the distributions more closely to data recorded in eye-tracking research, a collection of multimodal bivariate mixture distributions were chosen for the simulations discussed in Section 6.5. See Section 6.5.1 for more details.

Table 2: A table of the height parameter values (a_1, a_2, a_3, and a_4) which achieve a desired average number of points within the unit square (μ) for each respective intensity function.

μ	a_1	a_2	a_3	a_4
50	319	79	319	126
100	639	158	639	253
250	1597	395	1597	632
500	3193	790	3193	1264

6.1.2 Generated Data Reproducibility

In order to ensure reproducibility of the simulation results, the simulation data is generated up front with predefined random number seeds. The unique random number seeds allow the computational researcher the freedom to reproduce any of the individual data realizations or simulation statistics.

The default random number seeds used within the `R` computational environment consist of 626 32-bit integers (see the documentation for the `.Random.seed` object in `R` at `https://stat.ethz.ch/R-manual/R-devel/library/base/html/Random.html`). Since these can be unwieldy to handle directly, R provides a `set.seed()` function which maps individually provided 32-bit integers to well dispersed random number seeds. Hence, in order to preserve reproducibility, each of the four characteristics in the data generation process were mapped to integer values which were then concatenated into a single integer provided to the `set.seed()` function. The four characteristics are as follows: (1) the sample size of the first sample (n_1), (2) the distribution shape of the second sample (see Section 6.1.1 for more details), (3) the sample size of the second sample (n_2), and (4) a replication index number. However, the absolute value of the concatenated integers must be less than $\frac{2^{32}}{2} - 1 = 2147483647$ in order to fit within 32-bits of memory. (The division by two allows for one bit to represent the sign of the integer, and the subtraction of one allows for zero to be represented.) Hence, while allowing room for possible extensions of the simulation, yet staying within memory limits, the following number of digits, or "slots", were set aside for each data generation characteristic, respectively: two, two, two, and four assigned from left to right. Tables 3–6 show the assigned integer values for each level of the respective generated data characteristics.

For example, the first replication of two-sample data generated where the first sample has 500 CSR data values, and the second sample has 50 CSR data values had

Table 3: A table showing the two digit integer values assigned to the first two slots of a 32-bit integer for each level of n_1.

n_1	Assigned Integer
500	01
250	02
100	03
50	04

Table 4: A table showing the two digit integer values assigned to the the third and fourth slots of a 32-bit integer for each level of sample 2 distribution names.

Sample 2 Distribution Name	Assigned Integer
CSR	01
Center	02
Repel	03
Corner	04
Right	05

a `set.seed()` concatenated integer argument equal to "01-01-04-0001" or 101040001.

Thus, a researcher could easily generate the data necessary and extend the simulations to broader comparisons, e.g., if a single comparison was of interest between a CSR distribution with $n_1 = 25$ and a Repel distribution with $n_2 = 750$, additional two-slot integers could be assigned to these new levels of n_1 and n_2, say 06 and 07 respectively, and the resulting concatenated integer argument for `set.seed()` would be "06-03-07-0001" or 603070001. Hence, keeping track of these integer arguments for `set.seed()` allows for the research to be extended in a systematic way which avoids reusing identical random number seeds in generating additional simulated data scenarios.

6.1.3 Common Random Numbers

To reduce the overall variability of the statistics on the simulated data, the method of common random numbers (CRNs) is employed. CRNs (also called corre-

Table 5: A table showing the two digit integer values assigned to the the fifth and sixth slots of a 32-bit integer for each level of approximate n_2.

Approx. n_2	Assigned Integer
500	01
250	02
100	03
50	04

Table 6: A table showing the four digit integer values assigned to the the seventh through the tenth slots of a 32-bit integer for each replication number.

Replication Number	Assigned Integer
1	0001
2	0002
3	0003
4	0004
5	0005
⋮	⋮
10	0010

lated sampling, matched sampling, or matched pairs) are a variance reduction technique commonly employed in Monte Carlo simulations (Glasserman, 2013; Botev and Ridder, 2017), and are well established within the statistical simulation community (Kleijnen, 1975, 1976, 1979). In essence, when making comparisons between different configurations within a Monte Carlo simulation, CRNs ensure that any one realization of a random variable is used in the same way across all of the configurations. Hence, the same randomly generated numbers will be used across all of the configurations of the experiment, which reduces the overall variation in the simulation statistics.

In the context of our simulation, CRNs are used to reduce variation in the simulated statistics, such as the Syrjala and modified Syrjala statistics. For example, instead of producing ten realizations of 500 random points to compute the Syrjala statistic with one binning technique (described in more detail below), and then pro-

ducing another ten realizations of 500 random points to compute the Syrjala statistic with another binning technique, the same ten realizations for the first binning technique will also be used with the second binning technique. Thus, if an unusual observation happened to be sampled within the ten realizations, producing a higher amount of variation due to sampling, that same unusual observation would be used across the binning techniques reducing the overall variation due to sampling error in the individual statistics.

Similarly, CRNs also aids in the comparison of the modified Syrjala tests to one another. In later sections within this chapter, it is clear that some of the simulated data for the Repel distribution appears to look similar to the CSR distribution just by chance. However, all of the modified Syrjala tests being compared will have to handle this unusual sample simultaneously. Hence, unusual results exhibited across all of the tests in comparison can be attributed more to chance variation in the data generation process and less to other differences between the tests. Thus, CRNs also provide a decrease in computational time since fewer realizations of randomly generated data are necessary in order to obtain similarly stable results.

Specifically, let $X_1 = X_{1,1}, X_{1,2}, \ldots, X_{1,n}$ be a vector of independent and identically distributed random variables. Similarly, let $X_2 = X_{2,1}, X_{2,2}, \ldots, X_{2,m}$ be a second vector of independent and identically distributed random variables. Say we are interested in estimating the difference in two population parameters by use of the sample statistics $T_1 = \frac{1}{n}\sum_{i=1}^{n} f(X_{1,i})$ and $T_2 = \frac{1}{m}\sum_{j=1}^{m} g(X_{2,j})$, respectively. If common random numbers are not employed, and the samples are independent, then the covariance between all $X_{1,i}$ and $X_{2,j}$ is zero, and the variance of the difference in sample statistics is

$$
\begin{aligned}
Var(T_1 - T_2) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} f(X_{1,i}) - \frac{1}{m}\sum_{j=1}^{m} g(X_{2,j})\right) \\
&= Var\left(\frac{1}{n}\sum_{i=1}^{n} f(X_{1,i})\right) + Var\left(\frac{1}{m}\sum_{j=1}^{m} g(X_{2,j})\right) \\
&= \frac{1}{n^2} Var\left(\sum_{i=1}^{n} f(X_{1,i})\right) + \frac{1}{m^2} Var\left(\sum_{j=1}^{m} g(X_{2,j})\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} Var\left(f(X_{1,i})\right) + \frac{1}{m^2}\sum_{j=1}^{m} Var\left(g(X_{2,j})\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} Var\left(f(X_1)\right) + \frac{1}{m^2}\sum_{j=1}^{m} Var\left(g(X_2)\right) \\
&= \frac{1}{n^2} n Var\left(f(X_1)\right) + \frac{1}{m^2} m Var\left(g(X_2)\right) \\
&= \frac{1}{n} Var\left(f(X_1)\right) + \frac{1}{m} Var\left(g(X_2)\right)
\end{aligned}
$$

However, notice if we employ CRNs by using the same randomly generated data across all of our configurations, i.e., if we let $X_1 = X_2$ (and $n = m$), then this will tend to result in a positive correlation between X_1 and X_2. In such cases, the independence between the samples will be removed, and $Corr(X_1, X_2) > 0 \Rightarrow Cov(X_1, X_2) > 0 \Rightarrow Cov(f(X_1), g(X_2)) > 0 \Rightarrow -Cov(f(X_1), g(X_2)) < 0$.

Therefore, if we let T_1^* and T_2^* be the statistics while employing CRNs, then

$$
\begin{aligned}
Var(T_1^* - T_2^*) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} f(X_{1,i}) - \frac{1}{n}\sum_{j=1}^{n} g(X_{2,j})\right) \\
&= Var\left(\frac{1}{n}\sum_{i=1}^{n} f(X_{1,i})\right) + Var\left(\frac{1}{n}\sum_{j=1}^{n} g(X_{2,j})\right) \\
&\qquad - Cov\left(\frac{1}{n}\sum_{i=1}^{n} f(X_{1,i}), \frac{1}{n}\sum_{j=1}^{n} g(X_{2,j})\right) \\
&= \frac{1}{n^2} Var\left(\sum_{i=1}^{n} f(X_{1,i})\right) + \frac{1}{n^2} Var\left(\sum_{j=1}^{n} g(X_{2,j})\right) \\
&\qquad - \frac{1}{n^2} Cov\left(\sum_{i=1}^{n} f(X_{1,i}), \sum_{j=1}^{n} g(X_{2,j})\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} Var\left(f(X_{1,i})\right) + \frac{1}{n^2}\sum_{j=1}^{n} Var\left(g(X_{2,j})\right) \\
&\qquad - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} Cov\left(f(X_{1,i}), g(X_{2,j})\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} Var\left(f(X_1)\right) + \frac{1}{n^2}\sum_{j=1}^{n} Var\left(g(X_2)\right) \\
&\qquad - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} Cov\left(f(X_1), g(X_2)\right) \\
&= \frac{1}{n^2} n Var\left(f(X_1)\right) + \frac{1}{n^2} n Var\left(g(X_2)\right) \\
&\qquad - \frac{1}{n^2} n^2 Cov\left(f(X_1), g(X_2)\right) \\
&= \frac{1}{n} Var\left(f(X_1)\right) + \frac{1}{n} Var\left(g(X_2)\right) - Cov\left(f(X_1), g(X_2)\right)
\end{aligned}
$$

will ensure that $Var(T_1^* - T_2^*) < Var(T_1 - T_2)$.

However, this is not always the case. If instead a removal of the independence between X_1 and X_2 results in a negative correlation, the use of common random

numbers will back-fire resulting in a greater amount of variability in the difference between test statistics (Botev and Ridder, 2017). Glasserman and Yao (1992) provided guidelines for avoiding these situations.

6.1.4 Simulated Test Result Reproducibility

Similar to how random number seeds are used in Section 6.1.2 to preserve the reproducibility of the generated simulation data, random number seeds are also set before each simulation test to preserve the reproducibility of the stochastic nature within each test. The test random number seeds are simply an extension of the generated data random number seeds, except that the third and fourth digits (hundreths and thousandths place) are changed from two zeros to a two digit number that corresponds to what test is being used. Table 7 shows the assigned integers for each of the simulation test types.

Thus, a researcher could easily reproduce the simulation test results by using the same random number seed. For example, the first replication of two-sample data generated where the first sample has 500 CSR data values, and the second sample has 50 CSR data values had a `set.seed()` concatenated integer argument equal to 101040001. Thus, if the modified Syrjala tests which employ both rotations and toroidal shifts with a toroidal shift threshold of 25 (RotToro25Thrshld) is applied to this case of generated data, 60 would be added to the third and fourth digits, and the random number seed for the simulation test would be "01-01-04-6001" or 101046001. Alternatively, if the modified Syrjala tests employ five rotations and 0.3 proportion of toroidal shifts, the third and forth digits will have an assigned integer of 23 along with a negative sign prepended to the seed number: -101042301.

Table 7: A table showing the two digit integer values assigned to the the third and fourth slots of a 32-bit signed integer (i.e., seed number) for each type of test. The * symbol indicates when a seed number was also negated (i.e., a negative sign prepended to the integer) to avoid overlap with other seeds. For all possible combined rotational and toroidal shift modified Syrjala tests (**) the assigned integer is simply a sum of the corresponding individual assigned integers for the rotational and toroidal shift test types, respectively. However, the combined rotational and toroidal shift modified Syrjala test assigned integers are only negated once, similar to the rotational or toroidal shift tests. For example, a test which uses five rotations and 0.3 proportion of toroidal shifts will have an assigned integer of 23 along with a negative sign prepended to the seed number.

Test Type	Assigned Integer
Syrjala	08
Rotational (4 rotations)	10*
Rotational (5 rotations)	20*
Rotational (6 rotations)	30*
Rotational (8 rotations)	40*
Rotational (10 rotations)	50*
Rotational (36 rotations)	60*
Rotational (45 rotations)	70*
Toroidal Shifts (0.1 proportion of shifts)	01*
Toroidal Shifts (0.2 proportion of shifts)	02*
Toroidal Shifts (0.3 proportion of shifts)	03*
Toroidal Shifts (0.5 proportion of shifts)	04*
Toroidal Shifts (0.75 proportion of shifts)	05*
Toroidal Shifts (0.9 proportion of shifts)	06*
Combined Rotational and Toroidal Shifts	**
RotToro25Thrshld	60
Energy	09
FR-KS	90
Kmmd	90*

6.2 The Effects of Data Binning on the Syrjala Test

Due to the necessity of identical sampling locations for the Syrjala test (see Section 2.2.1 for more details), binning of data has been used in the literature (Chetverikov et al., 2018; McAdam et al., 2012). Hence, two types of data binning techniques along with a spectrum of binning granularity are discussed in further detail in the next subsection. These are incorporated into the simulation study of the Syrjala test in order to further study their respective effects on test results. This simulation is an extension of the research presented by McKinney and Symanzik (2019, 2021).

6.2.1 Data Binning for Common Sampling Locations

Before applying the Syrjala test, two different types of binning were applied to the data, namely regular and random binning. Regular binning consists of dividing the sample region into a grid of equally sized rectangles. The density of all sample points within each rectangle was reported at the center of the respective rectangles. Random binning consists of randomly assigning binning points (using a simple sequential inhibition process) across the sample region, and assigning each sample point to the closest random binning point (using Euclidean distance). Within each of these binning approaches, three levels of granularity were used. Regular binning consisted of dividing the unit square into 5×5, 10×10, and 20×20 rectangular grids. Random binning involved randomly assigning 25, 100, and 400 random binning points across the sample region.

Figure 16 outlines the process in which random number seeds were used in generating the simulation data used in Sections 6.2–6.4. The different simulation scenarios consist of comparisons among the generated data structures outlined in Section 6.1.1. The different binning scenarios referred to in the figure (mainly random or regular

binning) are outlined above.

Specifically, the process for generating the raw data begins with selecting a two-sample simulation scenario and corresponding random number seed, e.g., a sample size is selected for the first sample (from Table 3) which will be CSR, along with the second sample's distribution (from Table 4) and size (from Table 5). These scenarios are replicated ten times each, and the process is repeated for every simulation scenario (i.e., for every combination of possible first sample size, second sample distribution, and approximate second sample size). This implies there are a total of 80 different simulation scenarios (four different values for the first sample size times five different distributions for the second sample times four different values for the second sample size). Hence, since each simulation scenario is replicated ten times, a total of 800 two-sample raw data comparisons are saved to disk.

Next, Figure 16 shows a fork in the workflow which represents loading the raw data in preparation for two subsequent processes. The process on the right side of the fork is for the simulations for modified Syrjala tests (which employ the processes outlined in Figures 12, 14, and 15) and the alternative tests (discussed in Sections 2.2.2, 2.2.3, and 2.2.4). The process on the left side of the fork is for the data binning process necessary for Syrjala test simulations. Each of the 800 two-sample raw data comparisons are binned twice (using one of the two methods discussed earlier in this section) for a total of 1600 two-sample binned data comparisons, which are also saved to disk before being loaded for the Syrjala test simulations.

6.2.2 Simulation Results

Figure 17 displays a grid of line graphs which summarize the results of a simulation comparing the effect of regular or random data binning (detailed in Section 6.2.1) on the Syrjala test. The horizontal axis displays which type of binning, either regular

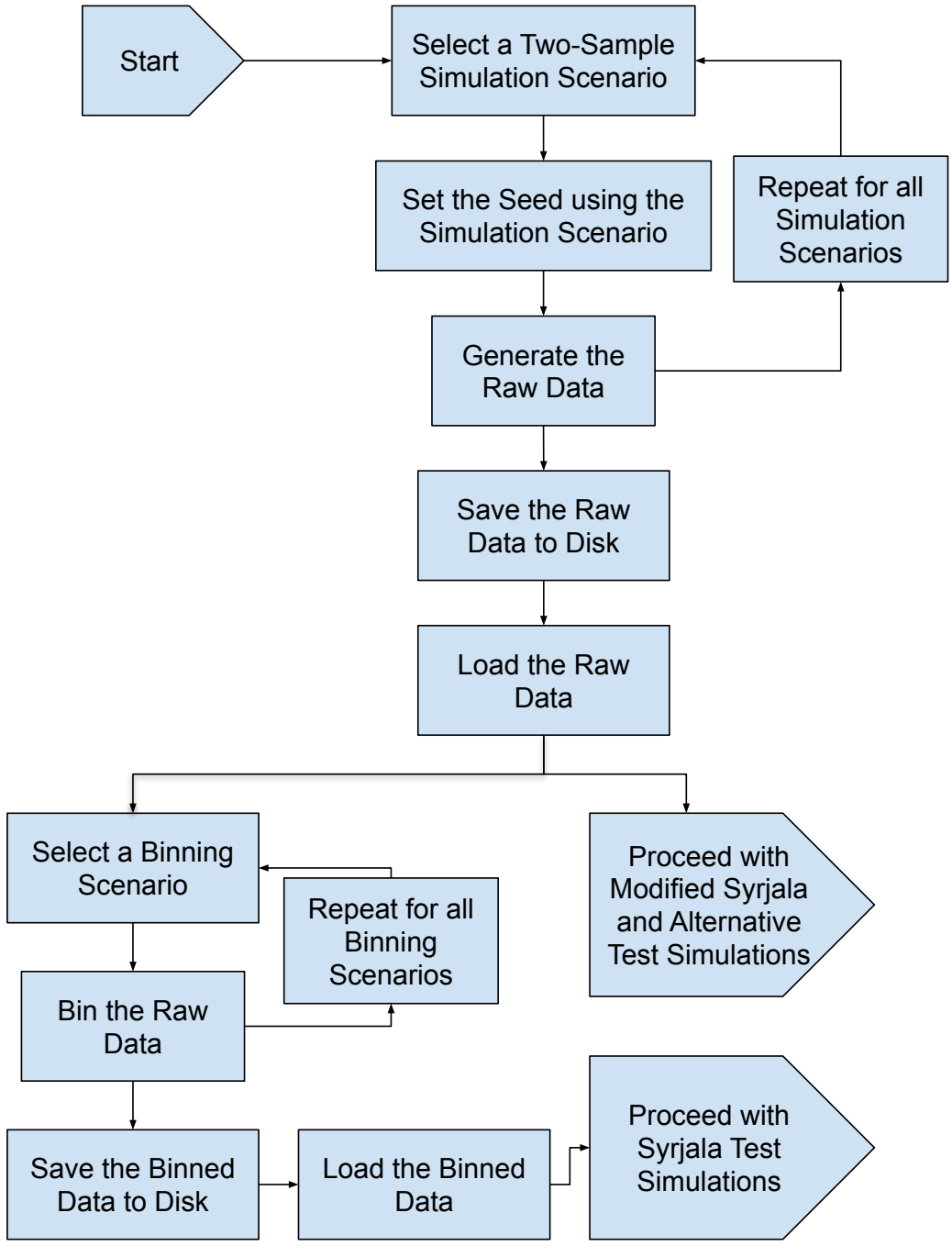


Fig. 16: A flowchart which displays the process in which random number seeds were used in generating the simulation data used in Sections 6.2–6.4. The different simulation scenarios consist of comparisons among the generated data structures outlined in Section 6.1.1. Similarly, the different binning scenarios, mainly random or regular binning, are detailed in Section 6.2.1.

(Reg) or random (Ran), was applied to the simulation data. The granularity of the binning is represented after the Reg (5×5, 10×10, or 20×20) or Ran (25, 100, or 400) horizontal axis tick labels, and are detailed in Section 6.2.1. The grid column name indicates the CSR sample size (n_1), and the grid row indicates the shape of the second sample. The colors of the lines and symbols indicate the second sample size (approximate n_2).

Looking at all of the comparisons of CSR vs CSR (all of the line graphs in the first row of Figure 17), we see that the Syrjala test rejected 67 out of 960 (ten iterations times six binning techniques times four n_1 sample size comparisons times four n_2 sample size comparisons) tests. In other words, the Syrjala test is rejecting around 7% of the tests. Since we are testing at the 5% significance level, we should expect to see roughly 5% of tests reject the null hypothesis when it is actually true. Hence, the Syrjala test is exhibiting anti-conservative behavior.

This result seems to contradict the conservative behavior of the Syrjala test exhibited in McKinney and Symanzik (2019). However, for the case when the null hypothesis is true, a much larger number of small sample comparisons are being made here (960 total tests) as compared to McKinney and Symanzik (2019) (240 total tests). The greater number of false positives here suggests that the performance of the test may behave differently than what has been previously observed in the literature (Fuller et al., 2006) when comparing two samples with relatively small samples sizes.

The remaining rows of graphs show comparisons between realizations of a CSR process with departures from CSR, i.e., when the null hypothesis is false. In the second row, realizations of CSR (with sample sizes of 50, 100, 250, and 500 points) were compared to realizations of a heterogeneous Poisson process called Center (with 50, 100, 250, and 500 sample points, respectively). Overall, the Syrjala test produced

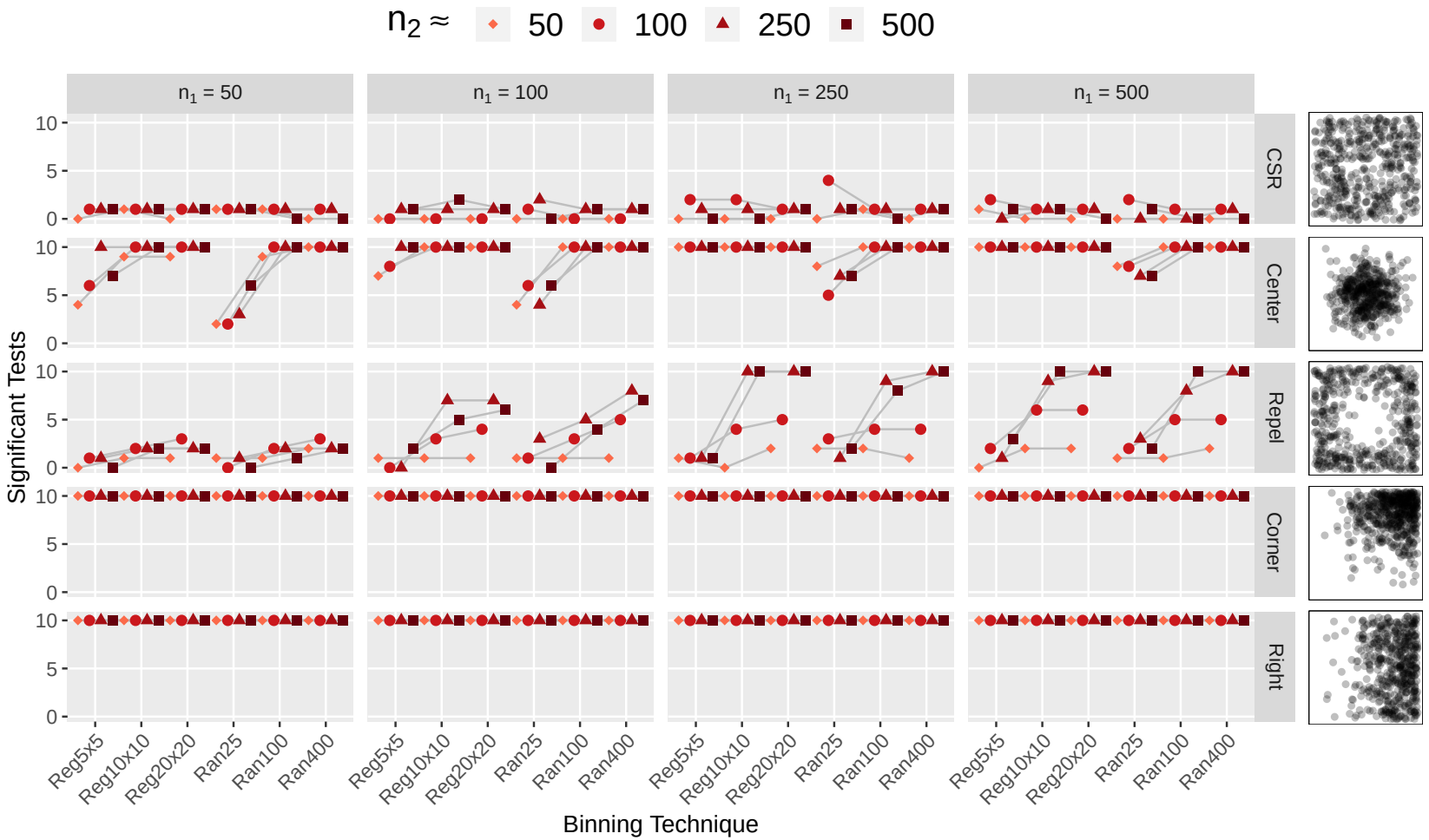


Fig. 17: A grid of line graphs showing the results of a simulation comparing the effect of two types of data binning (lower horizontal axis), abbreviated as Reg or Ran, on the Syrjala test. The grid column indicates the CSR sample size (n_1), and the grid row indicates the distribution of the second sample. The point shapes and colors indicate the second sample size (approximate n_2). For example, the bottom left graph shows the number of significant (p-values < 0.05) Syrjala tests (out of ten tests) on CSR realizations of 50 points with Right realizations of approximately 50, 100, 250, and 500 points. The granularity of the binning is represented after the Reg (5×5, 10×10, or 20×20) or Ran (25, 100, or 400) horizontal axis tick labels, and are detailed in Section 6.2.1

multiple non-significant test results depending on the binning technique and sample size.

Particularly, CSR vs Center (second row of graphs) shows that in all cases of random binning using only 25 points the Syrjala test fails to detect all of the differences. This is also exhibited in regular binning with only a 5×5 grid. However, the effect is overcome for regular binning as soon as both n_1 and n_2 are greater than 100. Nonetheless, this comparison (CSR vs Center) suggests a dependence of the Syrjala test on the data aggregation step, i.e., the binning must be granular enough to reflect the deviations from CSR.

This is further suggested in the third row of graphs where realizations of CSR were compared with departures from CSR called Repel. These comparisons provide an interesting case since the Syrjala test struggled to indicate every significant difference across the different sample size comparisons. Again, a dependence on the granularity of the binning is seen by the non-decreasing nature for all line graphs when both samples are greater than 50.

Furthermore, the general jump in significant test results becomes more stark as both sample sizes increase. For example, when $n_1 = 500$ and $n_2 = 50$ the Syrjala test is only able to detect zero, two, and two significant test results for the regular binning granularity levels of 5×5, 10×10, and 20×20, respectively, and only one, one, and two significant test results for the random binning granularity levels of 25, 100, and 400, respectively. However, when $n_1 = 500$ and $n_2 = 500$ the number of significant Syrjala tests jumps from two to ten and from three to ten when the binning granularity increases from the smallest to the largest levels of regular (i.e., from 5×5 to 20×20) and random (i.e., from 25 to 400) binning, respectively.

Additionally, the Syrjala test is better able to detect differences as the sample sizes increase. This can be seen in two ways. In general, there is a positive trend in

the number of significant test results across all values of n_1 for a fixed value of n_2 and binning technique. Alternatively, there is a positive trend in the number of significant test results across all values of n_2 for a fixed value of n_1 and binning technique. Also notable are the cases in which either sample size is less than 100. Here, the Syrjala test is only able to detect at most three out of ten cases across all of the other sample's sizes and binning techniques.

Overall, not only does row three in Figure 17 reinforce the observed dependence of the Syrjala test on the binning technique (also observed by McKinney and Symanzik (2019)), but it also confirms that the Syrjala test places less emphasis on differences located near the center of the bounding region which was observed by McAdam et al. (2012).

In the remaining two rows where realizations of CSR are compared with the Corner and Right distributions, the Syrjala test was able to detect all significant differences. This confirms the results established by McKinney and Symanzik (2019).

6.3 Modified Syrjala Tests Simulation Study

Each of the three modifications proposed to the Syrjala test, namely the rotational, toroidal shift, and combination of rotational and toroidal shift modifications (detailed in Chapter 5), has a parameter (or two in the case of the combined modifications) which is explored within this simulation study. When simulating the modification which extends the number of rotations of the data, 4, 5, 6, 8, 10, 36, and 45 rotations (within one 360° rotation) are employed in this simulation as compared to the 4, 6, 8, 10, and 36 rotations used in (McKinney and Symanzik, 2019). The additional rotations of 5 and 45 were included to further ensure stability among the results. Since the modification which involves only toroidal shifts of the data introduces considerable additional computational load, only subsets of the combined

two-sample data are randomly selected as origins to the toroidal shifts (instead of constructing a toroidal shift for each data point). The proportions of randomly selected data values explored in the simulation are 0.1, 0.2, 0.3, 0.5, 0.75, and 0.90. Naturally, the modification which includes both rotational and toroidal shifts is even more computationally intensive. While all of the previously employed rotations are still included, only the proportions 0.1, 0.2, and 0.3 (of randomly selected data values used as origins of the toroidal shifts) are explored. However, stable results are exhibited and discussed for these proportions in Section 6.3.3.

Additionally, Section 6.3.4 explores the performance of the combined modified Syrjala tests when the number of randomly chosen points as origins for the toroidal shifts are limited to some threshold which is typically much smaller than the pooled sample size. This restriction eases some of the computational load of the test which employs both rotational and toroidal shift modifications. It also justifies a default threshold value for the tests in the `R` package (see Chapter 8), which guides new users of the package toward relatively reasonable parameter values.

6.3.1 Rotational Modification Simulation Results

Figure 18 shows the results of the simulation study for the rotational modification when using the CWS statistic. The remaining five proposed versions of the statistic ξ (see Section 5.1 for more details) are also explored. However, since Figures 18 and 62–66 show almost the same behavior aside from some chance variation, the latter figures (Figures 62–66) for the DWS, UWS, DWA, UWA, and CWA simulations (respectively) are provided in Appendix B. Recall from Section 5.1 that the abbreviations DW, UW, and CW refer to the different types of weightings in the ξ statistics, i.e., double weighted, uniformly weighted, and complementary weighted, respectively. The S and A are abbreviations for squared or absolute differences, re-

spectively, computed between the ECDFs from the two samples. Figures 18 and 62–66 each display a grid of line graphs which depict the number of significant test results (p-values < 0.05) out of ten tests for each combination of rotational parameter level (horizontal axes), second sample size (n_2, represented by the different colors and shapes in the line graphs), distribution of the second sample (grid rows), and first sample size (n_1, shown in the grid columns).

In summary, while the squared statistics perform marginally better than the absolute value statistics (this is shown more explicitly in Section 6.4.2), the choice in proposed versions of the statistic ξ makes little difference in the overall performance of the rotational modification test. Additionally, Figure 18 shows little indication that an increase in the number of rotations within the test has any effect on the number of significant test results regardless of the version of the statistic ξ. This suggests that a lower number of rotations will achieve similar results while providing computational efficiency. This effect is investigated further in Section 6.3.3 when simulating the behavior of the test which involves a combination of the rotational and toroidal shift modifications.

For the cases when the null hypothesis is true (both samples were generated from the same CSR distribution), we can see that all of the rotational tests demonstrated roughly the same false positive rate (as seen in the first row of graphs across all of Figures 18 and 62–66). Overall, 67 out of 1120 tests resulted in false positives, the ratio of which gives a false positive rate of approximately 0.0598. This is also shown and discussed further in Section 6.4.2.

The cases when the null hypothesis is false are seen in the bottom four rows of graphs. Here, none of the rotational tests in Figure 18 exhibit any difficulty in correctly identifying all of the significant differences for both Corner and Right departures from CSR (as seen in the bottom two rows of graphs). When compared

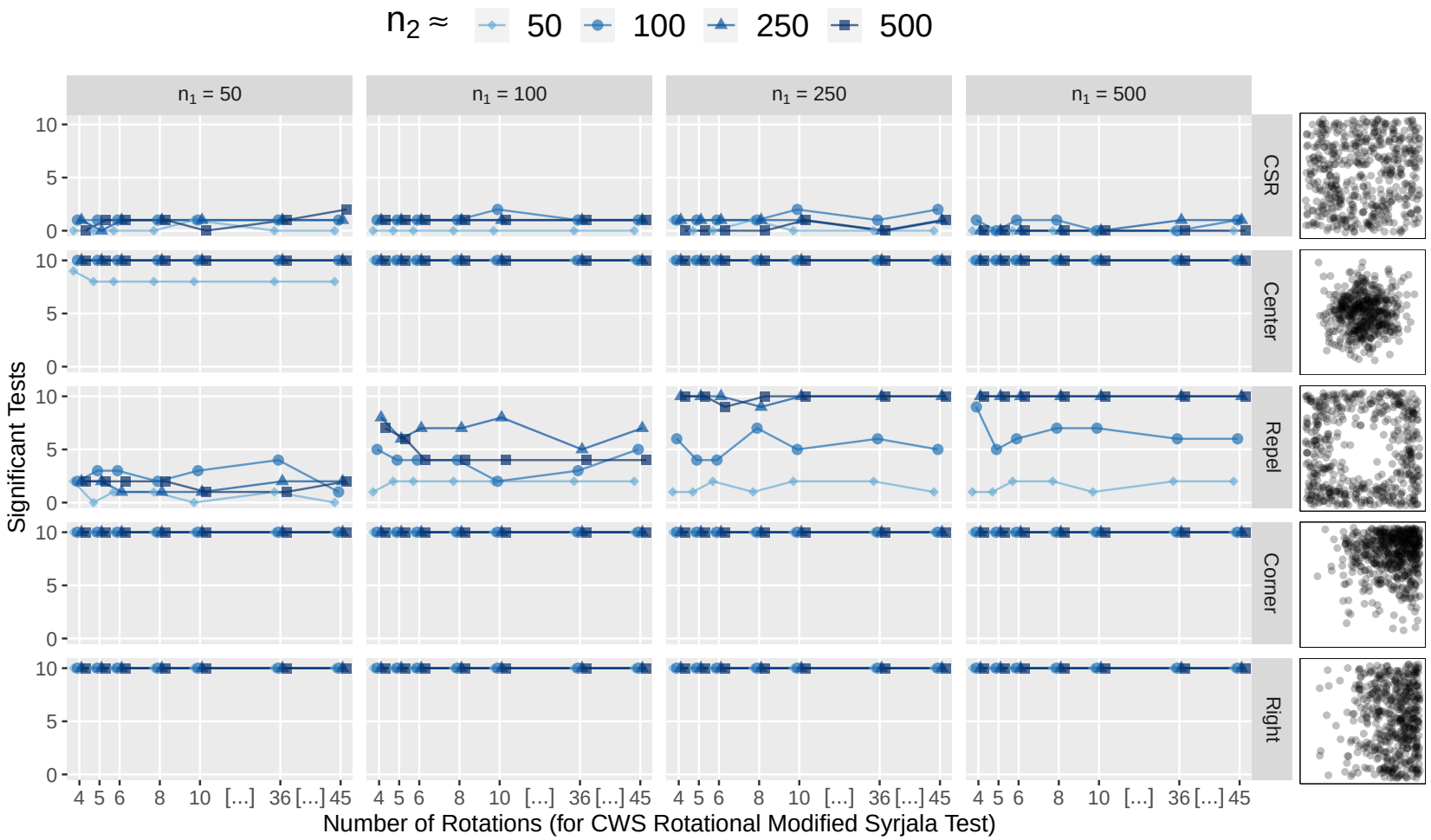


Fig. 18: A grid of line graphs showing the results of a simulation comparing multiple rotations of the modified Syrjala test using complementary weightings of the squared differences in the ECDFs (CWS). The grid column name indicates the CSR sample size (n_1), and the grid row indicates the shape of the second sample. The colors of the lines and symbols indicate the second sample size (approximate n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) on CSR realizations of 50 points with Right realizations of approximately 50, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

with Figure 17 it is clear that there are cases in which the Syrjala test and the rotational modified Syrjala test still agree. Furthermore, the rotational tests do well in correctly identifying significance for the Center distribution (as seen in the second row of graphs) except for the case when the second sample is small ($n_2 \approx 50$). In general, about two out of ten tests were labeled as non-significant in this case.

However, similar to the results shown in Section 6.2.2, the Repel case (as seen in the third row of graphs) proves to be more difficult for the test to correctly identify significant differences. This confirms that the rotational modification Syrjala tests also place less emphases on differences located near the center of the bounding region similar to the Syrjala test (McAdam et al., 2012). However, the rotational modification Syrjala tests overcome some of these issues (see Section 6.2.2). In general, as both of the sample sizes increase so do the number of significant results. At the point when both sample sizes are greater than 250, the test can identify almost all of the significant differences (see the $n_2 \approx 250$ and $n_2 \approx 500$ line graphs for the $n_1 = 250$ and $n_1 = 500$ columns in the third row of graphs).

Additionally, more detailed comparisons of the power and false positive rate of each of these tests as compared to the other tests employed in Section 6.3.2 and 6.3.3 are discussed in Section 6.4.2. In comparison, the far right column ($n_1 = 500$) of Figure 62 match the results found by McKinney and Symanzik (2019).

6.3.2 Toroidal Shift Modification Simulation Results

Similar to the previous section, Figure 19 shows the results of the simulation study for the rotational modification when using the CWS statistic. The remaining five proposed versions of the statistic ξ (see Section 5.1 for more details) are also explored. However, since Figures 19 and 67–71 show almost the same behavior aside from some chance variation, the latter figures (Figures 67–71) for the DWS, UWS,

DWA, UWA, and CWA simulations (respectively) are provided in Appendix B. The layout of these figures is identical to Figure 18 except that the horizontal axes display the proportions of points used as origins for the toroidal shifts ranging from 0.1, 0.2, 0.3, 0.5, 0.75, and 0.9.

Similar to Section 6.3.1, while the squared statistics perform marginally better than the absolute value statistics in Figure 19 (this is shown more explicitly in Section 6.4.2), the choice in proposed versions of the statistic ξ makes little difference in the overall performance of the toroidal shift modification test. Additionally, Figure 19 shows little indication that an increase in the proportion of randomly selected points (used for origins of the toroidal shifts) within the test has any effect on the number of significant test results regardless of the version of the statistic ξ. This suggests that a lower number of toroidal shifts will achieve similar results while providing computationally efficiency. This effect is investigated further in Section 6.3.3 when simulating the test behavior of the test which involves a combination of the rotational and toroidal shift modifications.

For the cases when the null hypothesis is true (both samples were generated from the same CSR distribution), we can see that all of the toroidal shift tests demonstrated roughly the same false positive rate (as seen in the first row of graphs). Overall, 30 out of 960 tests resulted in false positives, the ratio of which gives a false positive rate of approximately 0.0313. This is shown more explicitly and discussed further in Section 6.4.2.

The cases when the null hypothesis is false are seen in the bottom four rows of graphs. Here, none of the toroidal shift tests in Figure 19 exhibits any difficulty in correctly identifying all of the significant differences for the Center or Corner departures from CSR (as seen in the second and fourth rows of graphs, respectively). Similarly, the toroidal shift tests do well in correctly identifying significance for the

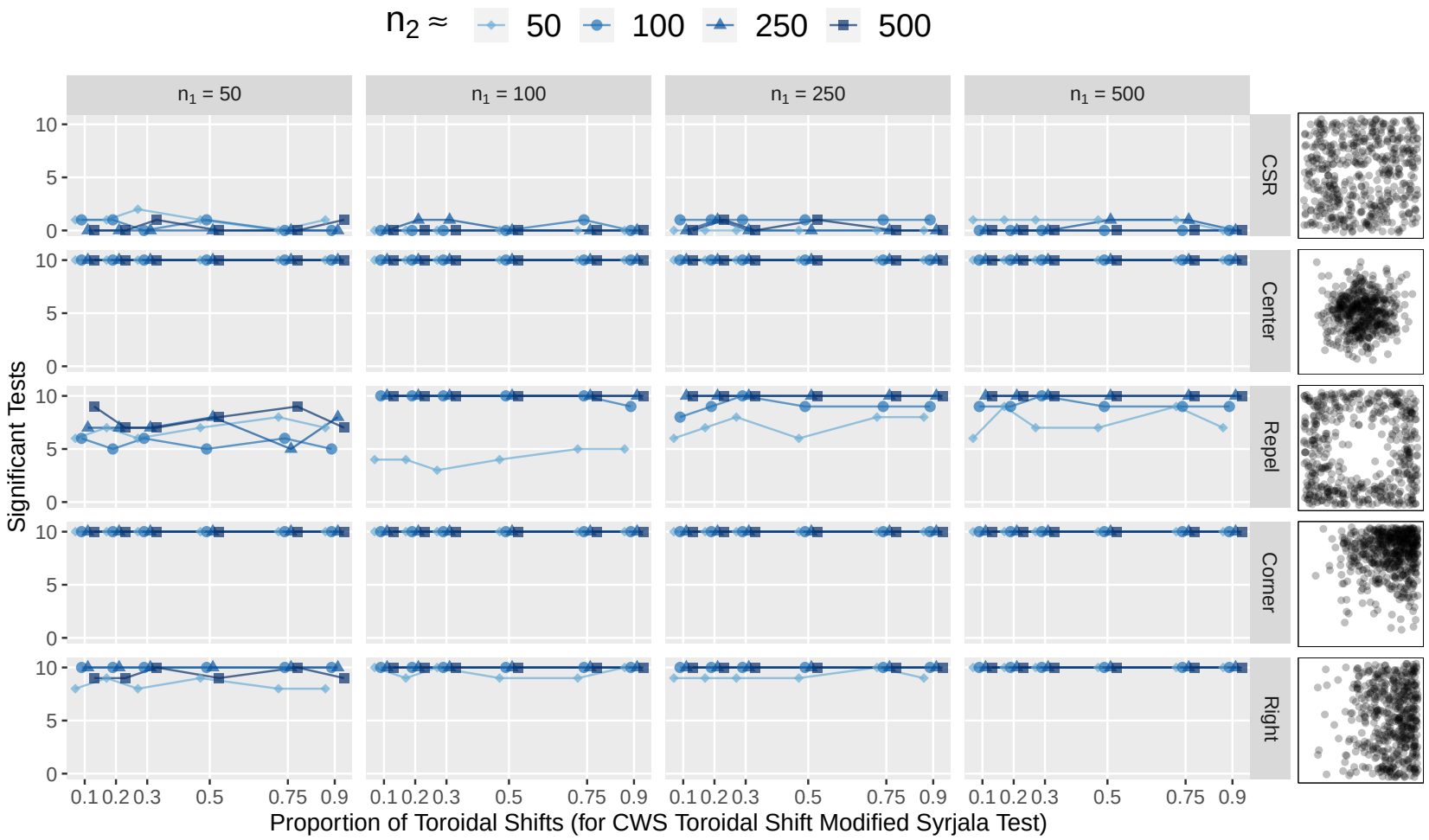


Fig. 19: A grid of line graphs showing the results of a simulation comparing multiple randomly selected proportions of points for the origins of the toroidal shifts of the modified Syrjala test using complementary weightings of the squared differences in the ECDFs (CWS). The grid column name indicates the CSR sample size (n_1), and the grid row indicates the shape of the second sample. The colors of the lines and symbols indicate the second sample size (approximate n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) on CSR realizations of 50 points with Right realizations of approximately 50, 100, 250, and 500 points.

Right distribution (as seen in the bottom row of graphs) except for a few cases most of which occur when the second sample is small. In comparison, the rotational modified Syrjala tests correctly computed significance for all of the Right distribution cases (as seen in the bottom row of graphs of Figure 18). As a reminder, common random numbers are being employed across all of the simulations (see Section 6.1.3). Thus, the same ten replications of CSR vs. Right pairs are being compared between the previous simulation (Figure 18) and this simulation. Therefore, it is not reasonable to say that these different results for CSR vs. Right are due to chance variation from a small sample size n_2. The slight decrease in performance can be attributed to the differences in test modifications (i.e., toroidal shifts vs. rotations).

Similar to the rotational test results (in Section 6.3.1), the Repel case (as seen in the third row of graphs) proves to be more difficult for the test to correctly identify significant differences. However, there is a noticeable improvement of the toroidal shift modification over the rotational modification. In general, as both of the sample sizes increase so do the number of significant results. At the point when both sample sizes are greater than 100, the test can identify almost all of the significant differences (see the $n_2 \approx 100$, $n_2 \approx 250$ and $n_2 \approx 500$ line graphs for the $n_1 = 100$, $n_1 = 250$ and $n_1 = 500$ columns in the third row of graphs).

When compared with Figure 18, Figure 19 shows that while the toroidal shift test increased the overall number of correct significant test results, the rotational test does produce more significant test results in a few cases. Specifically, if we compare the two tests one row at a time, it is clear that the toroidal shift is a better selection for the Center distribution as it handles the case when both sample sizes are equal to 50 better than the rotational test. Similarly, the toroidal shift test performs better in almost all cases of the Repel distribution. Both tests perform perfectly for the Corner distribution in identifying every test result as significant. However, the rotational

test outperforms the toroidal shift test for several small sample comparisons from the Right distribution. This is shown more explicitly in Section 6.4.2, in addition to more detailed comparisons of the power and false positive rate of each of toroidal shift modification tests as compared to the other tests employed in Section 6.3.3.

6.3.3 Simulation Results for the Modified Syrjala Tests which Combine both Rotational and Toroidal Shift Modifications

While little difference is observed in the versions of the ξ statistics across both the rotational (Figures 18 and 62–66) and toroidal shift (Figures 19 and 67–71) modifications, squared differences in the statistics perform marginally better than the absolute differences. Hence, the squared differences are only considered in the simulations of the test involving both rotational and toroidal shift modifications.

While Figures 20–22 show the results of the simulation study for the combination of both rotational and toroidal shift modifications using only the ξ^{CWS} statistic, the remaining five proposed versions of the statistic ξ (see Section 5.1 for more details) are also explored. However, since Figures 20 and 72–77 show almost the same behavior aside from some chance variation, the latter figures (Figures 72–77) for the DWS, and UWS simulations are provided for each of the proportions of toroidal shifts (0.1, 0.2, and 0.3) in Appendix B.

Similar to Figures 18, 19, and 62–71, each one of these figures also displays a grid of line graphs which depict the number of significant test results (p-values < 0.05) out of ten tests for each number of rotations (horizontal axes), second sample size (n_2, represented by the different colors and shapes in the line graphs), distribution of the second sample (grid rows), and first sample size (n_1, shown in the grid columns) for a given proportion of randomly selected points used as origins of the toroidal shifts. However, Figures 20–22 use the proportions of 0.1, 0.2, and 0.3, respectively. The

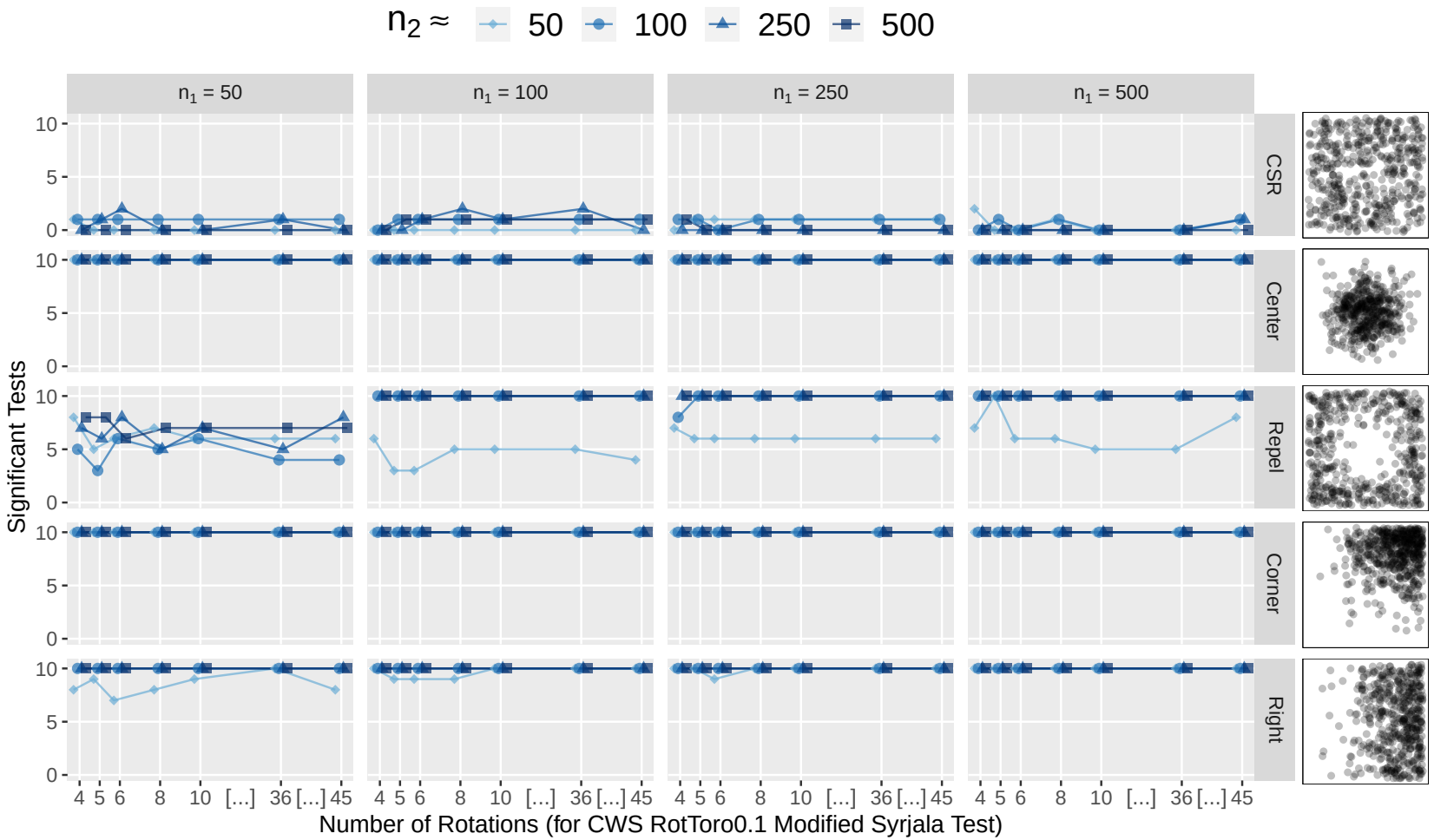


Fig. 20: A grid of line graphs showing the results of a simulation comparing multiple rotations of the modified Syrjala test using 0.1 proportion of points as origins of toroidal shifts, and complementary weightings of the squared differences in the ECDFs (CWS). The grid column name indicates the CSR sample size (n_1), and the grid row indicates the shape of the second sample. The colors of the lines and symbols indicate the second sample size (approximate n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) on CSR realizations of 50 points with Right realizations of approximately 50, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

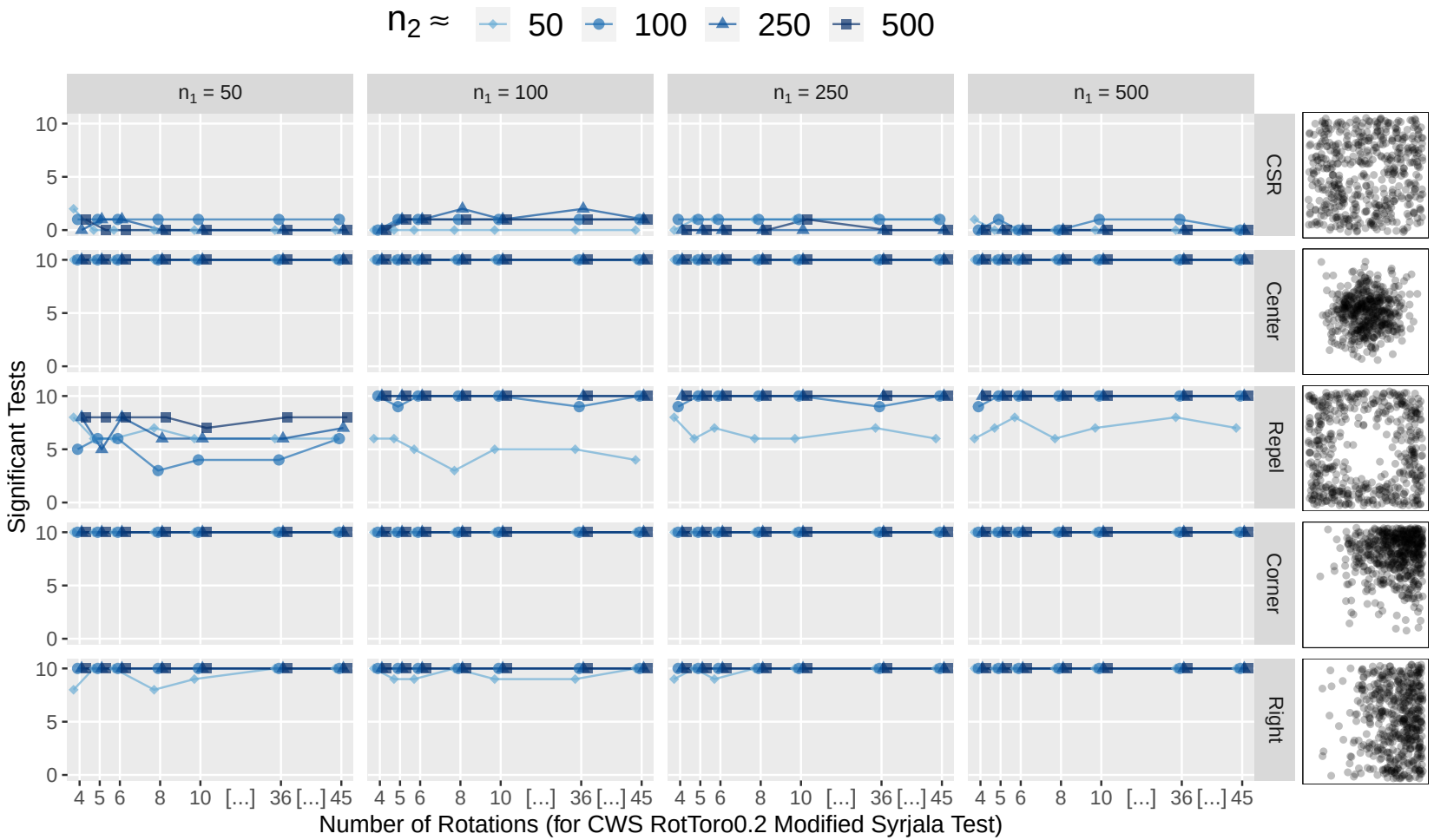


Fig. 21: A grid of line graphs showing the results of a simulation comparing multiple rotations of the modified Syrjala test using 0.2 proportion of points as origins of toroidal shifts, and complementary weightings of the squared differences in the ECDFs (CWS). The grid column name indicates the CSR sample size (n_1), and the grid row indicates the shape of the second sample. The colors of the lines and symbols indicate the second sample size (approximate n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) on CSR realizations of 50 points with Right realizations of approximately 50, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

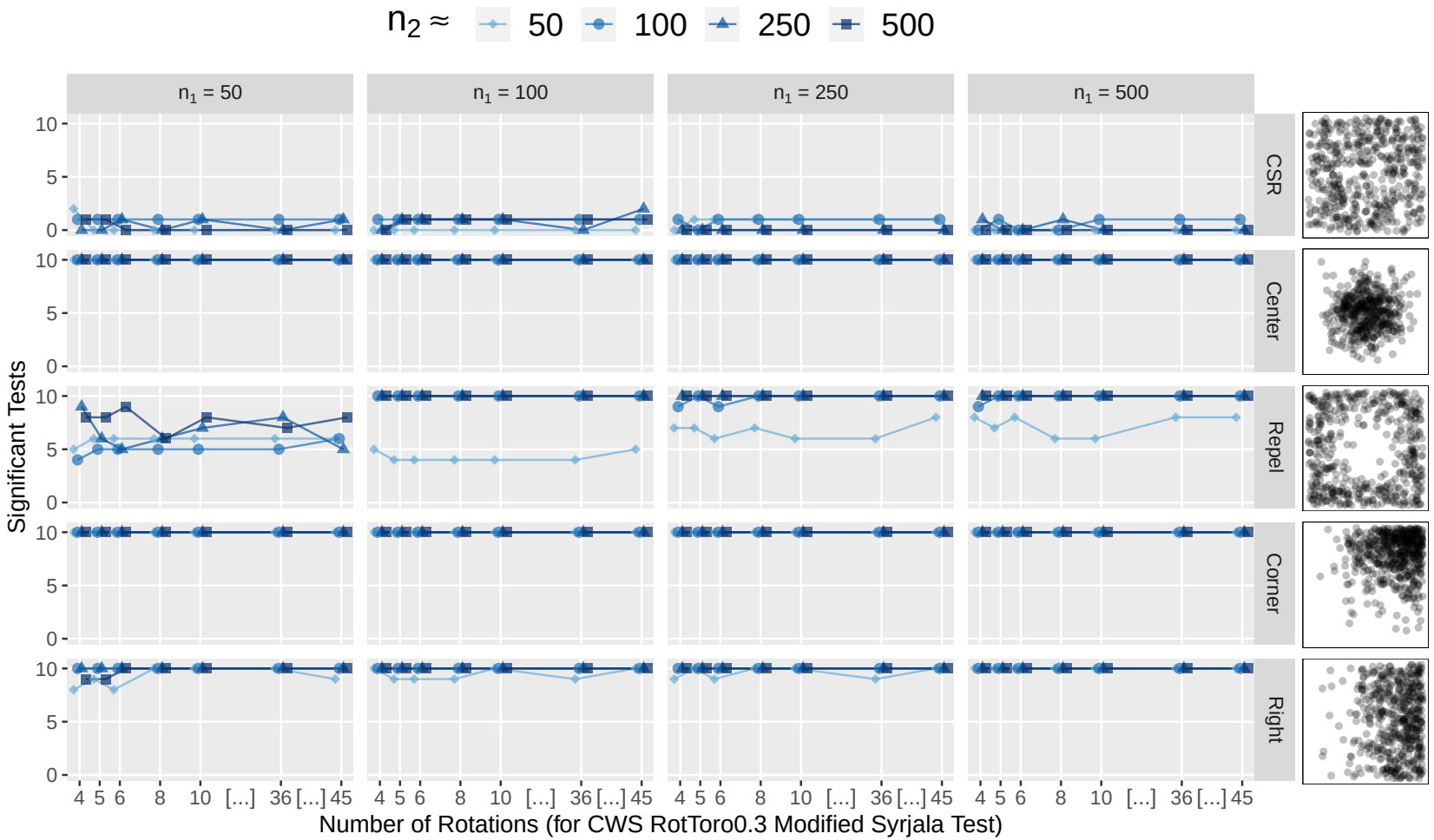


Fig. 22: A grid of line graphs showing the results of a simulation comparing multiple rotations of the modified Syrjala test using 0.3 proportion of points as origins of toroidal shifts, and complementary weightings of the squared differences in the ECDFs (CWS). The grid column name indicates the CSR sample size (n_1), and the grid row indicates the shape of the second sample. The colors of the lines and symbols indicate the second sample size (approximate n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) on CSR realizations of 50 points with Right realizations of approximately 50, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

remaining proportions explored in the test involving only the toroidal shifts (0.5, 0.75, and 0.9) are not included here due to the computational load imposed by the large proportions. However, stable results are shown across Figures 20–22 similar to those seen in Section 6.3.2.

Overall, Figures 20–22 show almost the same behavior aside from some chance variation, and the overall number of significant tests are almost identical to the toroidal shift test as seen in Section 6.3.2. This suggests that a smaller number of rotations as well as a smaller proportion of randomly selected points used as the origins of the toroidal shifts is sufficient for representative test results while also providing relief to the computational load.

For the cases when the null hypothesis is true (both samples were generated from the same CSR distribution), we can see that all of the combined rotational and toroidal shift tests demonstrated roughly the same false positive rate (as seen in the first row of graphs of each figure). However, the false positive rate expressed by this test is closer to the significance level of 0.05 than either the rotational or toroidal shift tests. Specifically, 50, 50, and 51 out of 1120 tests resulted in false positives, the ratio of which gives a false positive rate of approximately 0.045, 0.045, and 0.046 for each of the simulation results in Figures 20–22, respectively. This is compared more explicitly and discussed further in Section 6.4.2.

The cases when the null hypothesis is false are seen in the bottom four row of graphs. Here, none of the combined rotational and toroidal shift tests in Figures 20–22 exhibit any difficulty in correctly identifying all of the significant differences for the Center or Corner departures from CSR (as seen in the second and fourth rows of graphs, respectively). This behavior is identical to the toroidal shift test (see Section 6.3.2). Similarly, the combined rotational and toroidal shift tests do well in correctly identifying significance for the Right distribution (as seen in the bottom row

of graphs) except for a few cases when the second sample is small. This behavior is also similar to the toroidal shift test (see Section 6.3.2), except that the combined modification test is moderately better. However, the combined modification is still not as good as the test which employs only rotations for a few cases from the Right distribution. This may be due to the fact that the rotational test still emphasizes differences closer to the edge of the sample distributions, which proves to be a strength when faced with distributions similar to the Right case.

Similar to the toroidal shift test results (in Section 6.3.2), the Repel case (as seen in the third row of graphs) proves to be more difficult for the combined modification test to correctly identify significant differences. However, similar to the toroidal shift test, there is a noticeable improvement of the combined modification test over the rotational modification. In general, as both of the sample sizes increase so do the number of significant results. At the point when both sample sizes are greater than 100, the test can identify almost all of the significant differences (see the $n_2 \approx 100$, $n_2 \approx 250$ and $n_2 \approx 500$ line graphs for the $n_1 = 100$, $n_1 = 250$ and $n_1 = 500$ columns in the third row of graphs).

6.3.4 Simulation Results for the Combined Rotational and Toroidal Shift Modified Syrjala Tests which Employ Toroidal Shift Thresholds

While computing a number of toroidal shifts using a proportion of the pooled sample size has been studied in Sections 6.3.2 and 6.3.3, in this section additional functionality for computational efficiency has been integrated into the modified Syrjala test which employs both rotational and toroidal shifts. This functionality allows a limiting threshold to be set to the number of randomly selected points used as origins for toroidal shifts. If the pooled sample size is below the threshold, a toroidal shift will be computed for every point in both samples (within every rotation). However,

if the pooled sample size is over the threshold, only a random number of points equal to the threshold are used as origins for toroidal shifts. (Note that this random sample of points is redrawn for every rotation.) As a reminder, the same generated data used for previous modified Syrjala test simulations are being used here. Furthermore, the test results are reproducible using the random number seeds listed in Table 7.

Figure 23 shows the results of a simulation where a toroidal shift threshold was set at 25 points. While the combined sample size is always at least 100, and consequently always greater than a toroidal shift threshold of 25, later simulations (in Sections 6.5) include cases where the combined sample size is less than the toroidal shift threshold.

Furthermore, while Figure 23 only shows the results for the test which employs the CWS statistic, tests which employ the DWS and UWS statistics were also considered. However, the results are similar among these tests except for some chance variation. Hence, the test results for the latter two are included as Figures 78 and 79 in Appendix B. As a reminder, the tests with statistics which computed absolute differences in the ECDFs (i.e., the tests which used the DWA, UWA, and CWA statistics) were not considered here since they showed little difference to the squared statistics (DWS, UWS, and CWS), and the squared statistics achieved a marginally higher power. This is more clearly seen in Section 6.4.2.

Furthermore, the test results of Figure 23 as compared to Figures 20–22 are also similar except for some chance variation. Similar to Figures 20–22, Figure 23 also displays a grid of line graphs which depict the number of significant test results (p-values < 0.05) out of ten tests for each number of rotations (horizontal axes), second sample size (n_2, represented by the different colors and shapes in the line graphs), distribution of the second sample (grid rows), and first sample size (n_1, shown in the grid columns) for a limited number of 25 randomly selected points used as origins of the toroidal shifts.

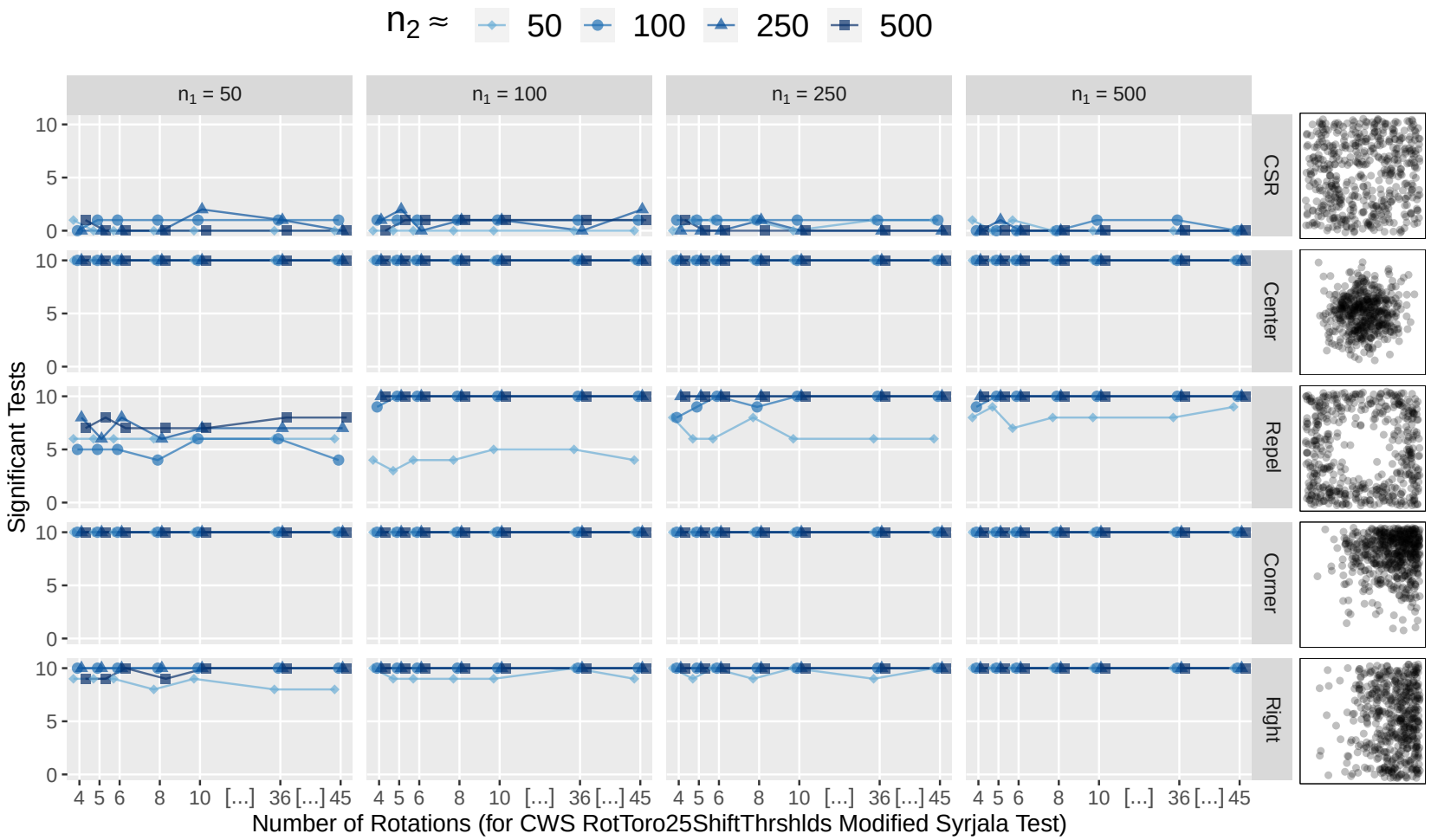


Fig. 23: A grid of line graphs showing the results of a simulation comparing toroidal shift thresholds of 25 points of the modified Syrjala test across a number of rotations using complementary weightings of the squared differences in the ECDFs (CWS). The grid column name indicates the CSR sample size (n_1), and the grid row indicates the shape of the second sample. The colors of the lines and symbols indicate the second sample size (approximate n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) on CSR realizations of 50 points with Right realizations of approximately 50, 100, 250, and 500 points.

For the cases when the null hypothesis is true (both samples were generated from the same CSR distribution), we can see that all of the tests demonstrated roughly the same false positive rate (as seen in the first row of graphs). Overall, 50 out of 1120 tests resulted in false positives, the ratio of which gives a false positive rate of approximately 0.045. This is shown more explicitly and discussed further in Section 6.4.2.

The cases when the null hypothesis is false are seen in the bottom four rows of graphs. Here, none of the tests in Figure 23 exhibits any difficulty in correctly identifying all of the significant differences for the Center or Corner departures from CSR (as seen in the second and fourth rows of graphs, respectively). Similarly, the toroidal shift tests do well in correctly identifying significance for the Right distribution (as seen in the bottom row of graphs) except for a few cases most of which occur when the second sample is small.

Similar to the combined rotational and toroidal shift test results (in Section 6.3.3), the Repel case (as seen in the third row of graphs) proves to be more difficult for the combined modification test with a toroidal shift threshold to correctly identify significant differences. However, similar to the combined rotational and toroidal shift tests, there is a noticeable improvement of the combined modification test over the rotational modification. In general, as both of the sample sizes increase so do the number of significant results. At the point when both sample sizes are greater than 100, the test can identify almost all of the significant differences (see the $n_2 \approx 100$, $n_2 \approx 250$ and $n_2 \approx 500$ line graphs for the $n_1 = 100$, $n_1 = 250$ and $n_1 = 500$ columns in the third row of graphs).

Hence, limiting the number of toroidal shifts to 25 per rotation does not change the results of the tests under these scenarios considerably. These results provide motivation to a computationally conscientious default threshold value for the tests in the `R` package (see Chapter 8). A more detailed comparison of the power and false

positive rates of these tests is made in Section 6.4.2.

6.4 Comparative Simulation Study

In this section, a simulation study is conducted which makes a comparison of the performance (both power and false positive rate) of five methods: (1) the Energy test (Rizzo and Székely, 2016), (2) the kernel maximum mean discrepancy test (Gretton et al., 2012), (3) the extension of the Kolmogorov (1933) test within the Friedman and Rafsky (1979) test (see Sections 2.2.1–2.2.4 for more details), and the combined rotational and toroidal shift modified Syrjala tests which use (4) proportions of points for toroidal shifts and (5) thresholds for the number of toroidal shifts. The simulations discussed in Section 6.3 involved exploring the rotational and toroidal shift modifications across multiple levels of number of rotations, proportion of randomly chosen points as origins of the toroidal shifts, and a threshold of 25 toroidal shifts, respectively. However, in this section the two versions of the test that are considered use 8 rotations and either 0.1 proportions of toroidal shifts or a threshold of 25 toroidal shifts. Additionally, these tests employ only the CWS test statistic. While arguments could be made for the other test statistics and modification parameters, it has been shown that the results will not change drastically from one choice to another (see Section 6.3).

6.4.1 Simulation Results

Figure 24 shows a grid of line graphs which depict the number of significant tests (p-values < 0.05) out of ten tests for each of the methods under consideration. The five methods, abbreviated on the horizontal axes, are (1) the Energy test, (2) the kernel maximum mean discrepancy test (Kmmd), (3) Friedman and Rafsky's extension of the Kolmogorov test (FR-KS), (4) the combined rotational and toroidal

shift modified Syrjala test using the CWS statistic, 8 rotations, and 0.1 proportion of points as origins of the toroidal shifts (Rot8Toro0.1), and (5) the combined rotational and toroidal shift modified Syrjala test using the CWS statistic, 8 rotations, and a toroidal shift threshold of 25 points (Rot8Toro25). Figure 24 also shows the second sample size (n_2, represented by the different colors and shapes in the line graphs), the distribution shape of the second sample (grid rows), and the first sample size (n_1, shown in the grid columns). Note that the Rot8Toro0.1 and Rot8Toro25 test results are identical to those previously presented in Figures 20 and 23, respectively, for the cases in which the number of rotations equals eight.

When the null hypothesis is true, and the two samples are being generated by the same CSR process (as seen in the top row of graphs in Figure 17), the methods should reject about one in twenty times since the significance level is 0.05. None of the methods appears to be obviously conservative (like the Syrjala test in Figure 17) or anti-conservative. However, a more detailed analysis of the false positive rate of each test is discussed in Section 6.4.2.

When the null hypothesis is false, and the second sample exhibits some departure from CSR, it is interesting to note that all of the tests were able to correctly identify all significant cases for the Corner distribution (see the second row from the bottom in Figure 24). This case shows that there is an amount of agreement between the bivariate two-sample tests.

Some additional amount of agreement between the tests is also seen in the Right and Center distributions. However, the FR-KS test failed to identify all of the cases, particularly when the sample sizes were smaller. As both sample sizes increased, the FR-KS test was better able to identify significant differences in the Right distribution (bottom row) than in the Center distribution (second row from the top).

In contrast, the Energy, Kmmd, Rot8Toro0.1 and Rot8Toro25 tests were able

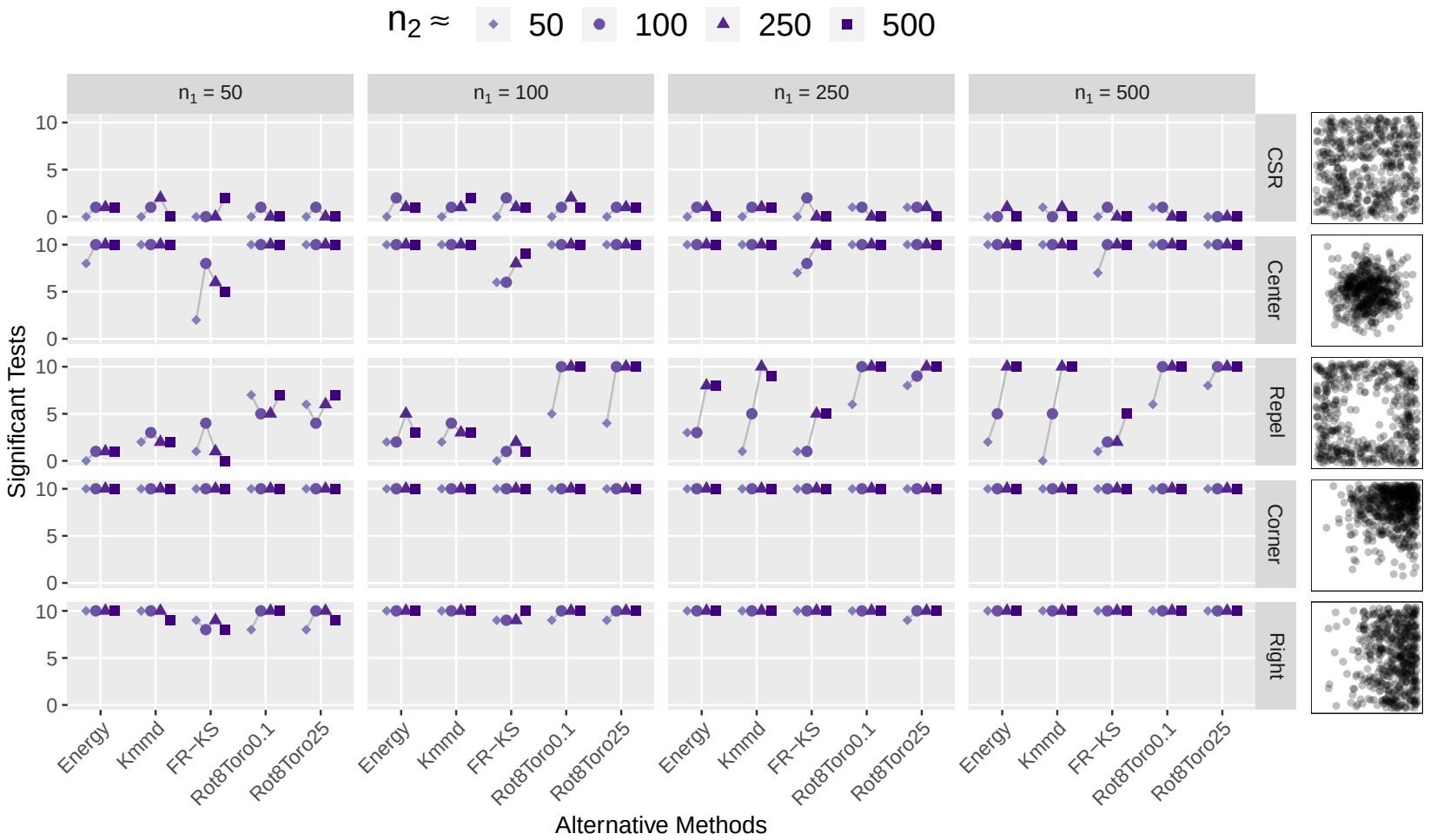


Fig. 24: A grid of line graphs showing the results of a simulation comparing alternative multivariate two-sample tests to the modified Syrjala test using the CWS statistic, eight rotations, and either 0.1 proportion of points as origins of toroidal shifts (Rot8Toro0.1), or a toroidal shift threshold of 25 points (Rot8Toro25). The grid column name indicates the CSR sample size (n_1), and the grid row indicates the shape of the second sample. The colors of the lines and symbols indicate the second sample size (approximate n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) on CSR realizations of 50 points with Right realizations of 50, 100, 250, and 500 points. Note that the Rot8Toro0.1 and Rot8Toro25 test results are identical to those previously presented in Figures 20 and 23, respectively.

to identify all of the differences in the Right and Center distributions except for a few cases. The Energy test misclassified two cases for when $n_1 = 50$ and $n_2 \approx 50$ in the Center distribution. The Kmmd test misclassified one case for when $n_1 = 50$ and $n_2 \approx 500$ in the Right distribution. The Rot8Toro0.1 modified Syrjala test misclassified two cases for when $n_1 = 50$ and $n_2 \approx 50$, and one case for when $n_1 = 100$ and $n_2 \approx 50$ in the Right distribution. In addition to the cases that the Rot8Mod0.1 test misclassified in the Right distribution, the Rot8Toro25 modified Syrjala test misclassified two additional cases when $n_1 = 50$ and $n_2 \approx 500$, and when $n_1 = 50$ and $n_2 \approx 250$. Otherwise, the two versions of the modified Syrjala test perform similarly.

The Repel departure from CSR provides an interesting case where all of the tests exhibited difficulty in correctly identifying all of the significant cases. While not always the case, in general, all of the tests performed better as either sample size increased. For the case when $n_1 = 50$ (first column in the third row), the Rot8Toro0.1 and Rot8Toro25 modified Syrjala tests are able to correctly identify more significant differences than all three of the other tests except for when $n_2 \approx 100$ where Rot8Toro0.1 performs better than the FR-KS test by only one more significant result and Rot8Toro25 performs the same as the FR-KS test. In all other cases, the Rot8Toro0.1 and Rot8Toro25 tests outperform the others by at least three significant tests or more.

A similar superiority is exhibited by Rot8Toro0.1 and Rot8Toro25 over the other three tests when $n_1 = 100$. However, as soon as $n_1 = 250$ or $n_1 = 500$ (third and fourth columns from the left of the third row) and $n_2 \approx 250$ or $n_2 \approx 500$, the Energy and Kmmd tests begin to perform comparatively to the Rot8Toro0.1 and Rot8Toro25 tests. While the FR-KS test is beginning to improve when $n_1 = 250$ and $n_1 = 500$, it under-performs all other tests, achieving significance in at most only five out of the ten tests in a few cases. Most notably, the Rot8Toro0.1 and Rot8Toro25 tests again

outperform all other tests when $n_1 = 250$ or $n_1 = 500$ and $n_2 \approx 50$ or $n_2 \approx 100$ by at least three significant test results or more.

6.4.2 Comparison of Power and False Positive Rates

This section summarizes the results of the simulations discussed in Sections 6.2–6.4. The power (see Figure 25 and Table 8) and false positive rate (see Figure 26 and Table 9) are computed and graphed for the rotational, toroidal shift, and combined modified Syrjala tests, as well as the Syrjala (across the different binning types and granularities), Energy, Kmmd, and FR-KS tests. For the combined rotational and toroidal shift modified Syrjala tests, both tests using proportions of points and thresholds are considered when limiting the number of toroidal shifts. Within each of the modified Syrjala tests shown in Figures 25 and 26, six different statistics were computed using double weighted (DW), uniformly weighted (UW), or complementary weighted (CW) differences in the sample ECDFs, which were either squared (S) or absolute valued (A). See Chapter 5 for more details. However, while the power and false positive rate of the rotational and toroidal shift modification tests are shown for all six of the proposed statistics (i.e., all combinations of weights with both types of differences [DWS, UWS, CWS, DWA, UWA, and CWA]), the combined modification tests only employ the squared differences in the test statistics (i.e., DWS, UWS, and CWS). As a reminder, the tests with statistics which computed absolute differences in the ECDFs (i.e., the tests which used the DWA, UWA, and CWA statistics) were not considered since they showed little difference to the squared statistics (DWS, UWS, and CWS), and the squared statistics each achieved a marginally higher power (as shown in Figure 25).

Specifically, the false positive rate is computed by dividing the number of significant test results by the total number of tests computed when both samples come

Fig. 25: A comparison of the power achieved by the tests discussed in Sections 6.3 and 6.4 via a Cleveland dot plot. The dot colors and shapes separate the results into three categories: (1) modifications to the Syrjala test (blue circles), (2) the Syrjala test (red triangles), and (3) alternative tests (purple squares). The tabs on the right further separate the modifications to the Syrjala test into rotations (Rot Mod), toroidal shifts (Toro Mod), and both rotations and toroidal shifts (RotToro Mod). The Syrjala test is also separated by regular (Syr Reg) and random binning (Syr Ran) tabs. The remaining alternative tests (Alt Tests) are also grouped together. DWS, UWS, CWS, DWA, UWA, and CWA refer to the six different proposed statistics (Section 5.1) for measuring the differences in the ECDFs.

Table 8: A table listing the test description, number of significant tests, total number of tests, and power (first four columns from the left) for all of the tests considered in Figure 25. DWS, UWS, CWS, DWA, UWA, and CWA refer to the six different proposed statistics (Section 5.1) for measuring the differences in the ECDFs.

Test Description	Sig. Tests	Total Tests	Power	Syrjala Test Modification
DWS Rot	3891	4480	≈ 0.869	Rot Mod
UWS Rot	3894	4480	≈ 0.869	Rot Mod
CWS Rot	3896	4480	≈ 0.870	Rot Mod
DWA Rot	3877	4480	≈ 0.865	Rot Mod
UWA Rot	3874	4480	≈ 0.865	Rot Mod
CWA Rot	3878	4480	≈ 0.866	Rot Mod
DWS Toro	3658	3840	≈ 0.953	Toro Mod
UWS Toro	3663	3840	≈ 0.954	Toro Mod
CWS Toro	3662	3840	≈ 0.954	Toro Mod
DWA Toro	3653	3840	≈ 0.951	Toro Mod
UWA Toro	3654	3840	≈ 0.952	Toro Mod
CWA Toro	3655	3840	≈ 0.952	Toro Mod
DWS RotToro0.1	4261	4480	≈ 0.951	RotToro Mod
UWS RotToro0.1	4262	4480	≈ 0.951	RotToro Mod
CWS RotToro0.1	4267	4480	≈ 0.952	RotToro Mod
DWS RotToro0.2	4279	4480	≈ 0.965	RotToro Mod
UWS RotToro0.2	4280	4480	≈ 0.965	RotToro Mod
CWS RotToro0.2	4283	4480	≈ 0.966	RotToro Mod
DWS RotToro0.3	4271	4480	≈ 0.953	RotToro Mod
UWS RotToro0.3	4275	4480	≈ 0.954	RotToro Mod
CWS RotToro0.3	4276	4480	≈ 0.954	RotToro Mod
DWS RotToro25Thrshld	4269	4480	≈ 0.953	RotToro Mod
UWS RotToro25Thrshld	4270	4480	≈ 0.953	RotToro Mod
CWS RotToro25Thrshld	4273	4480	≈ 0.954	RotToro Mod
Syr Reg 5×5 Bins	477	640	≈ 0.745	NA
Syr Reg 10×10 Bins	553	640	≈ 0.864	NA
Syr Reg 20×20 Bins	560	640	≈ 0.875	NA
Syr Ran 25 Bins	433	640	≈ 0.677	NA
Syr Ran 100 Bins	545	640	≈ 0.852	NA
Syr Ran 400 Bins	562	640	≈ 0.878	NA
Energy	542	640	≈ 0.847	NA
Kmmd	550	640	≈ 0.859	NA
FR-KS	465	640	≈ 0.727	NA

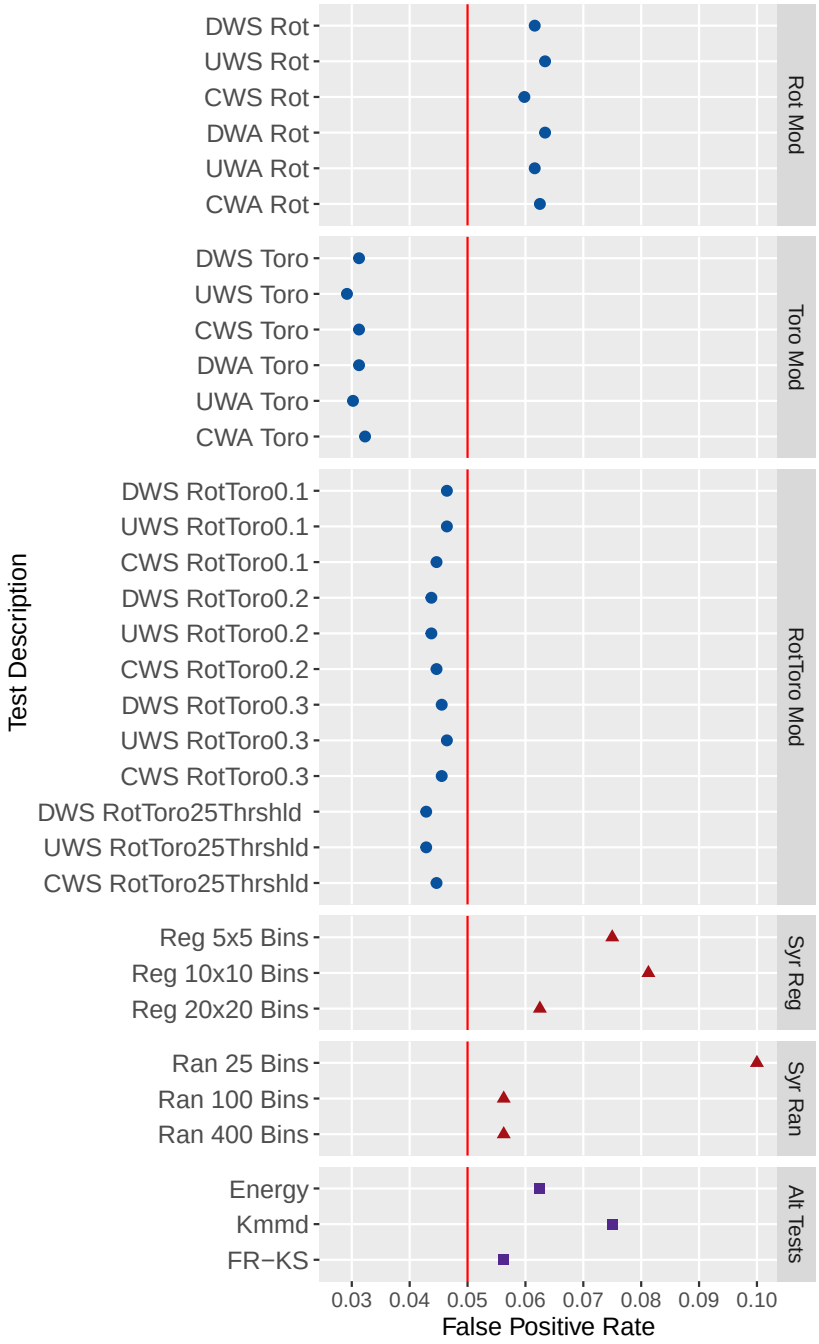


Fig. 26: A comparison of the false positive rates achieved by the tests discussed in Sections 6.3 and 6.4 via a Cleveland dot plot. The dot colors and shapes separate the results into three categories: (1) modifications to the Syrjala test (blue circles), (2) the Syrjala test (red triangles), and (3) alternative tests (purple squares). The tabs on the right further separate the modifications to the Syrjala test into rotations (Rot Mod), toroidal shifts (Toro Mod), and both rotations and toroidal shifts (RotToro Mod). The Syrjala test is also separated by regular (Syr Reg) and random binning (Syr Ran) tabs. The remaining alternative tests (Alt Tests) are also grouped together. DWS, UWS, CWS, DWA, UWA, and CWA refer to the six different proposed statistics (Section 5.1) for measuring the differences in the ECDFs. The vertical red line at 0.05 indicates the significance level of the tests.

Table 9: A table listing the test description, number of significant tests, total number of tests, and false positive rate (first four columns from the left) for all of the tests considered in Figure 26. DWS, UWS, CWS, DWA, UWA, and CWA refer to the six different proposed statistics (Section 5.1) for measuring the differences in the ECDFs.

Test Description	Sig. Tests	Total Tests	False Positive Rate	Syrjala Test Modification
DWS Rot	69	1120	≈ 0.062	Rot Mod
UWS Rot	71	1120	≈ 0.063	Rot Mod
CWS Rot	67	1120	≈ 0.060	Rot Mod
AbsV Rot	71	1120	≈ 0.063	Rot Mod
AbsV Rot	69	1120	≈ 0.062	Rot Mod
AbsV Rot	70	1120	≈ 0.062	Rot Mod
DWS Toro	30	960	≈ 0.031	Toro Mod
UWS Toro	28	960	≈ 0.029	Toro Mod
CWS Toro	30	960	≈ 0.031	Toro Mod
AbsV Toro	30	960	≈ 0.031	Toro Mod
AbsV Toro	29	960	≈ 0.030	Toro Mod
AbsV Toro	31	960	≈ 0.032	Toro Mod
DWS RotToro0.1	52	1120	≈ 0.046	RotToro Mod
UWS RotToro0.1	52	1120	≈ 0.046	RotToro Mod
CWS RotToro0.1	50	1120	≈ 0.045	RotToro Mod
DWS RotToro0.2	49	1120	≈ 0.044	RotToro Mod
UWS RotToro0.2	49	1120	≈ 0.044	RotToro Mod
CWS RotToro0.2	50	1120	≈ 0.045	RotToro Mod
DWS RotToro0.3	51	1120	≈ 0.046	RotToro Mod
UWS RotToro0.3	52	1120	≈ 0.046	RotToro Mod
CWS RotToro0.3	51	1120	≈ 0.046	RotToro Mod
DWS RotToro25Thrshld	48	1120	≈ 0.043	RotToro Mod
UWS RotToro25Thrshld	48	1120	≈ 0.043	RotToro Mod
CWS RotToro25Thrshld	50	1120	≈ 0.045	RotToro Mod
Syr Reg 5×5 Bins	12	160	$= 0.075$	NA
Syr Reg 10×10 Bins	13	160	≈ 0.081	NA
Syr Reg 20×20 Bins	10	160	≈ 0.062	NA
Syr Ran 25 Bins	16	160	≈ 0.100	NA
Syr Ran 100 Bins	9	160	≈ 0.056	NA
Syr Ran 400 Bins	9	160	≈ 0.056	NA
Energy	10	160	≈ 0.062	NA
Kmmd	12	160	$= 0.075$	NA
FR-KS	9	160	≈ 0.056	NA

from the same CSR distribution (first row of graphs in Figures 17–23 and 62–79). For example, the Syrjala test (see Figure 17) which employed 5×5 regular binning resulted in 12 out of 160 tests being false positives, i.e., a false positive rate of 0.075. For the modified Syrjala tests, this computation was applied to the aggregation of tests across all rotations or toroidal shifts values. For example, the rotational modified Syrjala test which used the CWS statistic resulted in 67 out of 1120 tests being false positives, i.e., a false positive rate of 0.05982.

For false positive rates in Figure 26, test results should be as close as possible to 0.05 (i.e., 5%, indicated by the horizontal line) when testing at the 5% significance level. Test results which fall below 0.05 are indications of a conservative nature in the test (i.e., a test which is less likely to reject the null when it is actually true). In Figure 25, the higher the power of a test the more likely the test is to reject the null when it is indeed false. Theoretically, the maximum power a test can achieve is one.

In Figure 25, the power was computed by dividing the number of significant tests by the total number of tests in which the null hypothesis was false. For the Energy, Kmmd, and FR-KS tests, the total number of tests was 640 (ten replications times four n_1 sample sizes times four n_2 sample sizes times four departures from CSR). Since the rotational, toroidal shift, combined modifications tests were simulated across a variety of number of rotations and proportions of randomly selected points used for the origins of the toroidal shifts, or both, respectively, the total number of tests for these cases were as follows: 4,480 (ten replications times seven rotational levels times four n_1 sample sizes times four n_2 sample sizes times four departures from CSR) rotational tests, 3,840 (ten replications times six proportion levels times four n_1 sample sizes times four n_2 sample sizes times four departures from CSR) toroidal tests, and 4,480 combined modification tests. The number of significant tests for all of the tests (given that the null hypothesis is false) are reported in Table 8.

The superiority of the toroidal shift and combined modifications over the rotational modification and other alternative methods (Energy, Kmmd, FR-KS, and binned Syrjala) is clearly seen in Figure 25. Additionally, while Figures 25 and 26 show a considerable difference between the toroidal shift and rotational modifications, little difference is seen across the different test statistics (DWS, UWS, CWS, DWA, UWA, and CWA). Specifically, the power of the tests which employ toroidal shift modifications is approximately 0.08 higher than the rotational modification tests on average. However, the power of tests which involve squared differences in the ECDF values is only approximately 0.003 higher than tests which use absolute differences in ECDF values on average. Additionally, the relative stability in results suggest less computationally intensive tests may be employed without a sacrifice in performance.

Additionally, while the toroidal shift modification clearly outperforms the rotational modification, the combination of both modifications did not achieve a considerably higher power than the toroidal shift modification alone. Specifically, the mean power across all of the rotational tests is 0.867, while the mean power for the toroidal shift and combined modification tests are 0.952 and 0.954, respectively. Hence, one could argue that the toroidal shift modification is sufficient since it will provide almost identical power as compared to the much more computationally intensive combination of modifications. However, Figure 26 shows that the combined test provides a more appropriately conservative (i.e., a false positive rate closer yet still below the significance level of 0.5) test as compared to the toroidal shift modification alone. Specifically, the average false positive rate of the toroidal shift modification is 0.03, which is more conservative than the average 0.045 false positive rate of the combined modification. Due to the trade-off between the false positive rate and power of a test, the test with false positive rate closer to the significance level is expected to be more powerful in the face of all departures from the null hypothesis. Hence, the

combined rotational and toroidal shift test is recommended as a more powerful, yet still conservative, choice in the face of all departures from the null.

Keep in mind that the number of total tests with respect to the false positive rate of each method in Figure 26 differs similar to Figure 25. For the Energy, Kmmd, and FR-KS tests, the total number of tests was 160 (ten replications times four n_1 sample sizes times four n_2 sample sizes). Similarly, the rotational, toroidal shift, and combined modifications tests had a total number of tests as follows: 1,120 (ten replications times seven rotational levels times four n_1 sample sizes times four n_2 sample sizes) rotational tests, 960 (ten replications times six proportion levels times four n_1 sample sizes times four n_2 sample sizes) toroidal tests, and 1,120 combined modification tests. The number of significant tests for all of the tests (given that the null hypothesis is true) are reported in Table 9.

Figure 26 shows that while the Energy, Kmmd, FR-KS and rotational modification tests all demonstrated false positive rate levels above the significance level, the toroidal and combined modification tests achieved conservative false positive rate levels (below the significance level). Furthermore, while the combined modification tests achieved about the same power as the toroidal shift modification tests (as seen in Figure 25), the combined modification test proves to be closer to the significance level than the toroidal modification test while still remaining conservative. The Kmmd test is the most anti-conservative test, while the FR-KS is the least anti-conservative. Hence, this figure, similar to Figure 25, demonstrates the superiority of the combined modification test over other methods as the most powerful yet conservative choice when applied to data of a similar nature.

Furthermore, there is no linear increase in performance as the proportion of points used for toroidal shifts increases for the combined rotational and toroidal shift modified Syrjala tests. Specifically, there is no improvement to the false positive

rate when increasing the proportion of points from 0.1 to 0.2. Additionally, fixing the number of toroidal shifts to a threshold of 25 randomly selected points (RotToro25Thrshld) performs similarly for all of the squared test statistics. Hence, these results provide motivation to a default threshold value of 25 for the tests in the `R` package (see Chapter 8).

6.5 Eye-Tracking Inspired Simulation Study

Another series of simulations were conducted to demonstrate the performance of the modified Syrjala tests on data which more closely represent cases taken from eye-tracking research. Due to the relatively stable results demonstrated by the modified Syrjala test (using both rotational and toroidal shifted modifications) in Sections 6.2–6.4, only the test which employs the CWS statistic (see Section 5.1) with eight rotations and uses 0.1 proportion of points as toroidal shift origins was used across these simulations. This modified Syrjala test has been shown to be one of the most powerful yet conservative choices among the tests which were previously compared in Section 6.4.2. The use of this version of the modified Syrjala tests is also justified by the stability of the results across a wider range of sample sizes within Sections 6.5.3 and 6.5.9. In Sections 6.5.3 and 6.5.9 the modified Syrjala test is again explored across rotations (namely, 4, 5, 6, 8, 10, 36, and 45), proportions of points for toroidal shifts (namely, 0.1, 0.2, 0.3, 0.5, 0.75, and 0.9), and thresholds for the number of toroidal shifts (namely, 15, 25, and 40).

6.5.1 Generated Data Structure

A variety of data structures were used across these simulations to demonstrate the performance of the modified Syrjala tests on scenarios which are patterned more closely to those seen in eye-tracking research. Specifically, gaze point fixations are

modeled as bivariate normal distributions with sample sizes of 15, 25, 40, 70, 100, 250, and 500 points. These sample sizes were chosen to provide realizations similar to those seen in the USU Posture Study data (analyzed in Chapter 7). Six of the following detailed simulations generate a single bivariate normal distribution as a null hypothesis which represents a subject concentrating their visual attention on one object. From here, departures from this behavior are compared as alternative hypotheses which exhibit differences in location, shape, allocation of visual attention between one and multiple objects, and the introduction of few or many additional noise gaze points. These are detailed in Sections 6.5.3–6.5.8.

Two additional simulations (in Sections 6.5.9 and 6.5.10) use mixture distributions to model a specific case taken from the USU Posture Study (Symanzik et al., 2017, 2018; Studenka et al., 2020; Coltrin et al., 2020; McKinney and Symanzik, 2019, 2021). Here, the collections of gaze points from treatment and control groups are modeled as clusters of bivariate normal distributions. Simulated data similar to posture ID 17 (see Appendix C.1) is compared with departures in fixation location, shape, and the introduction of few or many additional noise gaze points for the initial simulation, and differences in allocation of visual attention across the posture image for the latter simulation. The generated data for each simulation was produced on the $[0, 1] \times [0, 1]$ square. Additionally, common random numbers and random number seeds were employed similarly to previous simulations (see Section 6.1).

6.5.2 General Structure of Results

Many of the results in the following Sections 6.5.3–6.5.10 are of a similar format, except for the simulation where an equal amount of noise is generated for both samples in Section 6.5.8, and where the modified Syrjala tests are explored across a variety of parameters in Sections 6.5.3 and 6.5.9. These exceptions are explained in detail

in their respective sections. Otherwise, the figures which display the results have a similar layout. Each figure displays a grid of line graphs showing the performance of the modified Syrjala test (using 0.1 proportion of points as origins of toroidal shifts, 8 rotations, and the CWS statistic). The vertical axes display the number of significant tests out of ten tests (each conducted at the 0.05 significance level), the grid column name indicates the first sample size (n_1), and horizontal axes indicate the second sample size (n_2). The grid row indicates the shape of the second sample. However, the shape of the first sample follows the distribution exhibited in the first row in all cases (except for the figure in Section 6.5.8 where both samples shapes are indicated in the grid row).

Hence, the first row demonstrates the performance of the test when the null hypothesis is true, and the two samples originate from the same distribution. Whereas, the remaining four rows display the performance of the test for departures from the first sample's distribution (i.e., when the null hypothesis is false). In the first row, each significant test is a false positive since both samples are being drawn from the same distribution. Thus, it is expected to see roughly 5% of the 490 tests (ten tests times seven sample sizes for n_1 times seven sample sizes for n_2) as significant by chance variation. This is about 24.5 tests on average.

As an example, the bottom left subgraph shows the number of significant tests (out of ten tests) between the first samples with 15 points and second samples with 15, 25, 40, 70, 100, 250, and 500 points within the respective simulation. Additionally, each figure includes a column of scatterplots in the far right side of the figure which depict realizations of the second sample when the sample size is 500.

6.5.3 Simulating Differences in Fixation Location

The null hypothesis for this initial simulation assumes a subject is concentrating

their visual attention on a single object near the bottom left corner of the $[0,1]\times[0,1]$ square. This is modeled as a bivariate normal distribution with mean coordinates of $(0.25, 0.25)$ and variance-covariance matrix of $\left[\begin{smallmatrix} 0.005 & 0 \\ 0 & 0.005 \end{smallmatrix}\right]$.

Departures from the null hypothesis are modeled as identical distributions that are shifted up and to the right by increments of 0.025 in the vertical and horizontal directions (i.e., shifted diagonally toward the upper right corner by increments of $\sqrt{(0.025)^2 + (0.025)^2} = \frac{\sqrt{2}}{40} \approx 0.04$). Hence, four departures from the null hypothesis are modeled as bivariate normal distributions with variance-covariance matrices of $\left[\begin{smallmatrix} 0.005 & 0 \\ 0 & 0.005 \end{smallmatrix}\right]$ and bivariate mean coordinates of $(0.275, 0.275)$, $(0.3, 0.3)$, $(0.325, 0.325)$, $(0.35, 0.35)$, respectively.

Results Across All Modified Syrjala Tests

In Sections 6.3–6.4.2 the modified Syrjala tests were shown to exhibit stable results across an array of rotations, toroidal shifts, and both rotations and toroidal shifts in simulations which involved departures from CSR data. It was also shown that little difference in the test results are observed across six different statistics within the modified Syrjala tests. As the tests are being applied to eye-tracking inspired data throughout the subsections of Section 6.5, the six statistics together with the modifications (rotations, toroidal shifts, or both combined) are employed again to establish stable results across smaller sample sizes using eye-tracking inspired data, and to reaffirm the sensible default parameter values (initially proposed from the simulations with larger samples in Section 6.4.2) for the `distdiffR R` package functions (described in more detail in Chapter 8). The default parameters are the CWS statistic, eight rotations, and a threshold of 25 toroidal shifts or 0.1 for a proportion of points to be used as toroidal shifts. Similar results are also provided in Section 6.5.9 where the simulated data is patterned more closely to eye-tracking

data. These results also support the recommended default parameter values for the combined rotational and toroidal shift tests.

Consequently, the format of Figures 27–34 differs slightly from the others within Sections 6.3–6.4.2. While these figures still display grids of line graphs making comparisons between simulated differences in fixation location (detailed in Section 6.5.3) for the same sample sizes (15, 25, 40, 70, 100, 250, and 500), the horizontal axes differs from that described in Section 6.5.2. The horizontal axes show the number of rotations or proportions of toroidal shifts (similar to the simulation Figures 18–22 in Section 6.3). The abbreviation "cenbl" in the far right of the top row indicates that a bivariate normal distribution was used to generate the data in the bottom-left corner of the unit square. The number following cenbl in the far right of the remaining rows indicates the coordinates of the alternative distribution's bivariate mean, e.g., cenbl_0.325 indicates that the second sample was generated using a bivariate normal distribution centered at (0.325, 0.325).

While, Figure 27 shows the test results of the rotational modification for only the CWS statistic, the other five statistics were also explored (DWS, UWS, DWA, UWA, and CWA). However, similar to what has been seen in Section 6.3.1, the results of the tests displayed little difference regardless of the test statistic. Hence, the figures displaying those test results (Figures 80–84) are provided in Appendix B. Similarly, this is also true of the toroidal shift (Figure 28) in this section. Consequently, the performance of the other five statistics when used within the toroidal shift modification are also provided in Appendix B (as Figures 85–89). Due to the stability in results exhibited across the six statistics for the rotational and toroidal shift tests, only the CWS statistic was used when applying the combined rotational and toroidal shift modified Syrjala tests. These tests are explored using 0.1, 0.2, and 0.3 proportions of points for toroidal shift origins in Figures 29–31, respectively, and 15, 25, and 40 as

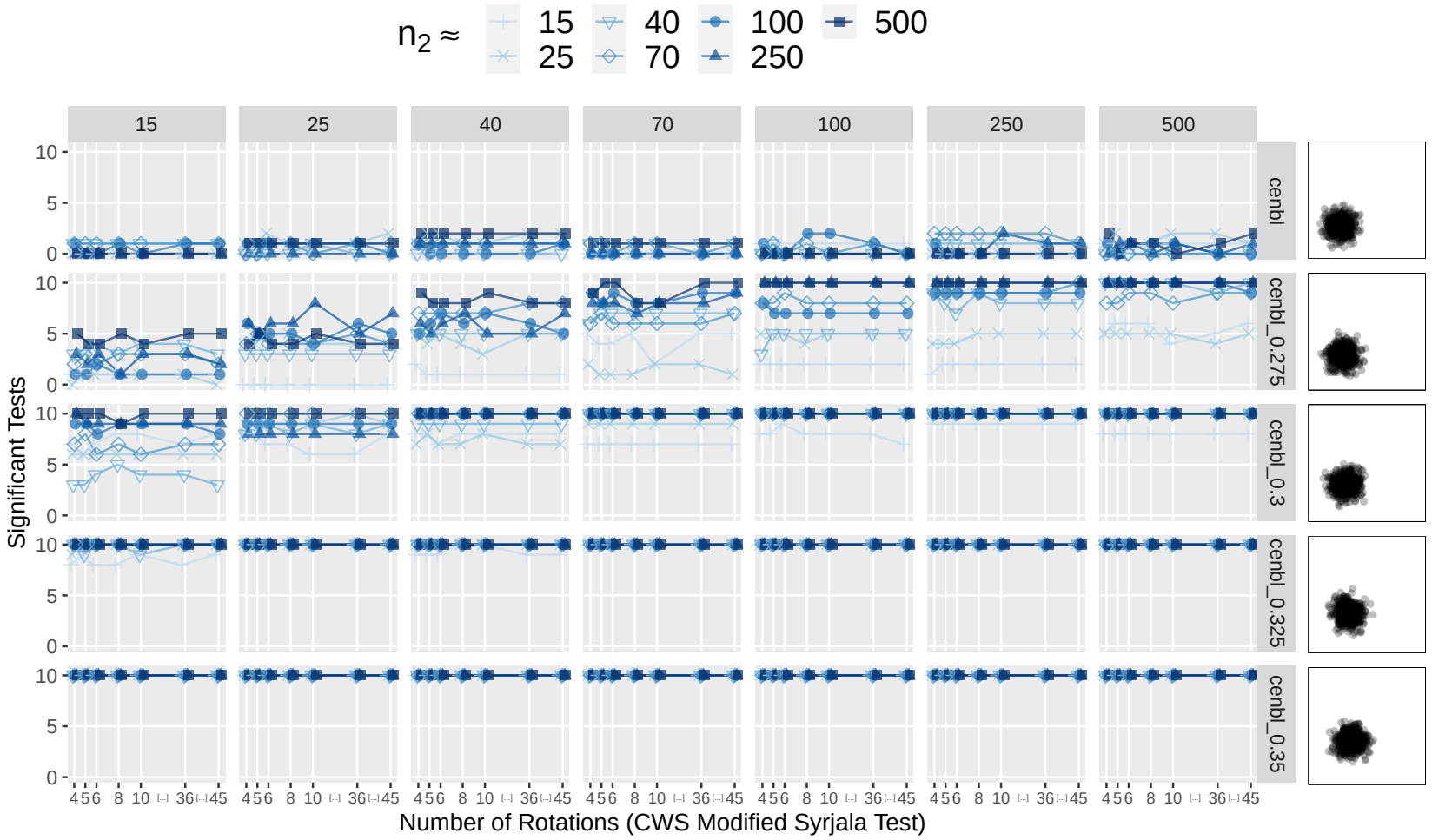


Fig. 27: A grid of line graphs showing the results of a simulation comparing multiple rotations of the modified Syrjala test (using the CWS statistic) on simulated eye-tracking data where subjects concentrate on a single object at differing locations. The grid column name indicates the first sample size (n_1), and the grid row indicates the shape of the second sample. The shape of the first sample follows the same distribution exhibited in the first row. The horizontal axis indicates the second sample size (n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) between the first samples with 15 points and second samples with 15, 25, 40, 70, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

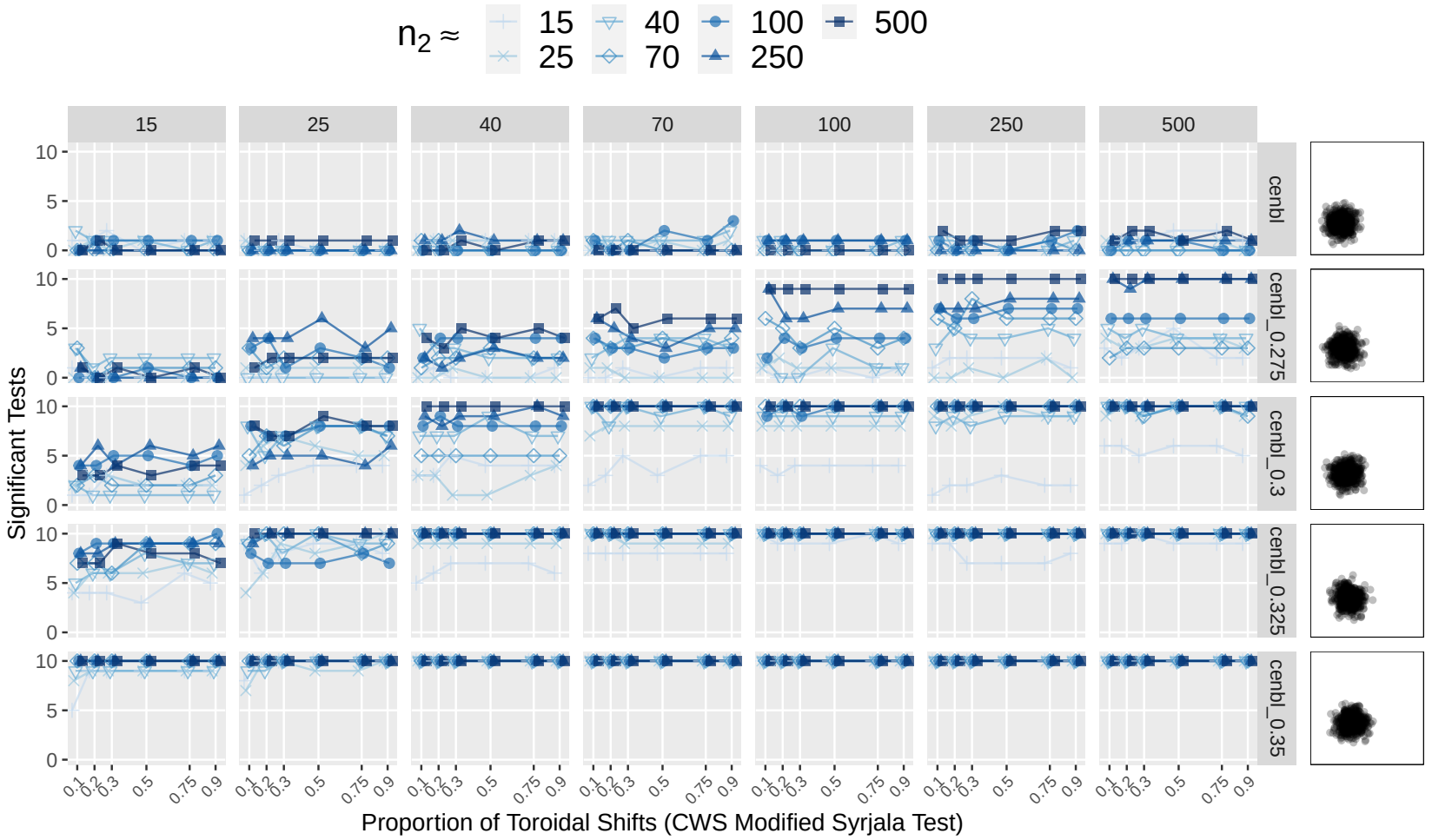


Fig. 28: A grid of line graphs showing the results of a simulation comparing multiple proportions of toroidal shifts of the modified Syrjala test (using the CWS statistic) on simulated eye-tracking data where subjects concentrate on a single object at differing locations. The grid column name indicates the first sample size (n_1), and the grid row indicates the shape of the second sample. The shape of the first sample follows the same distribution exhibited in the first row. The horizontal axis indicates the second sample size (n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) between the first samples with 15 points and second samples with 15, 25, 40, 70, 100, 250, and 500 points.

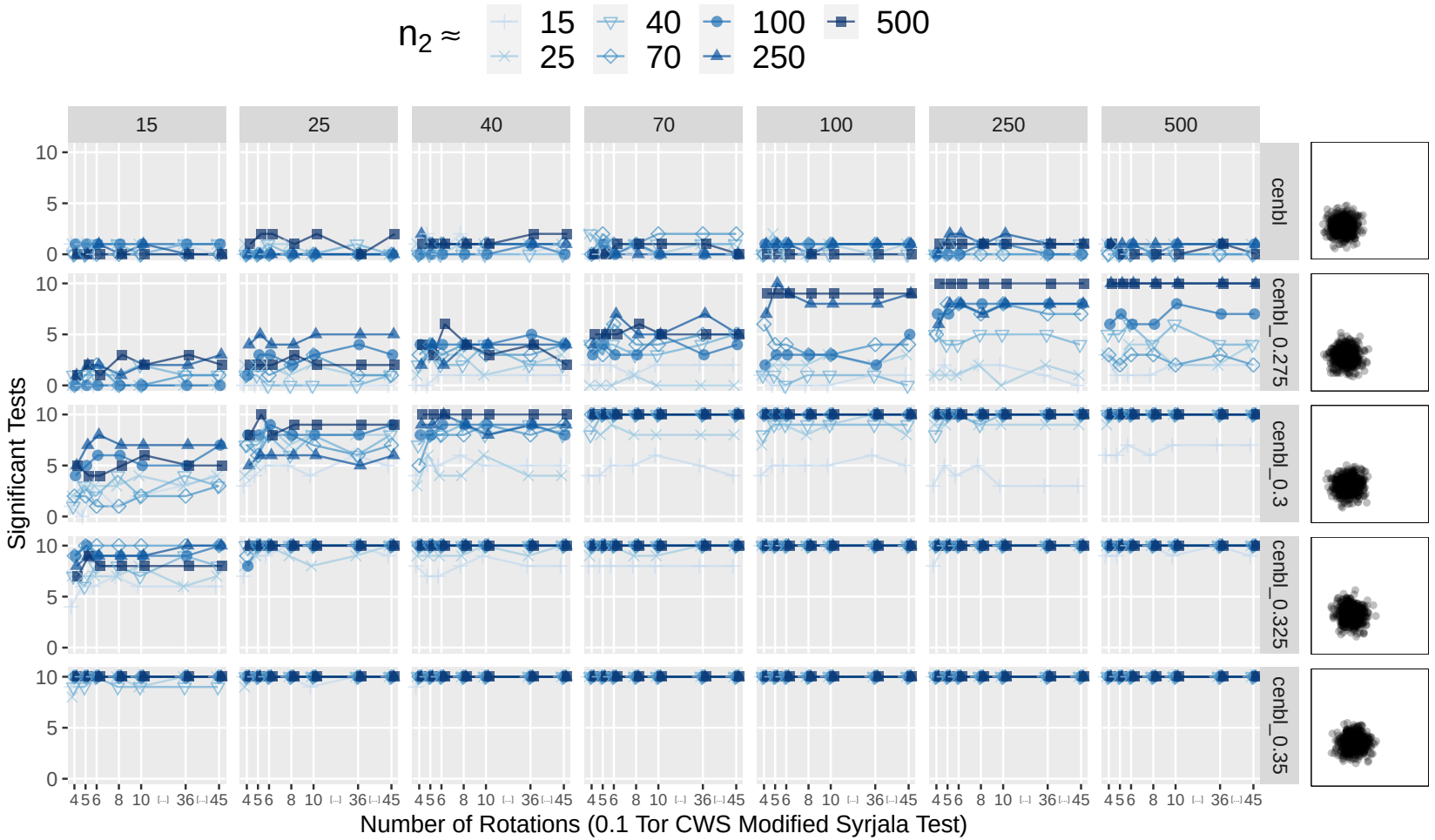


Fig. 29: A grid of line graphs showing the performance across a variety of rotations of the modified Syrjala test (using 0.1 proportion of points as origins of toroidal shifts, and the CWS statistic) on simulated eye-tracking data where subjects concentrate on a single object at differing locations. The grid column name indicates the first sample size (n_1), and the grid row indicates the shape of the second sample. The shape of the first sample follows the same distribution exhibited in the first row. The horizontal axis indicates the second sample size (n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) between the first samples with 15 points and second samples with 15, 25, 40, 70, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

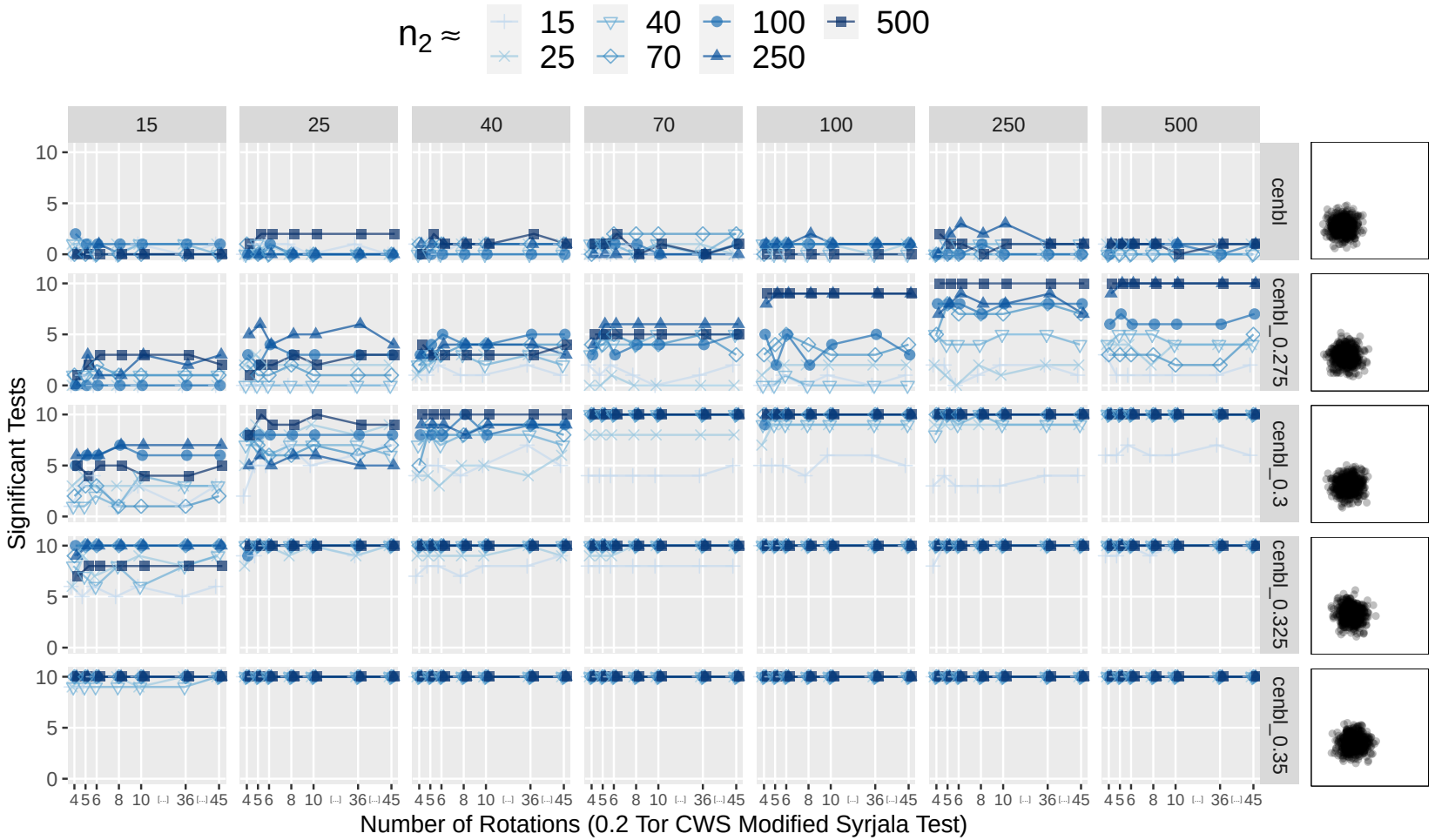


Fig. 30: A grid of line graphs showing the performance across a variety of rotations of the modified Syrjala test (using 0.2 proportion of points as origins of toroidal shifts, and the CWS statistic) on simulated eye-tracking data where subjects concentrate on a single object at differing locations. The grid column name indicates the first sample size (n_1), and the grid row indicates the shape of the second sample. The shape of the first sample follows the same distribution exhibited in the first row. The horizontal axis indicates the second sample size (n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) between the first samples with 15 points and second samples with 15, 25, 40, 70, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

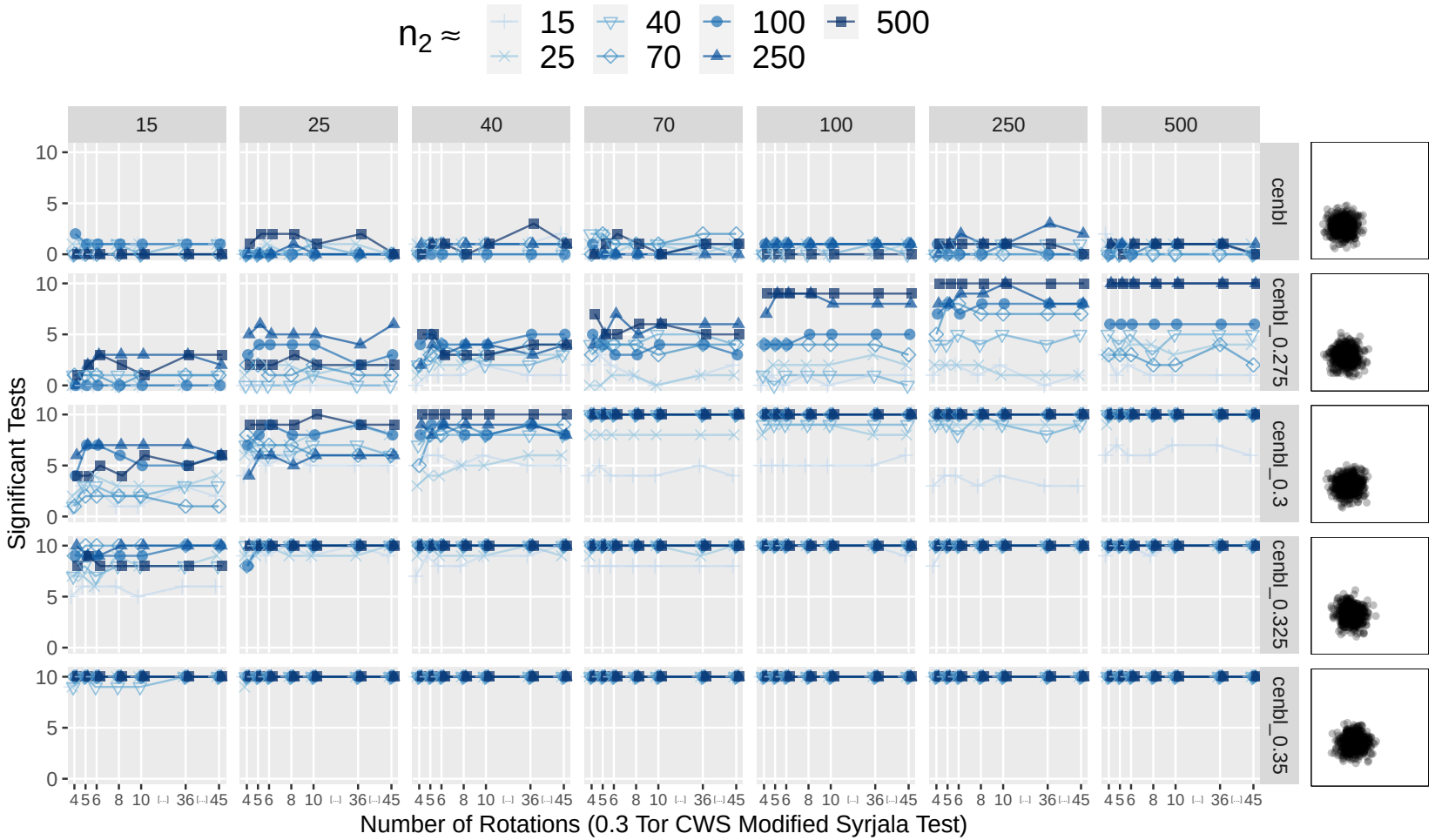


Fig. 31: A grid of line graphs showing the performance across a variety of rotations of the modified Syrjala test (using 0.3 proportion of points as origins of toroidal shifts, and the CWS statistic) on simulated eye-tracking data where subjects concentrate on a single object at differing locations. The grid column name indicates the first sample size (n_1), and the grid row indicates the shape of the second sample. The shape of the first sample follows the same distribution exhibited in the first row. The horizontal axis indicates the second sample size (n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) between the first samples with 15 points and second samples with 15, 25, 40, 70, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

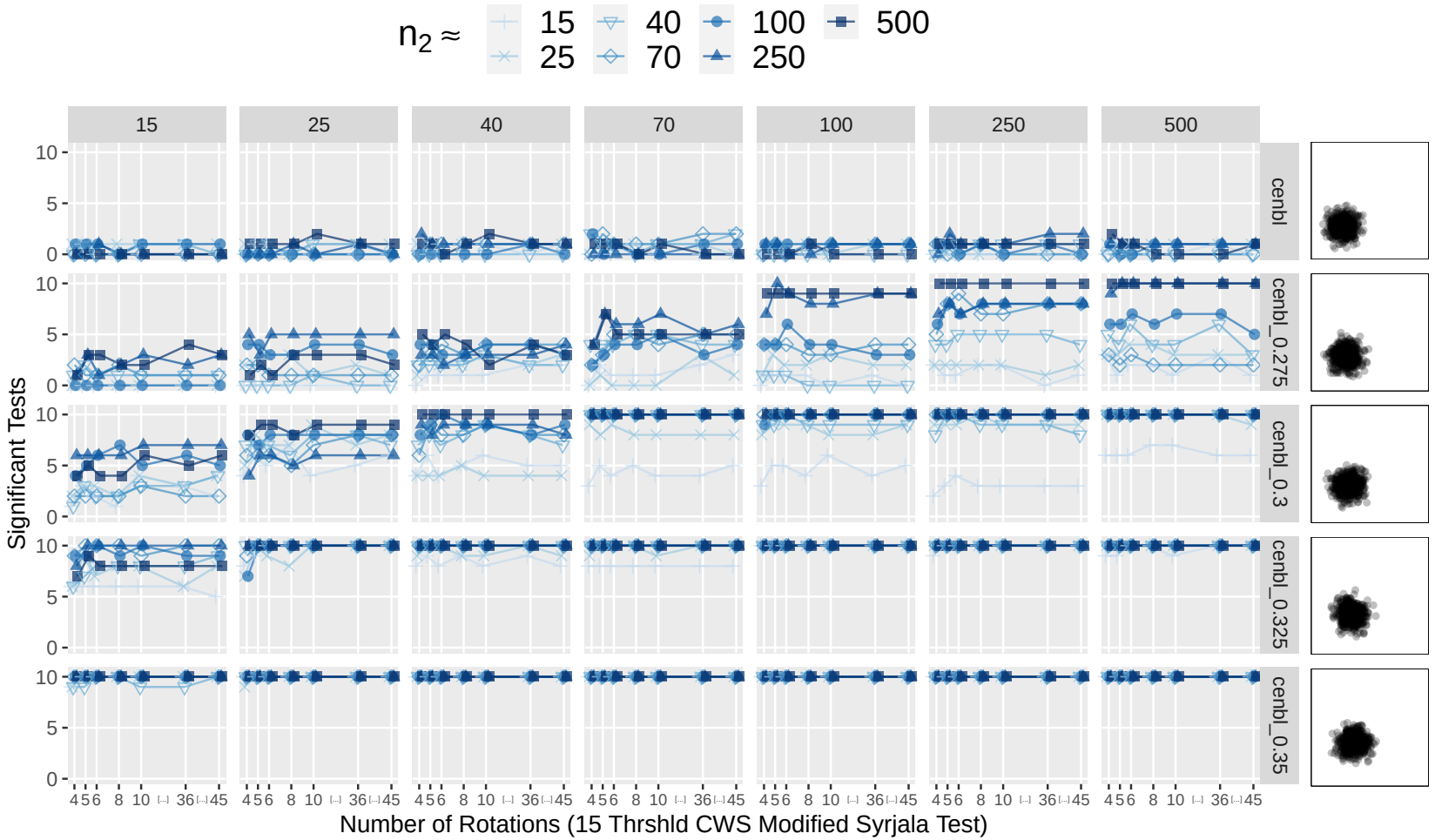


Fig. 32: A grid of line graphs showing the performance across a variety of rotations of the modified Syrjala test (using a threshold of 15 toroidal shifts, and the CWS statistic) on simulated eye-tracking data where subjects concentrate on a single object at differing locations. The grid column name indicates the first sample size (n_1), and the grid row indicates the shape of the second sample. The shape of the first sample follows the same distribution exhibited in the first row. The horizontal axis indicates the second sample size (n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) between the first samples with 15 points and second samples with 15, 25, 40, 70, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

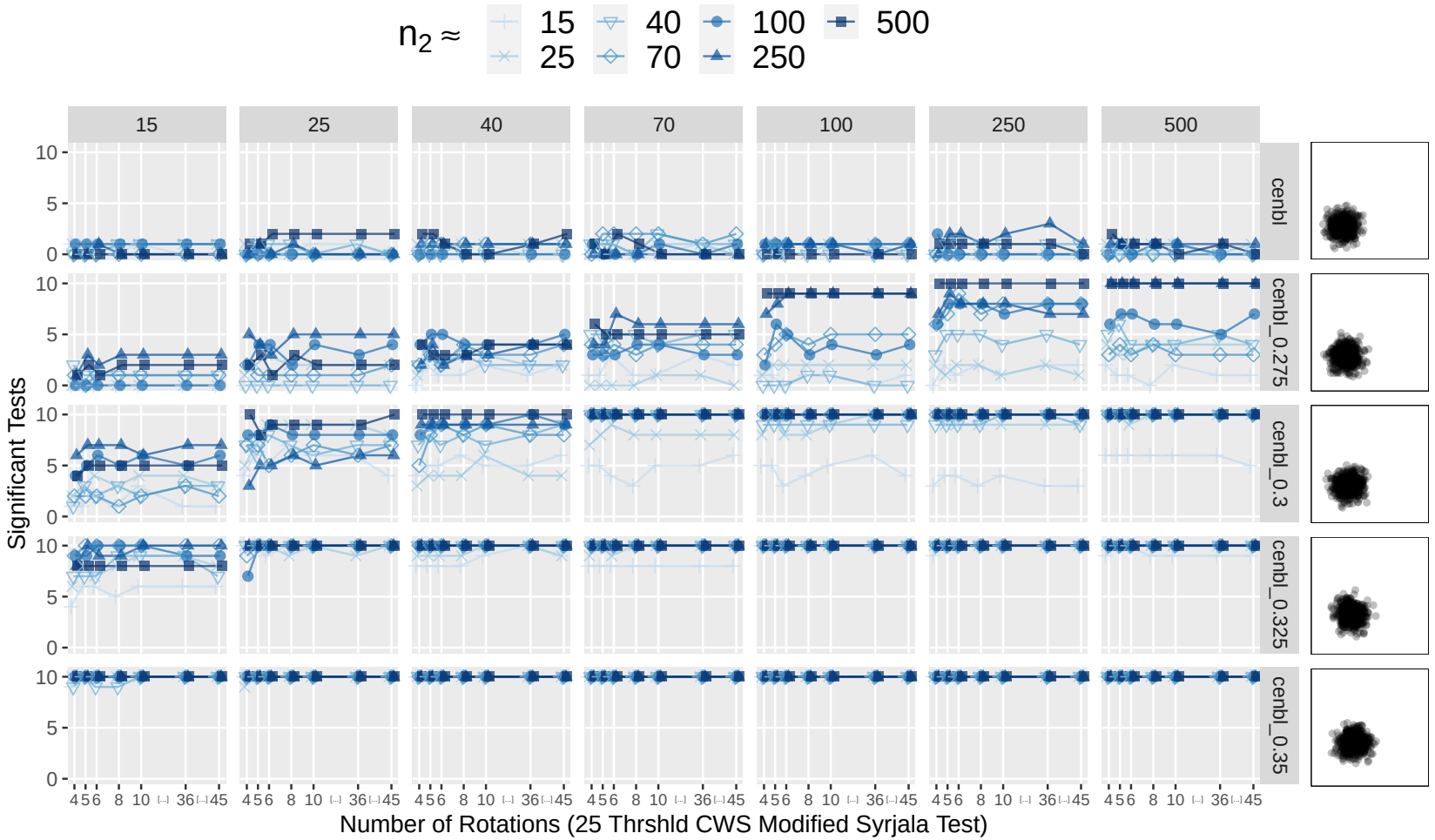


Fig. 33: A grid of line graphs showing the performance across a variety of rotations of the modified Syrjala test (using a threshold of 25 toroidal shifts, and the CWS statistic) on simulated eye-tracking data where subjects concentrate on a single object at differing locations. The grid column name indicates the first sample size (n_1), and the grid row indicates the shape of the second sample. The shape of the first sample follows the same distribution exhibited in the first row. The horizontal axis indicates the second sample size (n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) between the first samples with 15 points and second samples with 15, 25, 40, 70, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

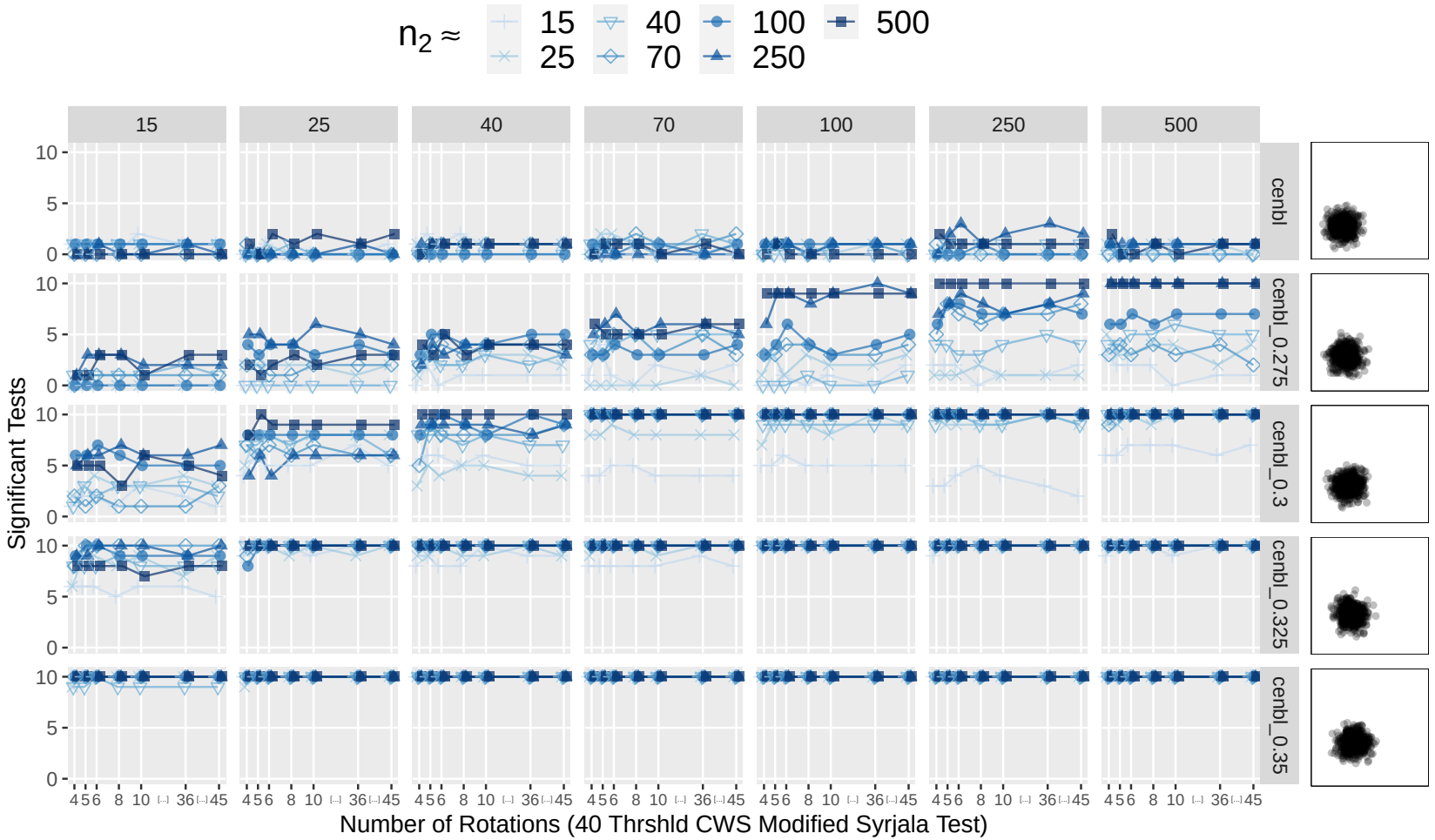


Fig. 34: A grid of line graphs showing the performance across a variety of rotations of the modified Syrjala test (using a threshold of 40 toroidal shifts, and the CWS statistic) on simulated eye-tracking data where subjects concentrate on a single object at differing locations. The grid column name indicates the first sample size (n_1), and the grid row indicates the shape of the second sample. The shape of the first sample follows the same distribution exhibited in the first row. The horizontal axis indicates the second sample size (n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) between the first samples with 15 points and second samples with 15, 25, 40, 70, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

thresholds for the number of toroidal shifts employed in Figures 32–34, respectively. Hence, as in Sections 6.3.1–6.3.4, only the results of tests using the CWS statistic are shown in this section for consistency.

Specifically, Figure 27 shows that across all of the significant tests out of the ten replications of the tests (vertical axes) across all of the number of rotations (horizontal axes) the rotational modification test returns stable results except for some chance variation. The test is capable of detecting almost all of the significant differences in the bottom two rows, and captures a majority of differences in the third row for all cases except for when $n_1 = 15$ and $n_2 \approx 40$. In the first row, 178 out of the 3430 tests are significant. This is roughly a 0.052 (or 5.2%) false positive rate. This result is on target given that the significance level of the test is 5%.

In Section 6.3.2, the toroidal shift was shown to be more powerful in general than the rotational test by simulation. However, Figure 28 shows the toroidal shift modification results in an overall decrease in the numbers of significant tests for the cases when the null hypothesis is false (bottom four rows) particularly for small sample sizes. Still, the test is capable of detecting almost all of the significant differences in the bottom row when the bivariate mean has shifted to the right and up by 0.35, respectively. In the first row, 137 out of the 2940 are significant. This is roughly a 0.045 (or 4.5%) false positive rate. This result is slightly conservative given that the significance level of the test is 5%.

Figures 29–31 show the test results for the combined rotational and toroidal shift modified Syrjala tests, where the horizontal axes shows the number of rotations and the proportions of toroidal shifts are 0.1, 0.2, and 0.3 in Figures 29–31, respectively. Although some increase in power is observed over using toroidal shifts alone, the test is still not as powerful as the rotational test for these simulated data. Again, the test results also confirm relatively stable performance across the number of rotations and

the proportions of points used as origins for the toroidal shifts. The false positive rates given by the proportion of significant tests to the total number of tests in the first row of subplots in Figures 29–31 is discussed more explicitly later in this section.

Instead of proportions of points for the toroidal shifts, the test results for the combined rotational and toroidal shift modified Syrjala tests, where the horizontal axes is the number of rotations and thresholds for the number of toroidal shifts are 15, 25, and 40 are shown in Figures 32–34, respectively. Again, some increase in power is observed over using toroidal shifts alone. However, the test is still not as powerful as the rotational test for these simulated data. Although these tests limit the number of toroidal shifts to the lower of the combined sample sizes or the shift threshold, the results are almost identical to that of the combined rotational and toroidal shift test where proportions of points are used for toroidal shifts. Also, the test results confirm relatively stable performance across the number of rotations and the thresholds of points used as origins for the toroidal shifts.

As a reminder, common random numbers are being employed across all of the simulations (see Section 6.1.3). Thus, the same ten replications of each simulation scenario (e.g., cenbl vs. cenbl_0.3) pairs are being compared across these simulations (Figures 27–34). Therefore, it is not reasonable to say that the unusually low number of significant tests for the case when $n_1 = 250$ and $n_2 \approx 15$ for combined modifications are due to the differences in test parameters (i.e., proportion of toroidal shifts or toroidal shift threshold size). The decrease in performance can be attributed to chance variation within the small second sample. However, the rotational test was notably able to detect all of the significant differences for these unusual cases.

An overview of the power and false positive rates of the tests employed in this section are provided in Figures 35 and 36 and Tables 10 and 11. While the rotational and toroidal shift modification tests are employed using all six of the test statistics

(top two subplots in Figures 35 and 36), due to the stability of the test results regardless of the test statistic used only the CWS statistic was used within the combined modification test (bottom subplot in Figures 35 and 36). As a reminder, the power was computed by dividing the number of significant tests by the total number of tests in which the null hypothesis was false (bottom four rows of graphs in Figures 27–34). In Figure 35, the higher the power of a test the more likely the test is to reject the null when it is indeed false. Theoretically, the maximum power a test can achieve is one.

The higher power of the rotational tests above the toroidal tests is clearly seen in Figure 35. Furthermore, if only the CWS statistic is compared across the tests, while some power is gained from using the combined test over the toroidal test, the rotational test is also shown to be more powerful than the combined rotational and toroidal tests. This is contrary to the results shown in Section 6.4.2 where the combined rotational and toroidal modifications test was shown to be the most powerful and appropriately conservative test. However, the results in Figure 35 make sense given the context of the simulations.

Recall that Section 6.4.2 provided the power and false positive rate summaries for simulations which involved departures from completely spatially random data. However, in this section the simulation involves shifts in circular bivariate normal distributions which model eye-tracking fixation distributions. Also recall that the Syrjala test has been shown to place a greater emphasis on data located close to the four corners of the bounding rectangle (see Section 6.2.2 and McAdam et al. (2012)). The rotational modified Syrjala tests have also been shown to place a greater emphasis on data near the outer edge of the pooled bivariate distributions (see Section 6.3.1). Hence, the high power of the rotational test in Figure 35 makes sense since the largest differences in the bivariate empirical cumulative density functions lie around

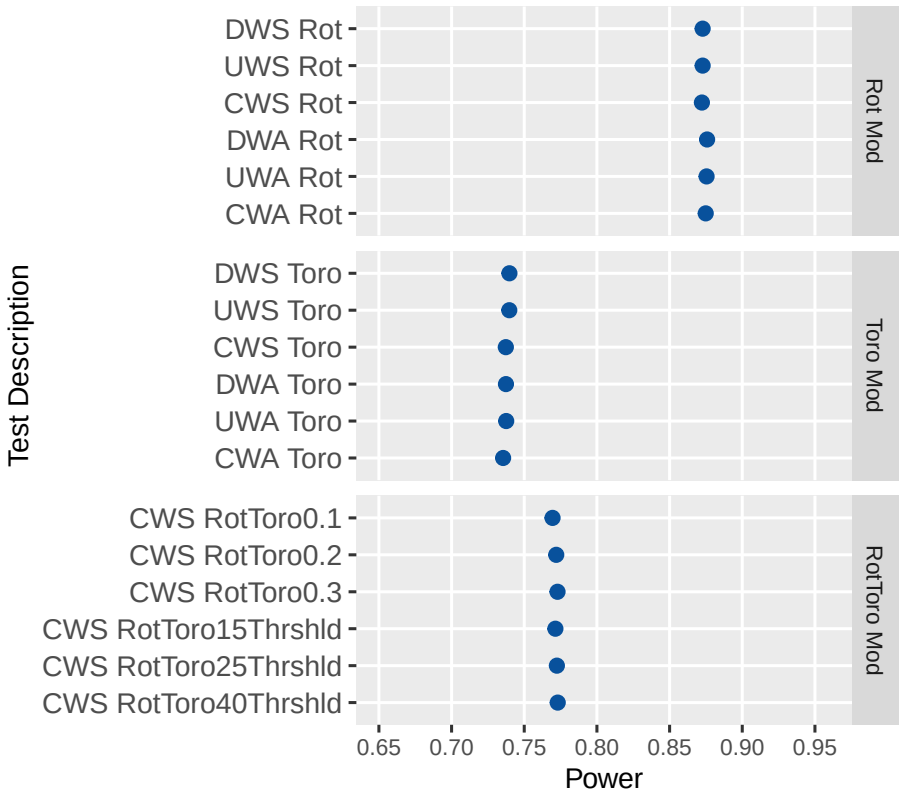


Fig. 35: A comparison of the power achieved by the tests discussed in this section via a Cleveland dot plot. The tabs on the right separate the modifications into rotations (Rot Mod), toroidal shifts (Toro Mod), and both rotations and toroidal shifts (RotToro Mod). DWS, UWS, CWS, DWA, UWA, and CWA refer to the six different proposed statistics (Section 5.1) for measuring the differences in the ECDFs. While the rotational and toroidal shift modification tests are employed using all six of the test statistics (top two subplots), only the CWS statistic was used within the combined modification test (bottom subplot).

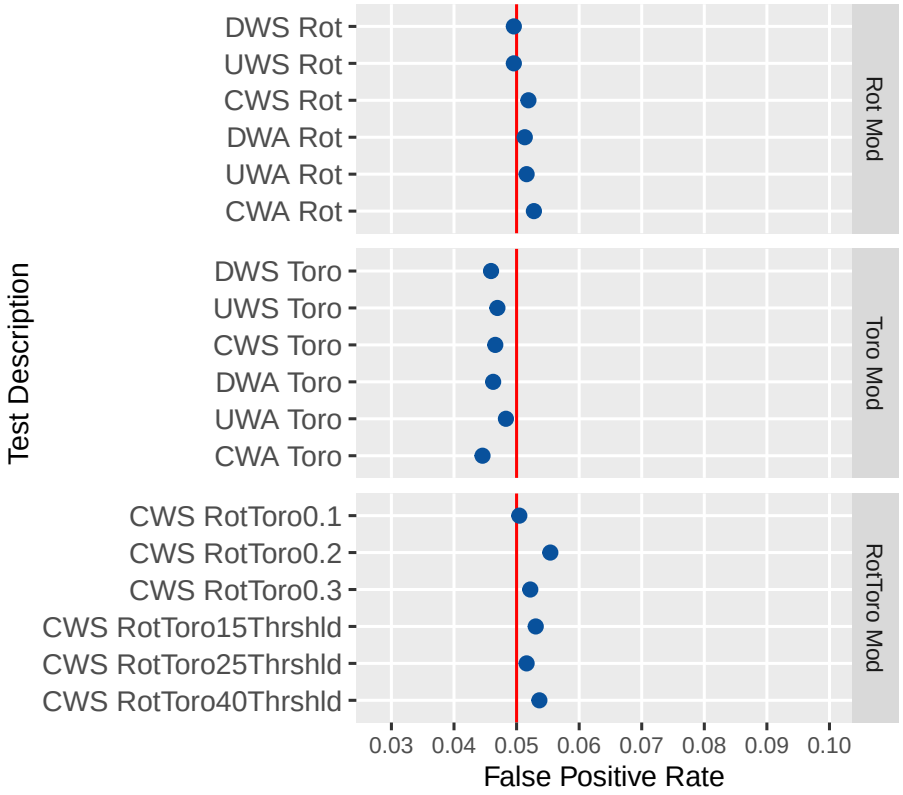


Fig. 36: A comparison of the false positive rates achieved by the tests discussed in this section via a Cleveland dot plot. The tabs on the right separate the modifications into rotations (Rot Mod), toroidal shifts (Toro Mod), and both rotations and toroidal shifts (RotToro Mod). DWS, UWS, CWS, DWA, UWA, and CWA refer to the six different proposed statistics (Section 5.1) for measuring the differences in the ECDFs. While the rotational and toroidal shift modification tests are employed using all six of the test statistics (top two subplots), only the CWS statistic was used within the combined modification test (bottom subplot). The vertical red line at 0.05 indicates the significance level of the tests.

Table 10: A table listing the test description, number of significant tests, total number of tests, and power (rounded to the third decimal place) for all of the tests considered in Figure 35.

Test Description	Sig. Tests	Total Tests	Power
DWS Rot	11 974	13 720	≈ 0.873
UWS Rot	11 974	13 720	≈ 0.873
CWS Rot	11 967	13 720	≈ 0.872
DWA Rot	12 016	13 720	≈ 0.876
UWA Rot	12 011	13 720	≈ 0.875
CWA Rot	12 003	13 720	≈ 0.875
DWS Toro	8700	11 760	≈ 0.740
UWS Toro	8699	11 760	≈ 0.740
CWS Toro	8671	11 760	≈ 0.737
DWA Toro	8672	11 760	≈ 0.737
UWA Toro	8674	11 760	≈ 0.738
CWA Toro	8649	11 760	≈ 0.735
CWS RotToro0.1	10 557	13 720	≈ 0.769
CWS RotToro0.2	10 592	13 720	≈ 0.772
CWS RotToro0.3	10 604	13 720	≈ 0.773
CWS RotToro15Thrshld	10 584	13 720	≈ 0.771
CWS RotToro25Thrshld	10 598	13 720	≈ 0.772
CWS RotToro40Thrshld	10 606	13 720	≈ 0.773

Table 11: A table listing the test description, number of significant tests, total number of tests, and false positive rates (rounded to the third decimal place) for all of the tests considered in Figure 36.

Test Description	Sig. Tests	Total Tests	False Positive Rate
DWS Rot	170	3430	≈ 0.050
UWS Rot	170	3430	≈ 0.050
CWS Rot	178	3430	≈ 0.052
DWA Rot	176	3430	≈ 0.051
UWA Rot	177	3430	≈ 0.052
CWA Rot	181	3430	≈ 0.053
DWS Toro	135	2940	≈ 0.046
UWS Toro	138	2940	≈ 0.047
CWS Toro	137	2940	≈ 0.047
DWA Toro	136	2940	≈ 0.046
UWA Toro	142	2940	≈ 0.048
CWA Toro	131	2940	≈ 0.045
CWS RotToro0.1	173	3430	≈ 0.050
CWS RotToro0.2	190	3430	≈ 0.055
CWS RotToro0.3	179	3430	≈ 0.052
CWS RotToro15Thrshld	182	3430	≈ 0.053
CWS RotToro25Thrshld	177	3430	≈ 0.052
CWS RotToro40Thrshld	184	3430	≈ 0.054

the outside edge of the two samples. However, since the true distributions are not known in practice, it is still recommended to use the combined rotational and toroidal modifications test as it is more generally suited to other types of data.

In Figure 36, the false positive rate is computed by dividing the number of significant test results by the total number of tests computed when the null hypothesis is true and both samples come from the same distribution (first row of graphs in Figures 27–34). For false positive rates in Figure 36, test results should be as close as possible to 0.05 (i.e., 5%, indicated by the horizontal line) when testing at the 5% significance level. Test results which fall below 0.05 are indications of a conservative nature in the test (i.e., a test which is less likely to reject the null when it is actually true).

While the rotational test false positive rates seem to be relatively on target with the significance level of 0.05 across all of the six test statistics, the toroidal test false positive rates seem to be a little more conservative (i.e., all are slightly below the significance level). If only the CWS statistic is compared across the tests, the false positive rates of the combined rotational and toroidal tests seem to be closer to the rotational false positive rate (which is slightly above the significance level) than the toroidal false positive rate (which is slightly below the significance level). Overall, the false positive rates do not differ considerably across the six test statistics within either the rotational or toroidal tests.

Additionally, little trend is seen in the false positive rates of the combined rotational and toroidal tests regardless of the proportions of points used for the origins of toroidal shifts or the thresholds used for the number of toroidal shifts. In fact, the 0.1 proportion of points gives a false positive rate closest to the significance level (0.05) among the tests which use proportions of points. Similarly, the threshold of 25 toroidal shifts gives a false positive rates closest to the significance level among tests

which use toroidal shift thresholds. These results provide additional evidence that a 0.1 proportion of points and a 25 toroidal shift threshold are sufficient.

Overall, while these simulations show that there are special cases in which the rotational modification alone is the most powerful test among the modified Syrjala tests, the combined rotational and toroidal test is still recommended, and the test results confirm that there is little difference across the six statistics. Furthermore, the tests are shown to be stable across an array of number of rotations, toroidal shifts, or both (using either proportions of points or thresholds for the number of toroidal shifts) in the case of the combined modified Syrjala tests, even for sample sizes as little as 15 in each sample (which is more commonly seen in some eye-tracking applications such as the USU Posture Study (see Chapters 4 and 7]).

6.5.4 Simulating Differences in Fixation Shape

The null hypothesis for this simulation is similar to Section 6.5.3. However, here it is assumed that the subject is concentrating their visual attention on a single object in the center of the $[0,1] \times [0,1]$ square. This is modeled as a bivariate normal distribution with a mean coordinates of $(0.5, 0.5)$ and variance-covariance matrix of $\left[\begin{smallmatrix} 0.005 & 0 \\ 0 & 0.005 \end{smallmatrix}\right]$.

Departures from the null hypothesis are modeled as distributions centered at the same bivariate mean as the null distribution, but exhibiting departures from the null covariance structure. Specifically, four departures from the null hypothesis are modeled as bivariate normal distributions with mean coordinates as $(0.5, 0.5)$ and variance-covariance matrices as $\left[\begin{smallmatrix} 0.005 & 0.000625 \\ 0.000625 & 0.005 \end{smallmatrix}\right]$, $\left[\begin{smallmatrix} 0.005 & 0.00125 \\ 0.00125 & 0.005 \end{smallmatrix}\right]$, $\left[\begin{smallmatrix} 0.005 & 0.001875 \\ 0.001875 & 0.005 \end{smallmatrix}\right]$, and $\left[\begin{smallmatrix} 0.005 & 0.0025 \\ 0.0025 & 0.005 \end{smallmatrix}\right]$, respectively. These variance-covariance matrices were chosen in order to induce an approximate positive linear correlation between the horizontal and vertical variables equal to 0.125, 0.25, 0.375, and 0.5, respectively.

Results

Figure 37 also follows the general structure of result visualizations described in Section 6.5.2. The abbreviation "cen" in the far right row labels indicates that a bivariate normal distribution was used to generate the data in the center of the unit square. The label "cov" and number following cov indicates the approximate positive linear correlation introduced between the horizontal and vertical variables, e.g., cov_0.125 indicates that the second sample was generated using a bivariate normal distribution with a variance-covariance matrix of $\left[\begin{smallmatrix} 0.005 & 0.000625 \\ 0.000625 & 0.005 \end{smallmatrix}\right]$ (since a correlation of 0.125 produces a covariance of $0.125 \times \sqrt{0.005 \times 0.005} = 0.000625$). In the first row, 24 out of the 490 are significant. This is roughly 0.049 or 4.9%. This result also agrees with the conservative nature of the test as seen in previous simulations (see Sections 6.4.1 and 6.4.2).

The remaining four rows (when the null hypothesis is false) demonstrate an increasing relationship between the number of significant tests and both the sample size and magnitude of the change in covariance structure in the second sample's distribution. In one extreme, the far left line graph in the second row demonstrates that for a small departure in the null distribution (specifically, an increase in the off diagonal entries of the variance-covariance matrix by 0.125) and a small sample size for the first sample, there are only three significant tests across all of the second sample sizes. This is not much different from the subgraphs in the first row. This is also evident when observing all of the cases in which the second sample size is 15 within the second row. However, for sufficiently large sample sizes, i.e. when both sample sizes are larger than 250, the number of significant tests begins to depart from the behavior exhibited in the first two rows.

The third row of line graphs demonstrates that for larger departures from the null distribution (specifically, an increase in the off diagonal entries of the variance-

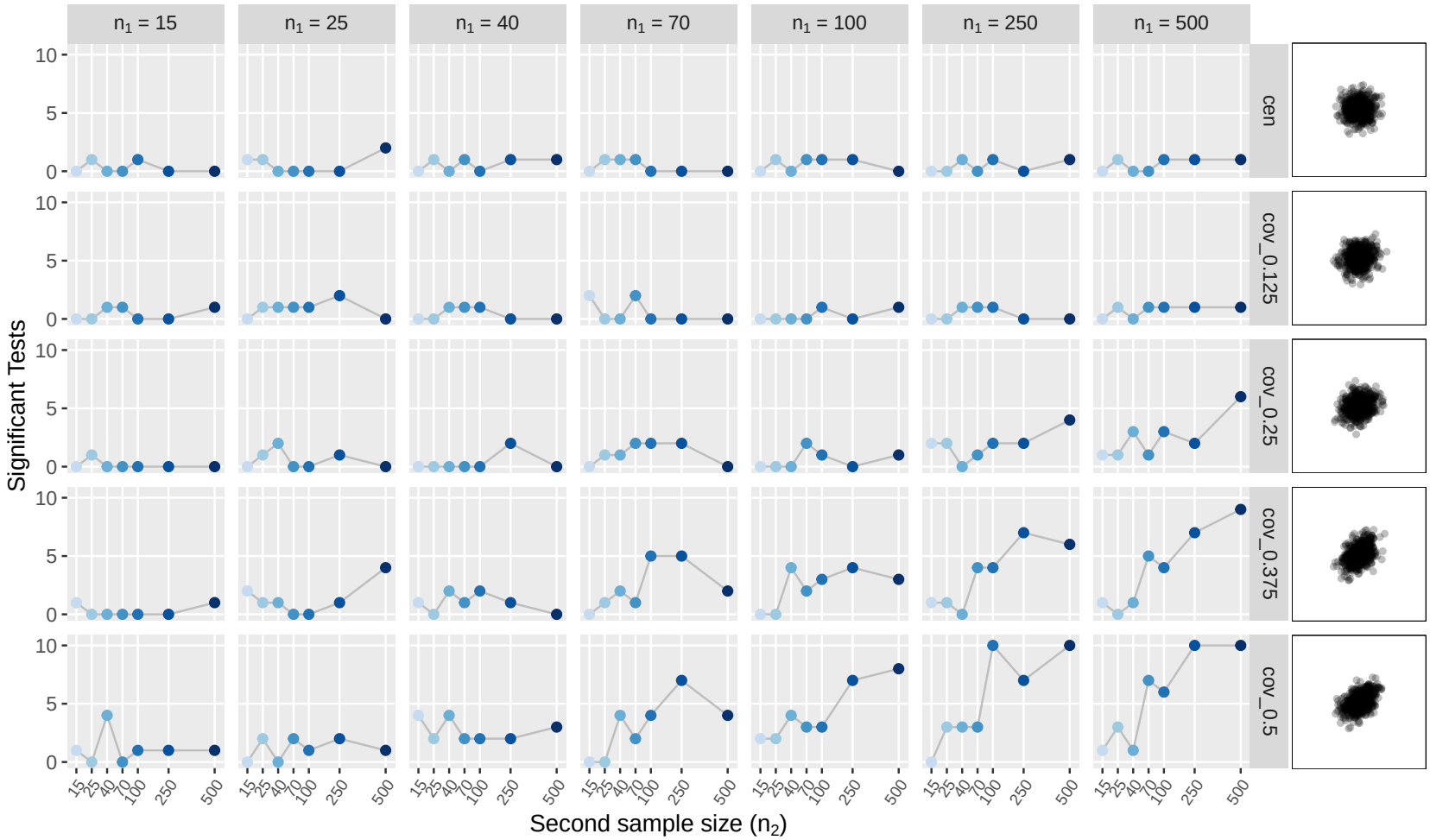


Fig. 37: A grid of line graphs showing the performance of the modified Syrjala test (using 0.1 proportion of points as origins of toroidal shifts, 8 rotations, and the CWS statistic) on simulated eye-tracking data where subjects concentrate on a single object with differing fixation shapes. The grid column name indicates the first sample size (n_1), and the grid row indicates the shape of the second sample. The shape of the first sample follows the same distribution exhibited in the first row. The horizontal axis indicates the second sample size (n_2). For example, the bottom left graph shows the number of significant tests (out of ten tests) between the first samples with 15 points and second samples with 15, 25, 40, 70, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

covariance matrix by 0.375), a majority of tests are significant for sample sizes larger than 250. In the fourth row, this effect is also seen but for samples sizes greater than 100 (except for the case when both sample sizes are 100).

6.5.5 Simulating Differences in Fixation Allocation

While the null hypothesis is modeled as a bivariate normal distribution with a mean coordinates of $(0.25, 0.5)$ and a variance-covariance matrix of $\left[\begin{smallmatrix} 0.005 & 0 \\ 0 & 0.005 \end{smallmatrix}\right]$, this simulation focuses on a case where a subject gradually splits their attention between two objects. Hence, the alternative distribution is modeled as a mixture of two bivariate normal distributions with identical variance-covariance matrices $\left[\begin{smallmatrix} 0.005 & 0 \\ 0 & 0.005 \end{smallmatrix}\right]$, and with one centered at $(0.25, 0.5)$ (similar to the null hypothesis) while the second is centered at $(0.75, 0.5)$. Consequently, the second sample (for the alternative hypothesis) is split between the two bivariate normal distributions in the second sample using a binomial distribution. Four departures from the null are modeled this way such that each exhibits an increased amount of attention being given to the right object (while still maintaining most attention on the left object).

Specifically, let n_2 be the sample size of the second sample with n_{2a} and n_{2b} being the integer subsample sizes belonging to the left and right individual distributions, respectively, that make up the bimodal mixture distribution for the second sample. Then the subsample of data which belongs to the right distribution is modeled as $n_{2b} \sim \text{binom}(n_2, p_i)$, such that $p_i = i * 0.05$ for i^{th} departure from the null hypothesis where $i = 1, 2, 3, 4$. Hence, the subsample of data which belongs to the left distribution can simply be computed as $n_{2a} = n_2 - n_{2b}$.

Results

Here, Figure 38 also displays a grid of line graphs as described in Section 6.5.2.

The abbreviation "cen_l" in the far right row labels indicates that a bivariate normal distribution was used to generate the data on the left side of the unit square. The label "cen_spt_" and subsequent number indicate the proportion of points in the second sample being allocated to the bivariate normal distribution on the right side of the unit square, e.g., cen_spt_0.20 indicates that approximately 20% of the second sample size is being allocated to the right bivariate normal distribution while the remaining 80% will be allocated to the left distribution. The first row shows 24 out of the 490 test results are significant. This is roughly 0.049 or 4.9%. This result also agrees with the conservative nature of the test as seen in previous simulations (see Sections 6.4.1 and 6.4.2).

The remaining four rows (when the null hypothesis is false) demonstrate an increasing relationship between the number of significant tests and both the sample size and proportion of points which move from the left object to the right object in the second sample's distribution. In one extreme, the far left line graph in the second row (from the top) demonstrates that for a small departure in the null distribution (specifically, assigning a binomial probability of 0.05 to the right distribution) and a small sample size for the first sample, there is only one significant tests across all of the second sample sizes. This is not much different from the subgraphs in the first row. Similarly, the remaining subgraphs in the second row for larger sample n_1 sample sizes show that there is not enough of a difference between the two samples for the test to label many of the tests as significant. However, the third row of tests which employ a binomial probability of 0.10, when both sample sizes are greater than 250 a majority of the test results are labeled as significant. In the third row, this effect is also seen but for samples sizes greater than 100, and for only sample sizes of greater than 70 in the forth row.

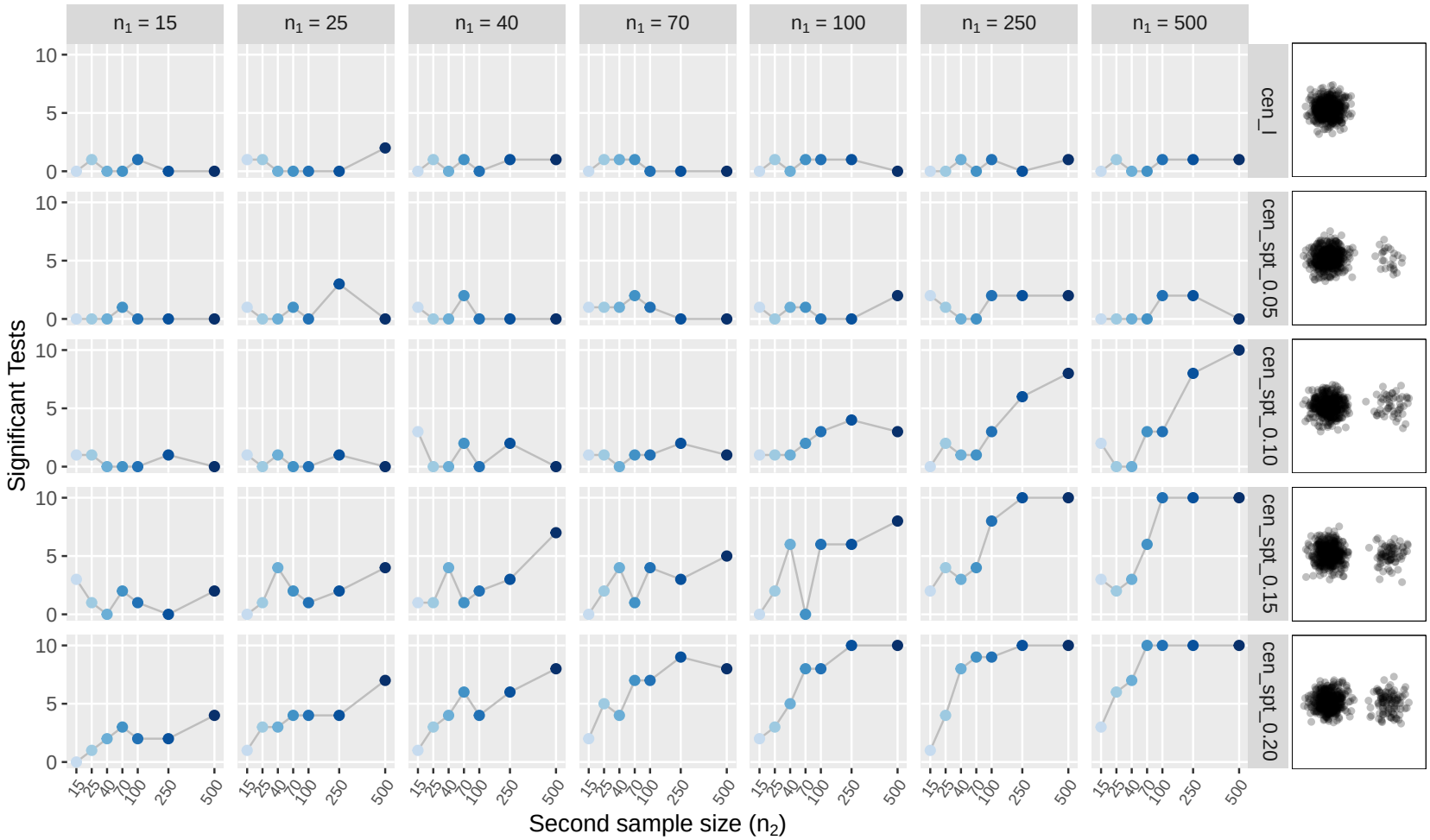


Fig. 38: A grid of line graphs showing the performance of the modified Syrjala test (using 0.1 proportion of points as origins of toroidal shifts, 8 rotations, and the CWS statistic) on simulated eye-tracking data where subjects concentrate on a single object at differing locations. The grid column name indicates the first sample size (n_1), and the grid row indicates the shape of the second sample. The shape of the first sample follows the same distribution exhibited in the first row. The horizontal axis indicates the second sample size (n_2). For example, the bottom left graph shows the number of significant tests (out of ten tests) between the first samples with 15 points and second samples with 15, 25, 40, 70, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

6.5.6 Simulating the Introduction of a Single Outlier

The null hypothesis for this simulation is identical to Section 6.5.5 (i.e., it is modeled as a bivariate normal distribution with a mean coordinates as $(0.25, 0.5)$ and variance-covariance matrix of $\left[\begin{smallmatrix} 0.005 & 0 \\ 0 & 0.005 \end{smallmatrix}\right]$).

Four alternative hypothesis are modeled using the same distribution except that a single outlier gaze point is introduced at increasing distances to the right of the distribution. The coordinates of the outliers for the four departures are $(0.375, 0.5)$, $(0.5, 0.5)$, $(0.625, 0.5)$, and $(0.75, 0.5)$, respectively.

Results

Figure 39 follows the general structure of result visualizations described in Section 6.5.2. The abbreviation "mvnorm" in the far right row labels indicates that a bivariate normal distribution was used to generate the data on the left side of the unit square. The label "pnt_" and subsequent number indicate the horizontal coordinates of the single outlier being generated, e.g., pnt_0.75 that a single outlier is being included in the second sample with coordinates (0.5, 0.75). The first row shows 24 out of the 490 test results are significant. This is roughly 0.049 or 4.9%. This result also agrees with the conservative nature of the test as seen in previous simulations (see Sections 6.4.1 and 6.4.2).

However, while the four bottom rows in Figure 39 compare the effect of a single outlier which strays further from the null distribution, the results are the same regardless of the distance of the outlier from the null distribution. Namely, the number of significant tests remains around the significance level (of 0.05) regardless of sample sizes or magnitude of the departure of the outlier from the null distribution.

This result makes sense. Since the modified Syrjala tests compare differences between the empirical cumulative distribution functions, the magnitude of space be-

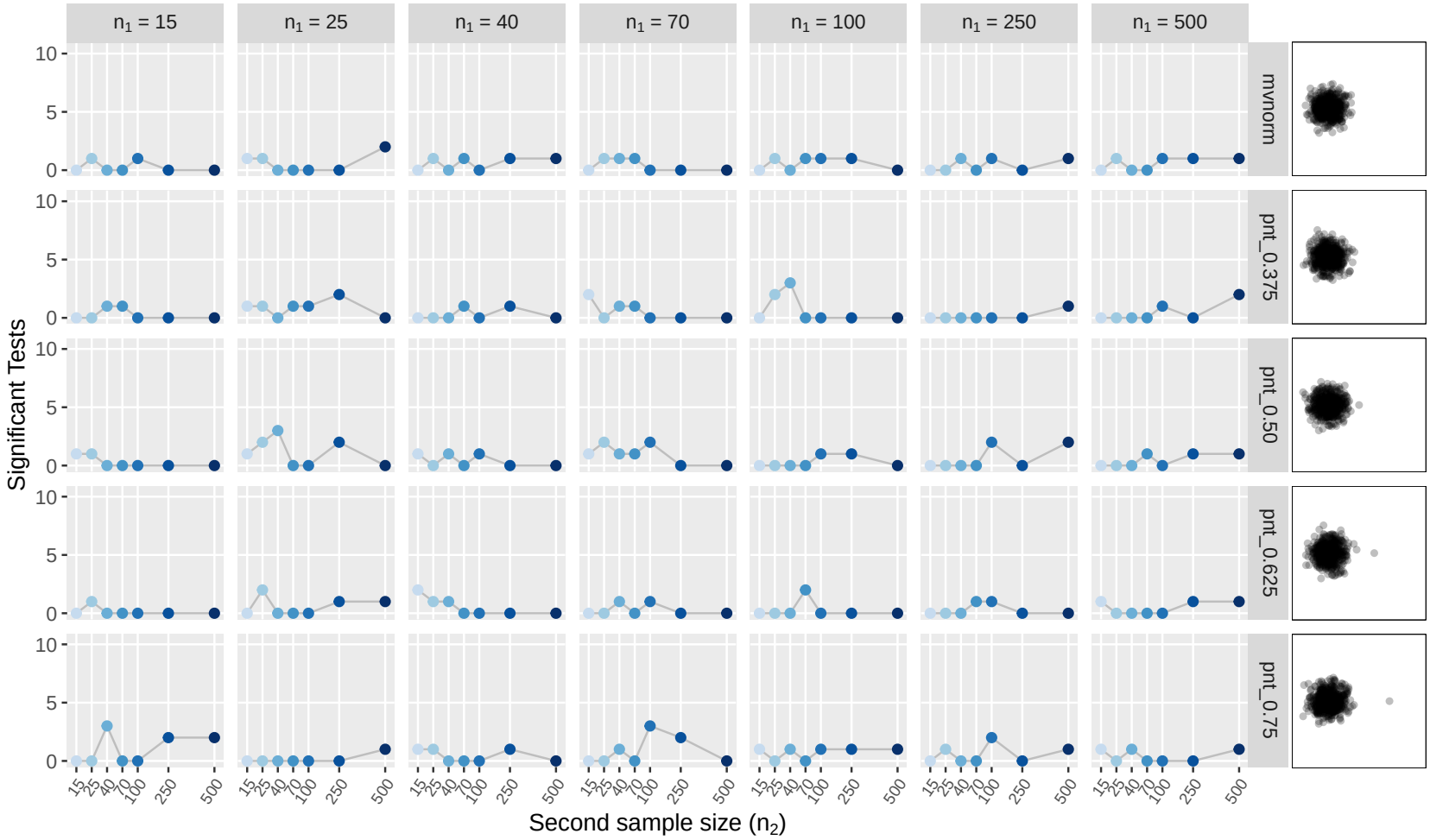


Fig. 39: A grid of line graphs showing the performance of the modified Syrjala test (using 0.1 proportion of points as origins of toroidal shifts, 8 rotations, and the CWS statistic) on simulated eye-tracking data where subjects concentrate on a single object across cases where a single straying outlier is also observed. The grid column name indicates the first sample size (n_1), and the grid row indicates the shape of the second sample. The shape of the first sample follows the same distribution exhibited in the first row. The horizontal axis indicates the second sample size (n_2). For example, the bottom left graph shows the number of significant tests (out of ten tests) between the first samples with 15 points and second samples with 15, 25, 40, 70, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

tween points is not so important as the bivariate ordering of the points between the two samples. Hence, since moving the outlier further from the null distribution does not change the ranking of the outlier there is not enough departure from the null distribution for the modified Syrjala test to label a proportion of the tests as significant much more than the significance level, i.e., the number of significant tests in the bottom four rows is similar that that described in the first row on average.

6.5.7 Simulating the Introductions of Many Outliers in a Single Sample

This simulation is related to the previous simulation discussed in Section 6.5.6, except that instead of introducing only one fixed outlier into the second sample many random outliers are introduced. The null distribution is identical (i.e., it is modeled as a bivariate normal distribution with a mean coordinates as $(0.25, 0.5)$ and variance-covariance matrix of $\left[\begin{smallmatrix} 0.005 & 0 \\ 0 & 0.005 \end{smallmatrix}\right]$). However, the four alternative hypothesis are modeled using mixture distributions which contain both the null distribution and a distribution for the random outliers. Specifically, the random outliers are modeled using a inhomogeneous Poisson point process with an intensity function, f_o, of

$$f_o(x, y) = 790 \cdot \Big(1 - \exp\Big\{ -80 \cdot \big[(x - 0.5)^4 + (y - 0.5)^4\big]\Big\}\Big).$$

This is similar to the intensity function used to generate the Repel distribution described in Section 6.1.1. This distribution was chosen to simulate the creation of outliers by undesirable subject interactions with the eye-tracking device, e.g., blinking. The four departures from the null distribution exhibit an increasing number of outliers, specifically, 2, 4, 8, and 16 outlier points, respectively.

Results

Figure 40 also displays a grid of line graphs as described in Section 6.5.2. However, the sample sizes listed in the figure do not include the added outliers included in the departures from the initial distribution. As an example, the bottom left graph shows the number of significant tests (out of ten tests) between the first samples with 15 points and second samples (including outliers) with 31, 41, 56, 86, 116, 366, and 516 points (or n_2 = 15, 25, 40, 70, 100, 250, and 500 points when outliers are not counted). The abbreviation "mvnorm" in the far right row labels indicates that a bivariate normal distribution was used to generate the data in the center of the unit square. The label "nse_" and subsequent number indicate the number of additional outliers being generated, e.g., nse_16 indicates that 16 outliers are being generated in addition to a bivariate normal distribution.

In the first row, 24 out of the 490 test results are significant. This is roughly 0.049 or 4.9%. This result also agrees with the conservative nature of the test as seen in previous simulations (see Sections 6.4.1 and 6.4.2).

The remaining four rows (when the null hypothesis is false) demonstrate a positive association between the number of significant tests and the number of outliers in addition to a negative association between the number of significant tests and the sample sizes. When only two outliers are present (as seen in the second row of subgraphs from the top), the performance of the test is similar to the null case (in the first row). This makes sense given the results of the simulation of discussed in Section 6.5.6. Little effect is also seen (in the third row) by four introduced outliers except for a few cases, e.g., the four significant tests when $n_1 = 250$ and $n_2 = 15$.

The combined effect of sample size with number of outliers begins to be evident in the forth row of line graphs. Note that for smaller samples sizes in the second sample eight outliers can represent a large proportion of the overall distribution, i.e.,

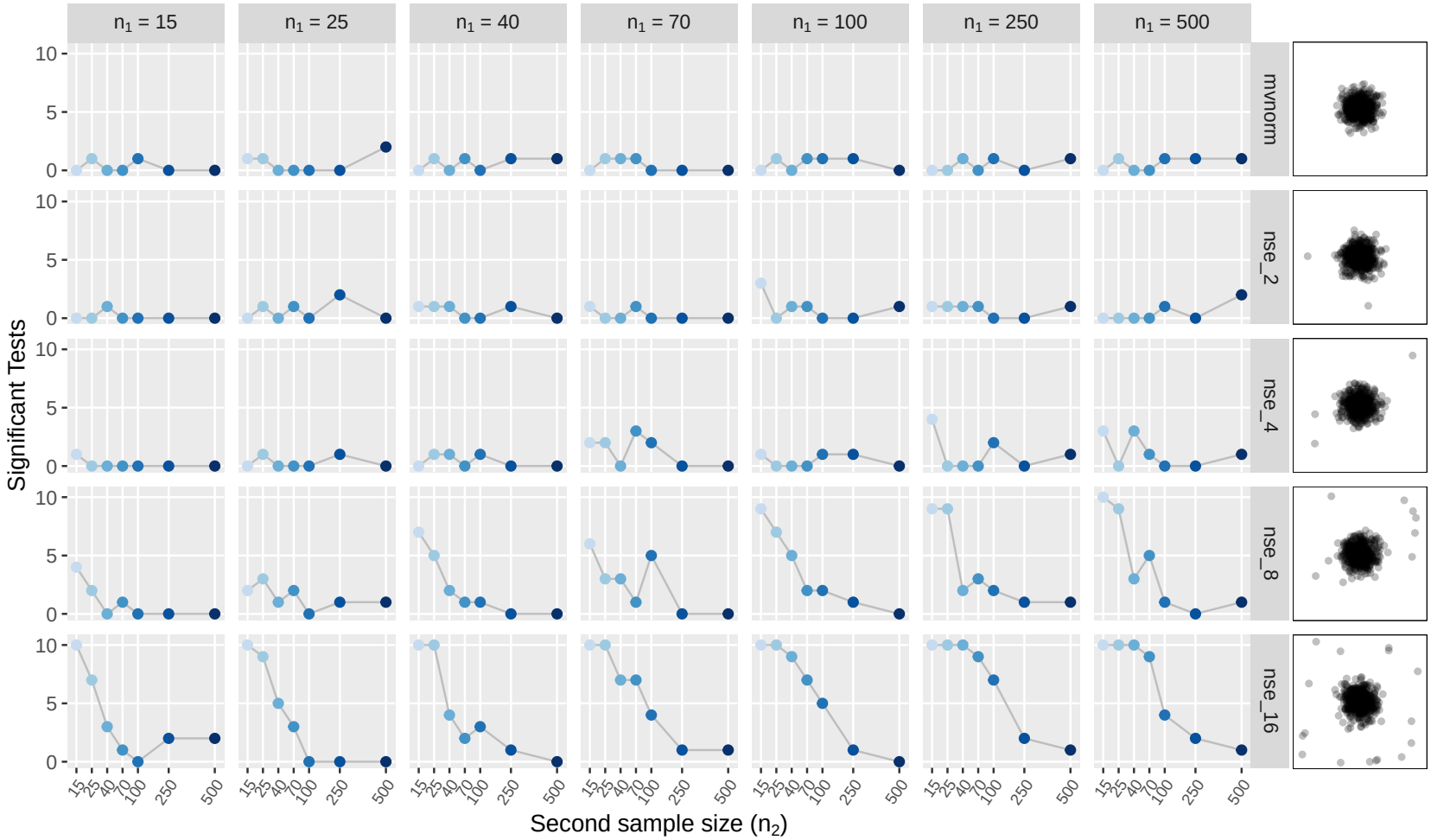


Fig. 40: A grid of line graphs showing the performance of the modified Syrjala test (using 0.1 proportion of points as origins of toroidal shifts, 8 rotations, and the CWS statistic) on simulated eye-tracking data where subjects concentrate on a single object while the second subject also exhibits increasing amounts of noise. The grid column name indicates the first sample size (n_1), and the grid row indicates the shape of the second sample. The shape of the first sample follows the same distribution exhibited in the first row. The horizontal axis indicates the second sample size (n_2). For example, the bottom left graph shows the number of significant tests (out of ten tests) between the first samples with 15 points and second samples with 15, 25, 40, 70, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

when $n_2 = 15$ the total sample size including the 16 outliers is 31, and the proportion of outliers in the sample is approximately 0.52. Hence, when the first sample size is greater than 100, a majority of tests are significant for $n_2 \leq 25$. In other words, when a relatively large n_1 provides a clear picture of what the null distribution is, eight outliers make up a large enough component of the second sample for small n_2 to trigger many significant results among the tests.

However, this effect is less pronounced as n_2 grows. When $n_2 \geq 40$ eight outliers represent less than 17% of the mixture distribution resulting in less than half of the tests being significant in most cases (a few are exactly half of the tests). This overall relationship between significant tests, sample sizes, and the number of outliers is only emphasized further in the bottom row of graphs.

6.5.8 Simulating the Introductions of Many Outliers in Both Samples

This simulation is closely related to the previous simulation discussed in Section 6.5.7, except that instead of introducing many outliers into the second sample alone many random outliers are introduced within both samples. The initial distribution remains the same (i.e., it is modeled as a bivariate normal distribution with a mean coordinates as $(0.25, 0.5)$ and variance-covariance matrix of $\left[\begin{smallmatrix} 0.005 & 0 \\ 0 & 0.005 \end{smallmatrix}\right]$). However, the four mixture distributions which introduce outliers are drawn from when creating the two independent samples (instead of the first sample always being drawn from the initial distribution). Hence, all comparisons within this simulation only consider cases for which the null hypothesis is true, and five null hypotheses are being tested:

The random outliers are still modeled using a inhomogeneous Poisson point process with an intensity function, f_o, of

$$f_o(x, y) = 790 \cdot \Big(1 - \exp\Big\{ -80 \cdot \big[(x - 0.5)^4 + (y - 0.5)^4\big]\Big\}\Big),$$

similar to the intensity function used to generate the Repel distribution described in Section 6.1.1. This simulation compares the performance of the modified Syrjala test when assuming two subjects are focusing on a single object while also exhibiting a similar number of outliers during data collection. The four numbers of outliers are 2, 4, 8, and 16 outlier points, respectively.

Results

Figure 41 shows the results of the simulation. Similar to the previous figures, this figure displays a grid of line graphs showing the performance of the modified Syrjala test (using 0.1 proportion of points as origins of toroidal shifts, 8 rotations, and the CWS statistic) for this simulation. The grid column name indicates the first sample size (n_1), horizontal axes indicate the second sample size (n_2). Note that neither n_1 nor n_2 include the introduced outliers, but only represent the number of points used to generate the bivariate normal distributions depicted in the center of the unit square. Additionally, unlike the figures in Sections 6.5.3–6.5.7, the grid row indicates the shape of both the first and second samples. The abbreviation "mvnorm" in the far right row labels indicates that a bivariate normal distribution was used to generate the data in the center of the unit square. The label "nse_" and subsequent number indicate the number of additional outliers being generated in both samples, e.g., nse_16 indicates that 16 outliers are being generated in addition to a bivariate normal distribution for both samples. Hence, while the sample sizes vary across the

columns and horizontal axes, each row demonstrates the performance of the test when the null hypothesis is true. As an example, the bottom left graph shows the number of significant tests (out of ten tests) between the first samples with 15 points (and 16 outliers) and second samples (including outliers) with 31, 41, 56, 86, 116, 366, and 516 points (or $n_2 = 15$, 25, 40, 70, 100, 250, and 500 points when outliers are not counted).

Overall, this simulation shows that when the proportion of outliers is about the same between samples, the majority of tests indicate non-significance. Additionally, when the proportion of outliers is relatively small (i.e., when the number of outliers is less than four), the effect of outliers is negligible across all sample sizes, and a majority of all tests are non-significant. This latter result can be seen in the top three rows of subgraphs in Figure 41.

The former result is most noticeable in the bottom two rows of subgraphs. Notice that for sample sizes relatively close to each other, the proportion of outliers within the samples are relatively similar. Consequently, a majority of tests are non-significant. In the other extreme, when the sample sizes are quite different, then the proportion of outliers within the samples differ considerably. This results in a majority of tests being significant. For example, in the bottom row of subgraphs where both samples have 16 outliers, when $n_1 = 70$ most of the tests are non-significant for $25 \leq n_2 \leq 250$. However, for $n_2 = 15$ or $n_2 = 500$, the proportion of outliers to the sample sizes are approximately 0.52 and 0.03, respectively. This results in a majority of the tests being significant.

6.5.9 Simulating Differences in Fixation Location, Shape, and Outliers within the USU Posture Study Data

While the simulations discussed in Sections 6.5.3–6.5.8 show the performance of

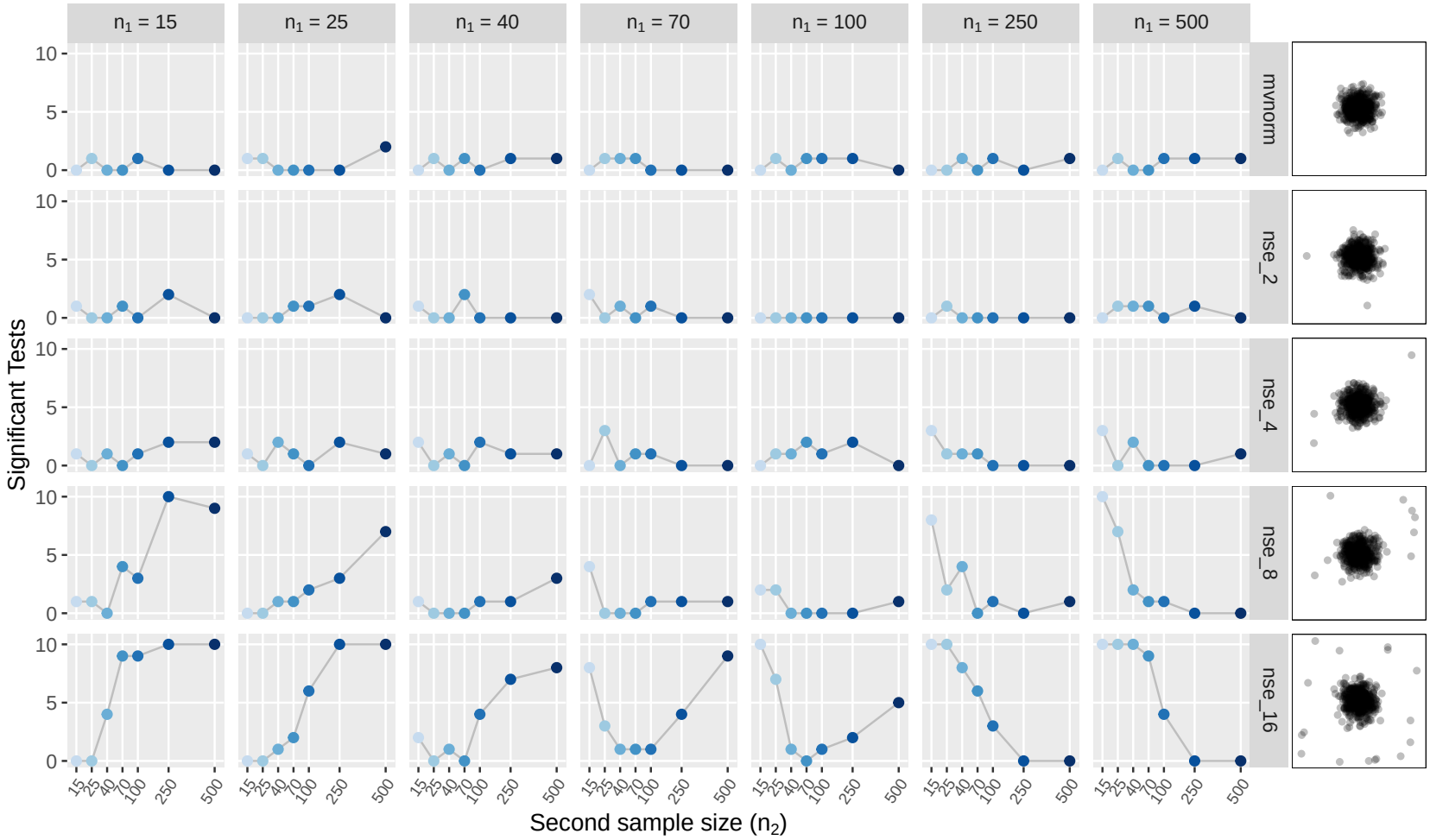


Fig. 41: A grid of line graphs showing the performance of the modified Syrjala test (using 0.1 proportion of points as origins of toroidal shifts, 8 rotations, and the CWS statistic) on simulated eye-tracking data where subjects concentrate on a single object while both subjects also exhibit increasing amounts of noise. The grid column name indicates the first sample size (n_1), and the grid row indicates the shape of the second sample. The shape of the first sample follows the same distribution exhibited in the first row. The horizontal axis indicates the second sample size (n_2). For example, the bottom left graph shows the number of significant tests (out of ten tests) between the first samples with 15 points and second samples with 15, 25, 40, 70, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

the modified Syrjala tests on generated data which are more closely representative of eye-tracking realizations, the scenarios modeled are simple and only meant to show the performance of the test for isolated incremental changes between the samples. The simulation in this section, however, generates data which models more complicated real-life data taken from the USU Posture Study (see Chapter 4 for more details).

However, this simulation builds upon those discussed in Sections 6.5.3–6.5.8 by demonstrating how changes in fixation location, fixation shape, and the introduction of outliers (simulated individually in Sections 6.5.3, 6.5.4, and 6.5.7, respectively) have upon the performance of the test. The impact of changes to all three simultaneously is also considered. For a closer look at the affect of changes in proportions of points allocated between the various gaze point clusters (simulated on a simpler lever in Section 6.5.5), see Section 6.5.10.

Consequently, the comparison between the aggregated gaze points from the treatment and control groups for posture ID 17 within the USU Posture Study was selected as a basis from which to construct the generated data for this simulation. Figure 42 compares the scatterplots of the gaze points from these two groups side-by-side. This data was selected to show how a collective change in fixation location, fixation shape, and introduction of additional outliers contribute to the significant differences in gaze point distributions between these two groups.

To generate data similar to that of the treatment and control groups, eleven components were identified as contributors to the mixture distributions, namely, component distributions which model the clusters located at the crown of the head, center of the neck, surrounding neck region, right shoulder, surrounding right shoulder region, between the knees and thighs, right foot, surrounding right foot region, left foot, between the feet, as well as a distribution modeling the outliers caused by the eye-tracking equipment occasionally incorrectly assigning gaze point locations. These

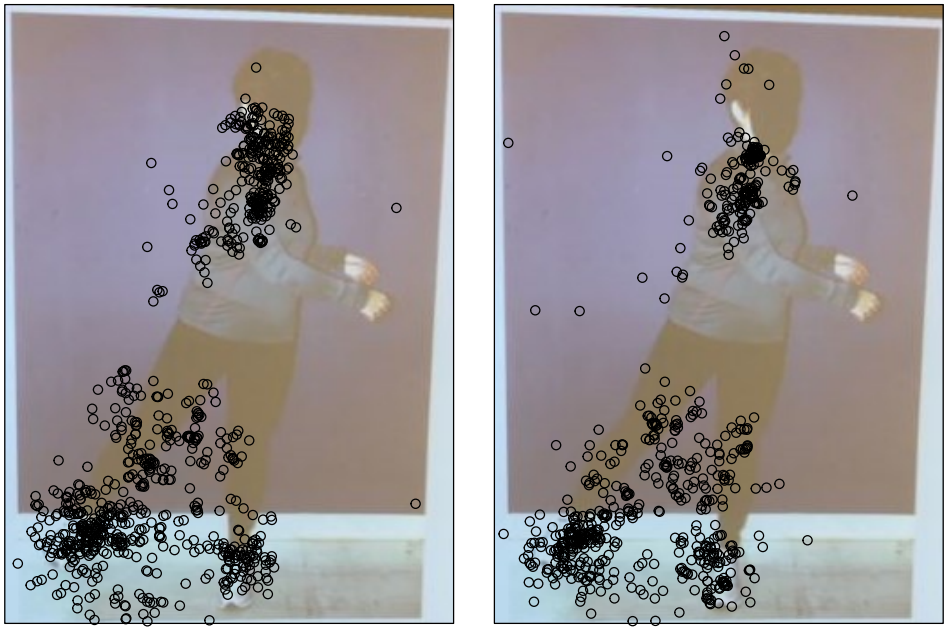


Fig. 42: Scatterplots of the aggregated gaze points for the treatment (left) and control (right) groups for posture ID 17 taken from the USU Posture Study (Symanzik et al., 2017, 2018; Studenka et al., 2020; Coltrin et al., 2020; McKinney and Symanzik, 2019, 2021).

clusters are referred to using the following abbreviations (and are listed in the same order as the brief cluster descriptions given above in the same paragraph): head, neck, sur. neck, shoulder, sur. shoulder, legs, r. foot, sur. r. foot, l. foot, b. feet, and noise.

From here, a null distribution was constructed in addition to four alternative distributions which depict departures from the null distribution similar to the differences seen between the two scatterplots in Figure 42. The individual component distributions were modeled using bivariate normal distributions except for the noise distribution which was a bivariate uniform distribution (that spans the unit square). The bivariate normal distributions were constructed using the bivariate means of the clusters as well as the two largest directions of variability within the clusters.

For example, after the data was rescaled to fit within the unit square, the center

of the legs cluster of the control group (left scatterplot in Figure 42) was identified as the coordinates $\mu_c = (0.54, 0.29)$. The two largest directions of variability are vectors which start from the center and point to the approximate 95^{th} bivariate percentile of the cluster located at $p_1 = (0.80, 0.20)$ and $p_2 = (0.63, 0.39)$, respectively. Hence the vectors are computed as $v_i = p_i - \mu_c$ (for $i = 1, 2$). The variance-covariance matrix can then be computed as follows: Let v_i be the largest two eigenvectors belonging to the eigen-decomposition of the variance-covariance matrix of the cluster ($\mathbf{\Sigma}$). Then the the corresponding eigenvalues can be computed as the Euclidean norm of the components of the eigenvectors: $\lambda_i = ||v_i||^2$. Then let $\mathbf{\Lambda} = \left[\begin{smallmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{smallmatrix}\right]$, and $\mathbf{E} = [v_1^T, v_2^T]$ such that T is the transpose operator. Using the eigen-decomposition of $\mathbf{\Sigma}$, $\mathbf{\Sigma}$ can be factorized as $\mathbf{\Sigma} = \mathbf{E\Lambda E}^{-1}$, where $\mathbf{E}^{-1}$ is the inverse of $\mathbf{E}$. Hence, for this example, $\mathbf{\Sigma} \approx \left[\begin{smallmatrix} 0.0196 & -0.0052 \\ -0.0052 & 0.0035 \end{smallmatrix}\right]$.

As a mixture distribution, the sample must be randomly split into eleven subsamples. Using the control groups aggregated data as a baseline, subsample proportions were identified for each of the component clusters. Using these subsample proportions, realizations from a multinomial distribution were drawn to allocate the sample sizes randomly to the different component distributions. The sample size was assigned as the number of trials for the multinomial distribution, while the subsample proportions were assigned as the probabilities of success for each of the multinomial events. The probabilities are listed in Table 12.

While the null distribution is patterned closely to the aggregated gaze points of the control group, the three departures from the null are expressed as (1) changes in the center of the legs distribution from $(0.54, 0.29)$ to $(0.41, 0.34)$, (2) changes in the covariance structure for the head, neck, sur. shoulder, and sur. r. foot distributions (see Table 13), (3) an increase in the multinomial probability of success for the event associated with the noise distribution from 0.004 to 0.0178. When

Table 12: A table of multinomial event success probabilities for the mixture distribution subsample allocation.

Cluster Name	Multinomial Probability
head	0.0118
neck	0.0811
sur. neck	0.0295
shoulder	0.0959
sur. shoulder	0.1032
legs	0.2065
r. foot	0.1032
sur. r. foot	0.2065
l. foot	0.1180
b. feet	0.0265
noise	0.0040

increasing the probability of success for the event associated with the noise distribution, the remaining probabilities in the multinomial distribution are decreased by $(0.0178 - 0.004)/10 = 0.00138$ in order to maintain a proper sum to 1 in the multinomial event probabilities. A combination of all three departures from the null distribution is also considered, and is patterned after the distribution of the aggregated gaze points for the treatment group.

Results Across All Modified Syrjala Tests

In Section 6.5.3 the modified Syrjala tests were shown to exhibit stable results across an array of rotations, toroidal shifts, and both rotations and toroidal shifts in eye-tracking inspired simulations which involved departures from bivariate normal data. It was also shown that little difference in the test results are observed across six different statistics within the modified Syrjala tests. Consequently, the combined rotational and toroidal shift modifications with the CWS statistic are employed again to establish stable results across the smaller sample sizes using more complex eye-tracking inspired data, and to reaffirm the sensible default parameter values (initially

Table 13: A table of approximate null and alternative variance-covariance matrices (rounded to the second decimal place) for bivariate normal distributions used to generate synthetic gaze point clusters. Note that while the noise cluster (not listed in this table) used the same random bivariate uniform distribution for both the null and alternative hypotheses, the multinomial probability assigned to the outliers subsample is 0.004 and 0.0178 for the null and alternative distributions, respectively. Additionally, while changes where exhibited in the covariance structures for the head, neck, sur. shoulder, and sur. r. foot clusters, the remaining variance-covariance matrices are the same between the null and alternative distributions.

Cluster Name	Null Covariance	Alternative Covariance
head	$\begin{bmatrix} 0.11 & 0 \\ 0 & 0.06 \end{bmatrix}$	$\begin{bmatrix} 0.11 & 0 \\ 0 & 0.15 \end{bmatrix}$
neck	$\begin{bmatrix} 0.06 & 0 \\ 0 & 0.05 \end{bmatrix}$	$\begin{bmatrix} 0.08 & 0 \\ 0 & 0.12 \end{bmatrix}$
sur. neck	$\begin{bmatrix} 0.12 & -0.04 \\ -0.02 & 0.05 \end{bmatrix}$	$\begin{bmatrix} 0.12 & -0.04 \\ -0.02 & 0.05 \end{bmatrix}$
shoulder	$\begin{bmatrix} 0.07 & 0 \\ 0 & 0.07 \end{bmatrix}$	$\begin{bmatrix} 0.07 & 0 \\ 0 & 0.07 \end{bmatrix}$
sur. shoulder	$\begin{bmatrix} 0.12 & 0 \\ 0 & 0.12 \end{bmatrix}$	$\begin{bmatrix} 0.13 & 0.03 \\ 0.02 & 0.11 \end{bmatrix}$
legs	$\begin{bmatrix} 0.24 & -0.04 \\ -0.1 & 0.17 \end{bmatrix}$	$\begin{bmatrix} 0.24 & -0.04 \\ -0.1 & 0.17 \end{bmatrix}$
r. foot	$\begin{bmatrix} 0.1 & 0.01 \\ 0.03 & 0.07 \end{bmatrix}$	$\begin{bmatrix} 0.1 & 0.01 \\ 0.03 & 0.07 \end{bmatrix}$
sur. r. foot	$\begin{bmatrix} 0.19 & 0.03 \\ 0.03 & 0.13 \end{bmatrix}$	$\begin{bmatrix} 0.2 & 0.01 \\ 0.01 & 0.18 \end{bmatrix}$
l. foot	$\begin{bmatrix} 0.14 & -0.03 \\ -0.04 & 0.13 \end{bmatrix}$	$\begin{bmatrix} 0.14 & -0.03 \\ -0.04 & 0.13 \end{bmatrix}$
b. feet	$\begin{bmatrix} 0.11 & 0 \\ -0.01 & 0.1 \end{bmatrix}$	$\begin{bmatrix} 0.11 & 0 \\ -0.01 & 0.1 \end{bmatrix}$

proposed from the simulations with larger samples in Section 6.4.2) for the `distdiffR` `R` package functions (described in more detail in Chapter 8). Only the results of tests using the CWS statistic are shown in this section for consistency.

Consequently, the format of Figures 43–48 differ slightly from the others within Sections 6.3–6.4.2. While these figures still display grids of line graphs making comparisons between simulated differences in aggregated gaze point distributions from the treatment and control groups for posture ID 17 within the USU Posture Study (detailed in Section 6.5.9) for the same sample sizes (15, 25, 40, 70, 100, 250, and 500), the horizontal axes differs from that described in Section 6.5.2. The horizontal axes show the number of rotations (similar to the simulation Figure 18 in Section 6.3).

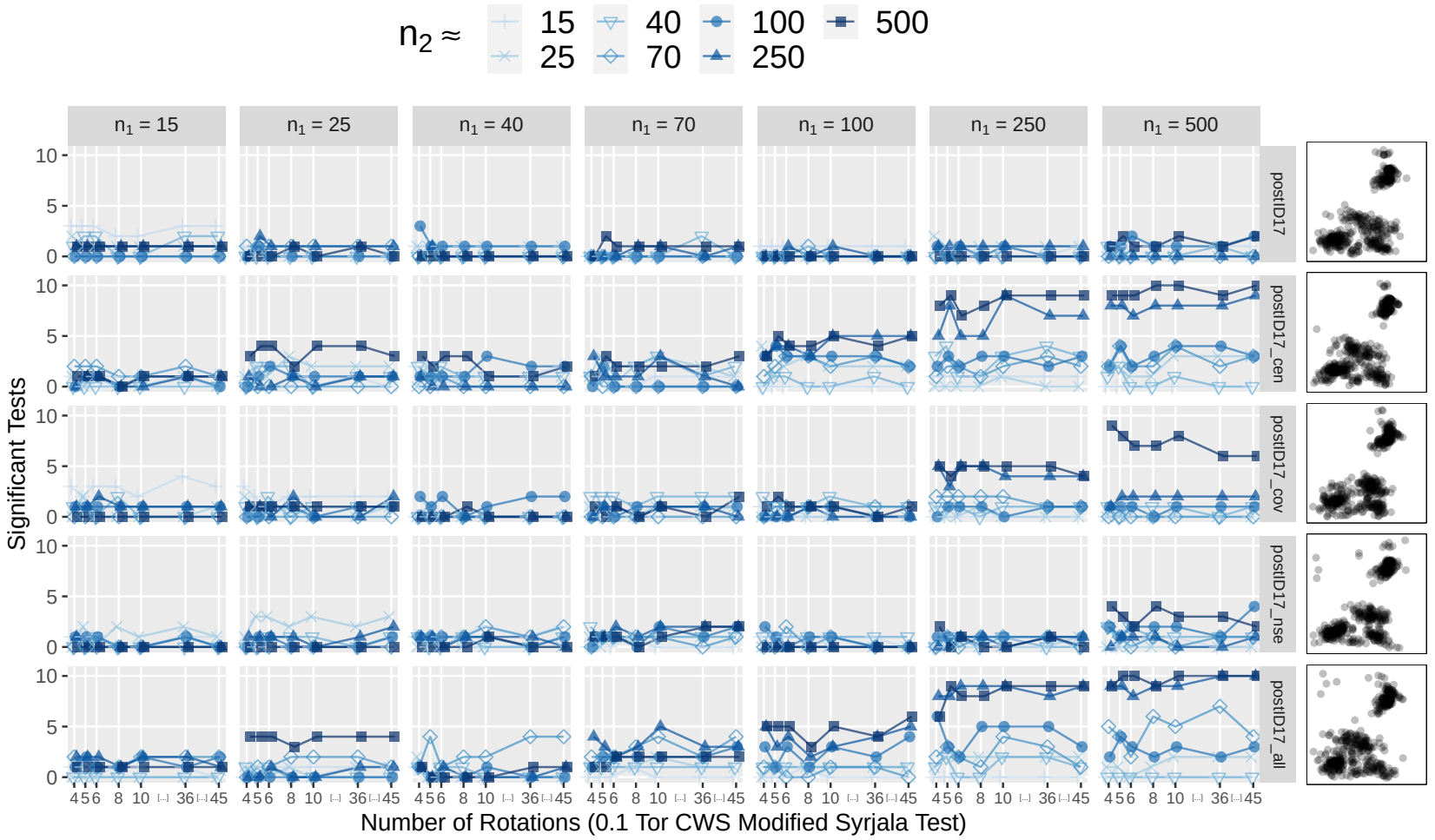


Fig. 43: A grid of line graphs showing the performance of the modified Syrjala test (using 0.1 proportion of points as origins of toroidal shifts, 8 rotations, and the CWS statistic) on simulated eye-tracking data which is similar to the aggregated subject contributions to Posture ID 17 (see Appendix C). The grid column name indicates the first sample size (n_1), and the grid row indicates the shape of the second sample. The shape of the first sample follows the same distribution exhibited in the first row. The horizontal axis indicates the second sample size (n_2). For example, the bottom left graph shows the number of significant tests (out of ten tests) between the first samples with 15 points and second samples with 15, 25, 40, 70, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

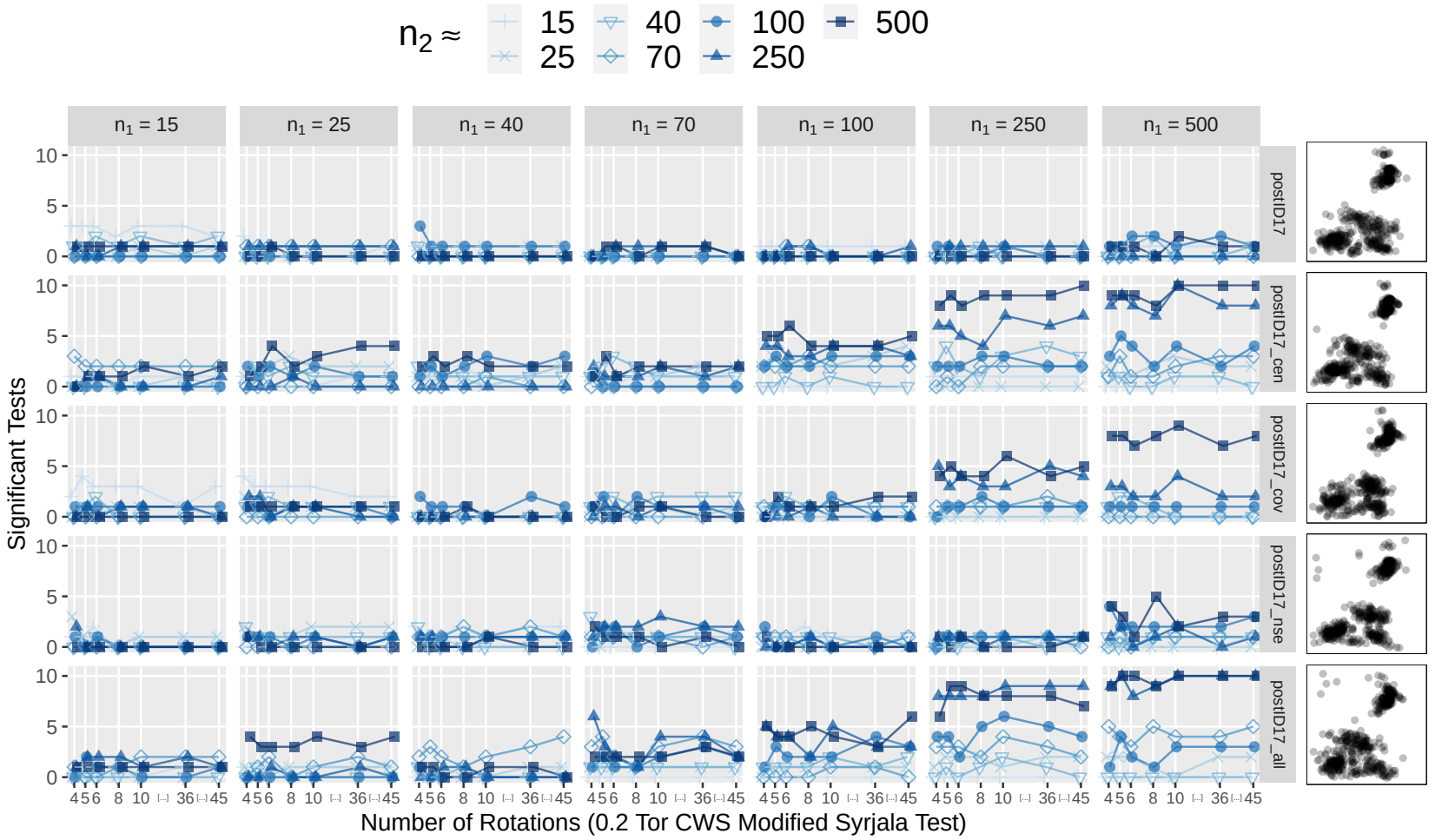


Fig. 44: A grid of line graphs showing the performance of the modified Syrjala test (using 0.2 proportion of points as origins of toroidal shifts, 8 rotations, and the CWS statistic) on simulated eye-tracking data which is similar to the aggregated subject contributions to Posture ID 17 (see Appendix C). The grid column name indicates the first sample size (n_1), and the grid row indicates the shape of the second sample. The shape of the first sample follows the same distribution exhibited in the first row. The horizontal axis indicates the second sample size (n_2). For example, the bottom left graph shows the number of significant tests (out of ten tests) between the first samples with 15 points and second samples with 15, 25, 40, 70, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

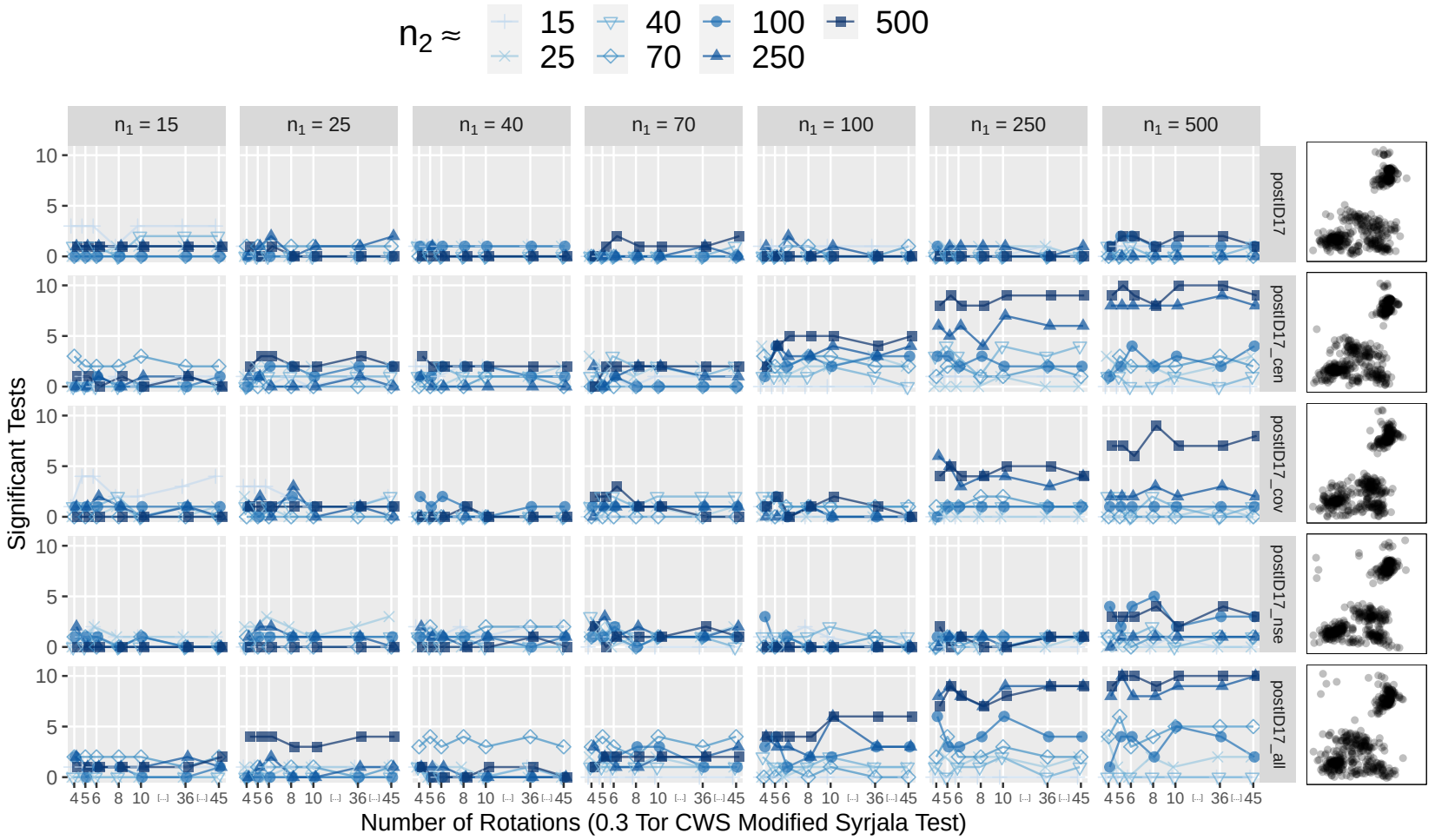


Fig. 45: A grid of line graphs showing the performance of the modified Syrjala test (using 0.3 proportion of points as origins of toroidal shifts, 8 rotations, and the CWS statistic) on simulated eye-tracking data which is similar to the aggregated subject contributions to Posture ID 17 (see Appendix C). The grid column name indicates the first sample size (n_1), and the grid row indicates the shape of the second sample. The shape of the first sample follows the same distribution exhibited in the first row. The horizontal axis indicates the second sample size (n_2). For example, the bottom left graph shows the number of significant tests (out of ten tests) between the first samples with 15 points and second samples with 15, 25, 40, 70, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

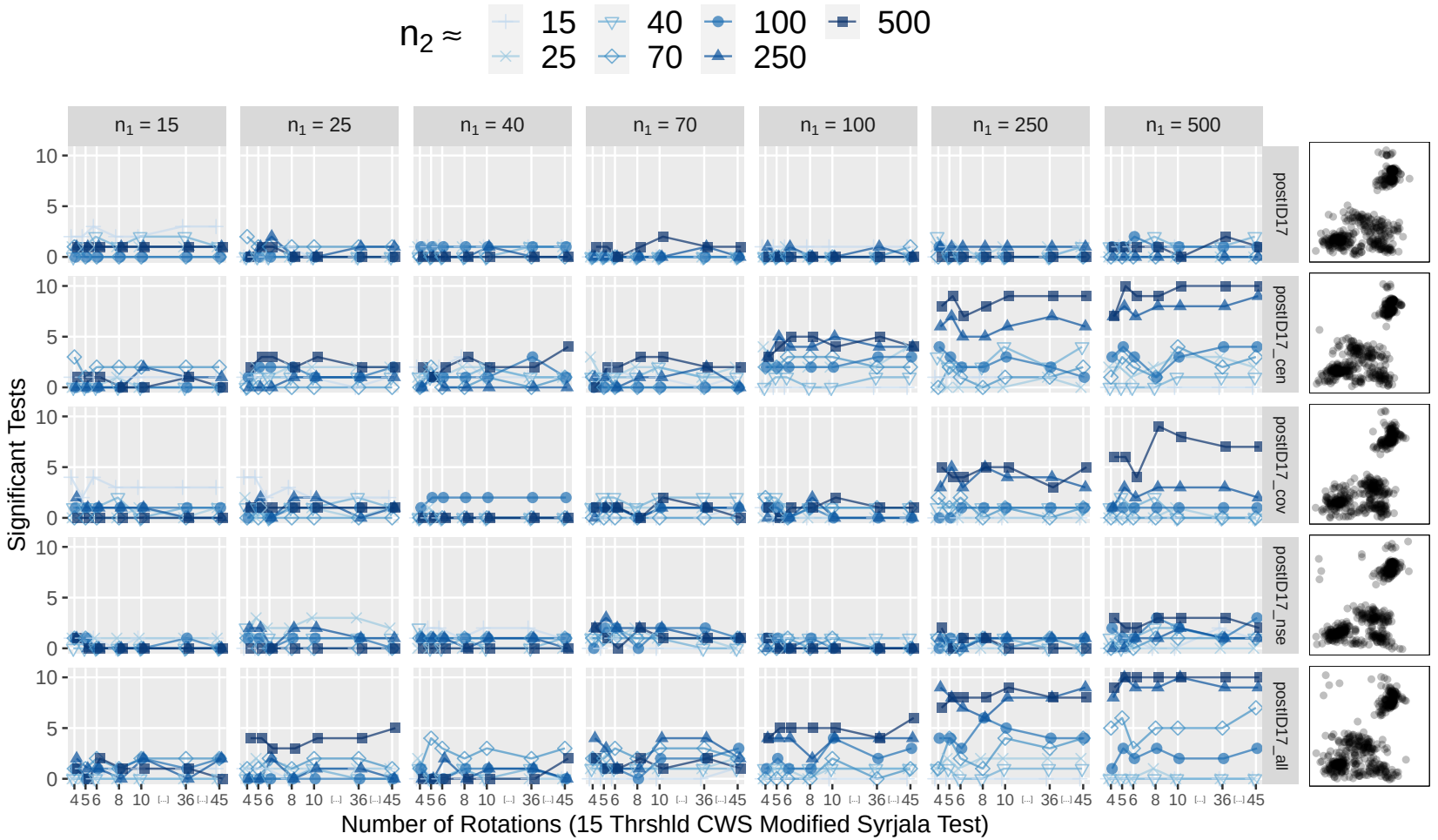


Fig. 46: A grid of line graphs showing the performance of the modified Syrjala test (using a threshold of 15 points as origins of toroidal shifts, 8 rotations, and the CWS statistic) on simulated eye-tracking data which is similar to the aggregated subject contributions to Posture ID 17 (see Appendix C). The grid column name indicates the first sample size (n_1), and the grid row indicates the shape of the second sample. The shape of the first sample follows the same distribution exhibited in the first row. The horizontal axis indicates the second sample size (n_2). For example, the bottom left graph shows the number of significant tests (out of ten tests) between the first samples with 15 points and second samples with 15, 25, 40, 70, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

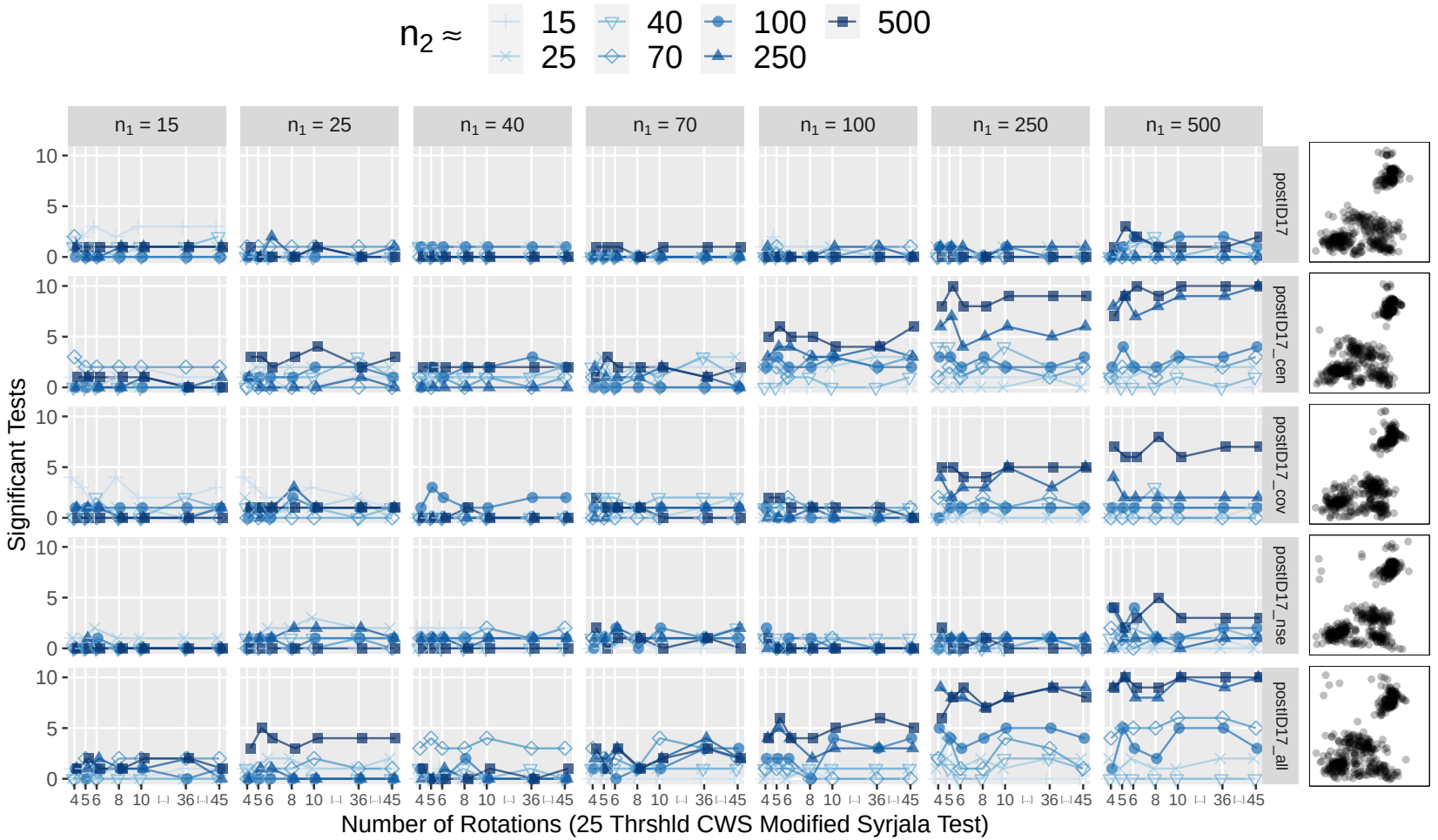


Fig. 47: A grid of line graphs showing the performance of the modified Syrjala test (using a threshold of 25 points as origins of toroidal shifts, 8 rotations, and the CWS statistic) on simulated eye-tracking data which is similar to the aggregated subject contributions to Posture ID 17 (see Appendix C). The grid column name indicates the first sample size (n_1), and the grid row indicates the shape of the second sample. The shape of the first sample follows the same distribution exhibited in the first row. The horizontal axis indicates the second sample size (n_2). For example, the bottom left graph shows the number of significant tests (out of ten tests) between the first samples with 15 points and second samples with 15, 25, 40, 70, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

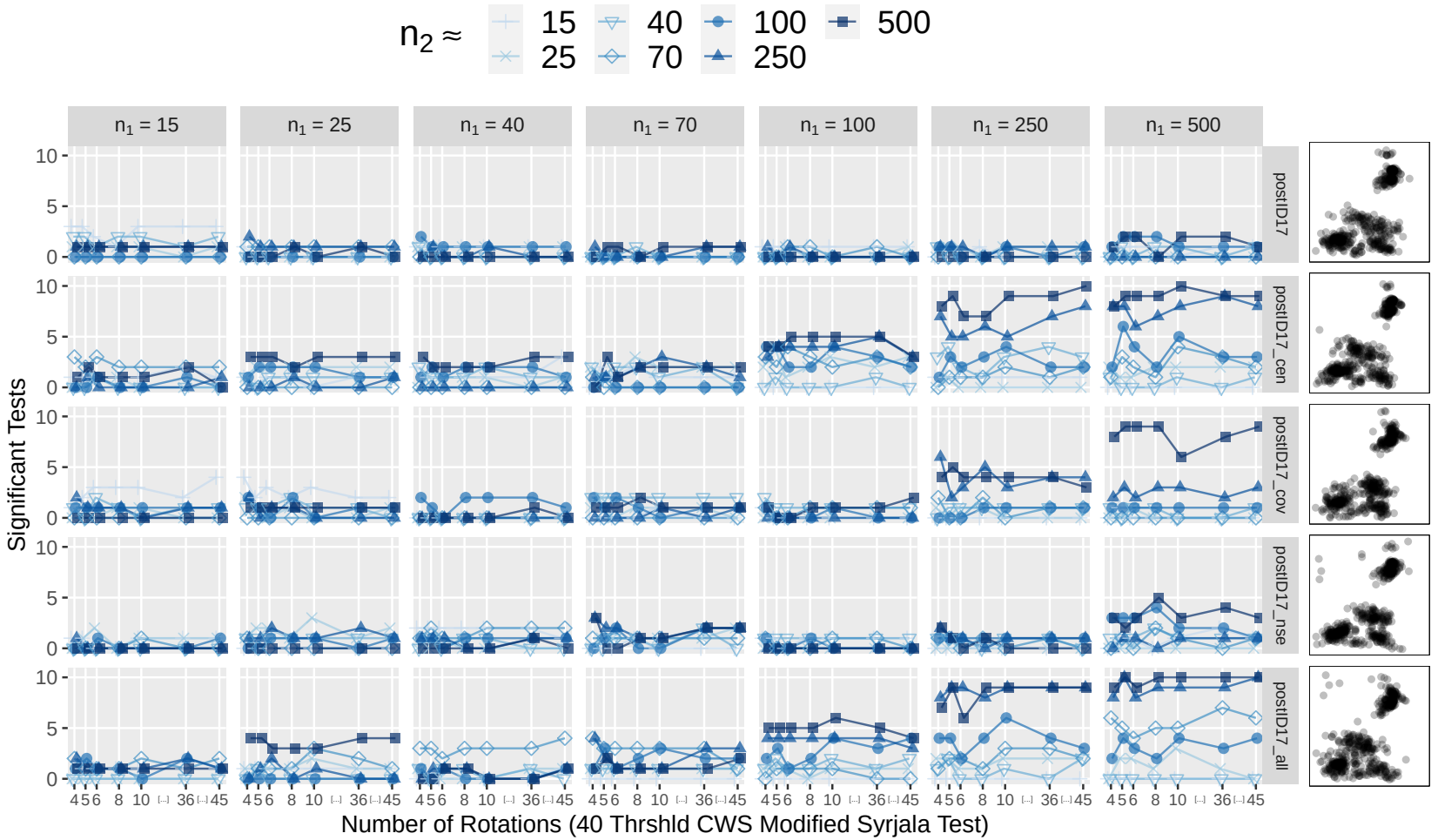


Fig. 48: A grid of line graphs showing the performance of the modified Syrjala test (using a threshold of 40 points as origins of toroidal shifts, 8 rotations, and the CWS statistic) on simulated eye-tracking data which is similar to the aggregated subject contributions to Posture ID 17 (see Appendix C). The grid column name indicates the first sample size (n_1), and the grid row indicates the shape of the second sample. The shape of the first sample follows the same distribution exhibited in the first row. The horizontal axis indicates the second sample size (n_2). For example, the bottom left graph shows the number of significant tests (out of ten tests) between the first samples with 15 points and second samples with 15, 25, 40, 70, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

Furthermore, the abbreviation "postID17" on the far right of the top row of Figures 43–48 (for the null distribution) indicates that a bivariate normal mixture distribution (similar to posture ID 17 from the USU Posture Study control group) was used to generate the data (see Section 6.5.9 for more details). The three departures from the null (shown on the far right of the second through fourth rows) are expressed as (1) changes in the center of the legs distribution from $(0.54, 0.29)$ to $(0.41, 0.34)$ abbreviated as "postID17_cen", (2) changes in the covariance structure for the head, neck, sur. shoulder, and sur. r. foot distributions (see Table 13) abbreviated as "postID17_cov", and (3) an increase in the multinomial probability of success for the event associated with the noise distribution from 0.004 to 0.0178 abbreviated as "postID17_nse". The remaining departure from the null distribution in the fifth row is a combination of all three of the other departures ("postID17_cen", "postID17_cov", and "postID17_nse") and is abbreviated as "postID17_all".

Overall, regardless of the number of rotations, proportions of points used for toroidal shifts, or toroidal shift thresholds in each test, Figures 43–48 exhibit almost the same test results aside from chance variation. For the postID17_cen case, all of the tests exhibit difficulty in detecting at least a majority (greater than or equal to five out of ten) of significant differences until both sample sizes are at least 250 (except for a few exceptions). A similar majority is not achieved in the postID17_cov case until both sample sizes are 500. However, the postID17_nse case never achieves consistently above six out of ten tests across all of the test modifications. Additionally, while all of the departures from the null distribution are combined in the postID17_all case, only a few additional cases achieve above five out of ten tests as statistically significant as compared to the postID17_cen case. Hence, the postID17_cen departures within the postID17_all case seem to be contributing the most to the total number of significant differences being detected.

As a reminder, common random numbers are being employed across all of the simulations (see Section 6.1.3). Thus, the same ten replications of each simulation scenario (e.g., postID17 vs. postID17_cen) pairs are being compared across these simulations (Figures 27–34). Therefore, it is not reasonable to say that the increases in the number of significant tests for the case when $n_1 = 25$ and $n_2 \approx 250$ for the postID17_all case are due to the test modifications. The increase in performance can be attributed to chance variation in the samples.

An overview of the power and false positive rates of the tests employed in this section are provided in Figures 49 and 50 and Tables 14 and 15. Due to the stability of the test results regardless of the test statistic used shown in Sections 6.4.2 and 6.5.3, only the CWS statistic was used within the combined modification test in Figures 35 and 36. As a reminder, the power was computed by dividing the number of significant tests by the total number of tests in which the null hypothesis was false (bottom four rows of graphs in Figures 43–48). In Figure 49, the higher the power of a test the more likely the test is to reject the null when it is indeed false. Theoretically, the maximum power a test can achieve is one.

Table 14: A table listing the test abbreviation, number of significant tests, total number of tests, and power (rounded to the third decimal place) for all of the tests considered in Figure 49.

Test (Abbreviation)	Sig. Tests	Total Tests	Power
CWS RotToro0.1	1897	13 720	≈ 0.138
CWS RotToro0.2	1869	13 720	≈ 0.136
CWS RotToro0.3	1902	13 720	≈ 0.139
CWS RotToro15Thrshld	1862	13 720	≈ 0.136
CWS RotToro25Thrshld	1877	13 720	≈ 0.137
CWS RotToro40Thrshld	1902	13 720	≈ 0.139

Overall, the power of the tests in Figure 49 are all much lower than the power shown in Section 6.4.2. However, the results in Figure 49 make sense given the context

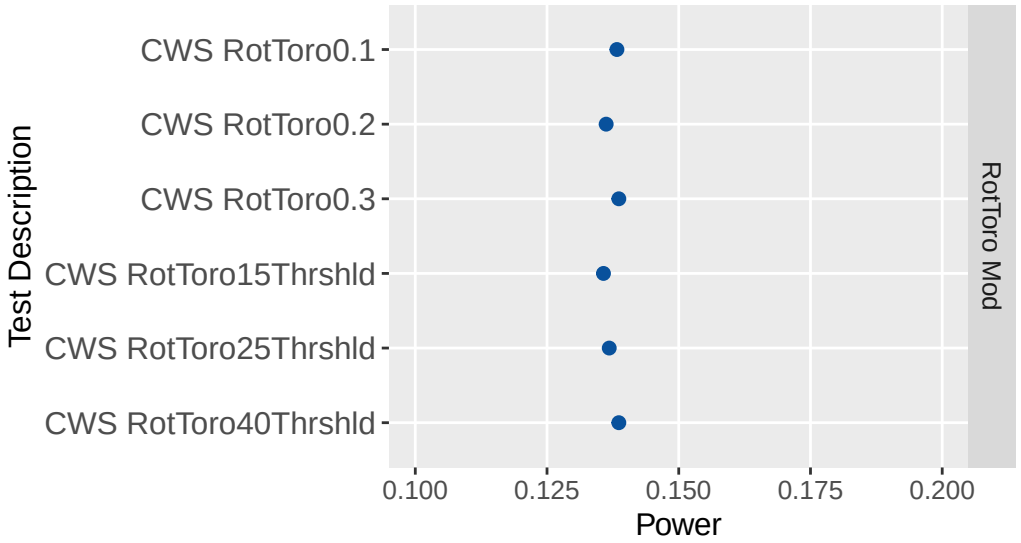


Fig. 49: A comparison of the power achieved by the tests discussed in this section via a horizontal dot plot. The tab on the right indicates that both rotations and toroidal shifts (RotToro Mod) are being used within the tests.

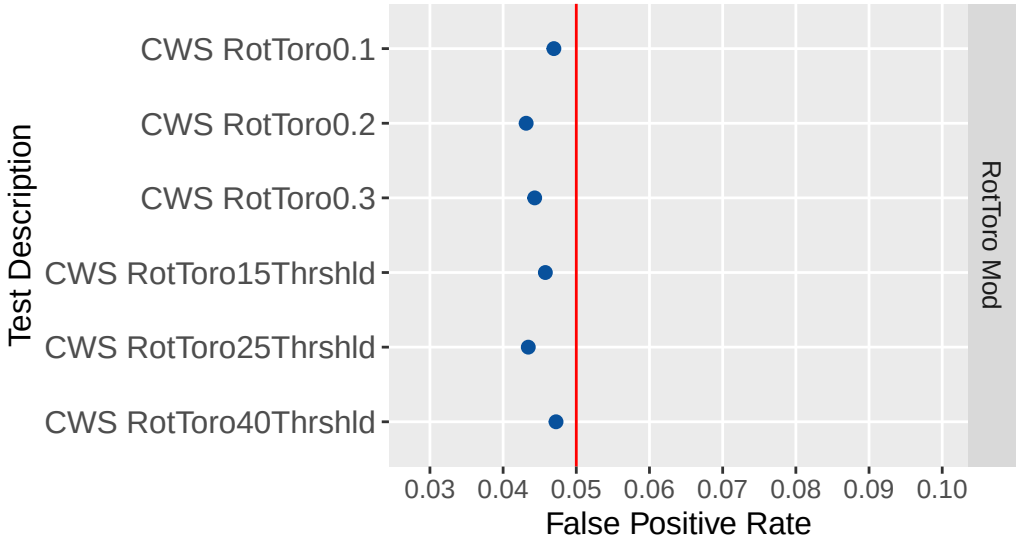


Fig. 50: A comparison of the false positive rates achieved by the tests discussed in this section via a horizontal dot plot. The tab on the right indicates that both rotations and toroidal shifts (RotToro Mod) are being used within the tests. The vertical red line at 0.05 indicates the significance level of the tests.

Table 15: A table listing the test abbreviation, number of significant tests, total number of tests, and false positive rates (rounded to the third decimal place) for all of the tests considered in Figure 50.

Test (Abbreviation)	Sig. Tests	Total Tests	False Positive Rate
CWS RotToro0.1	161	3430	≈ 0.047
CWS RotToro0.2	148	3430	≈ 0.043
CWS RotToro0.3	152	3430	≈ 0.044
CWS RotToro15Thrshld	157	3430	≈ 0.046
CWS RotToro25Thrshld	149	3430	≈ 0.043
CWS RotToro40Thrshld	162	3430	≈ 0.047

of the simulations. The simulations analyzed in Section 6.4.2 involve strong departures from completely spatially random bivariate distributions, whereas the simulations in this section are comparing small changes to the postID17 mixture distribution. Additionally, these simulations explore a larger number of cases with smaller sample sizes. Similar to a standard power analysis (in the field of experimental design) where small differences will not be seen as statistically significant until a sufficient sample size is obtained, so are many of the test results in this section. Hence, it is little surprise that the tests resulted in such the low power. Nonetheless, stable results are seen in the relatively similar values for power across the tests in Figure 49.

In Figure 36, the false positive rate is computed by dividing the number of significant test results by the total number of tests computed when the null hypothesis is true and both samples come from the same distribution (first row of graphs in Figures 43–48). For false positive rates in Figure 50, test results should be as close as possible to 0.05 (i.e., 5%, indicated by the horizontal line) when testing at the 5% significance level. Test results which fall below 0.05 are indications of a conservative nature in the test (i.e., a test which is less likely to reject the null when it is actually true).

Additionally, little trend is seen in the false positive rates of the combined rotational and toroidal tests regardless of the proportions of points used for the origins of toroidal shifts or the thresholds used for the number of toroidal shifts. This case shows that all of the tests achieve a conservative false positive rate (below the significance level of 0.05). While the false positive rate of the 0.1 proportion test is closest to the significance level among the proportion tests with a false positive rate of 0.047, the 25 threshold is the most conservative with a false positive rate of approximately 0.043. Still, these results combined with the results of Sections 6.4.2 and 6.5.3 provide evidence that a 0.1 proportion of points and a 25 toroidal shift threshold are sensible default parameters for the `distdiffR` `R` package functions (described in more detail in Chapter 8).

Results for the Rot8Toro0.1 Version of the Modified Syrjala Test

Figure 51 also displays a grid of line graphs as described in Section 6.5.2. The first row shows 21 out of the 490 test results are significant. This is roughly 0.043 or 4.3%. This result agrees with the conservative nature of the test as seen in previous simulations (see Sections 6.4.1 and 6.4.2).

In the second row, a single departure from the null distribution is analyzed. Particularly, a shift of the legs distribution from a center coordinates of $(0.54, 0.29)$ to $(0.41, 0.34)$, i.e. a shift of $\sqrt{(0.54-0.41)^2+(0.29-0.34)^2} \approx 0.14$ in magnitude. This is approximately the same magnitude of shift exhibited in the most extreme departure from in null of for the previous simulation in Section 6.5.3. However, while a shift of this magnitude resulted in almost all of the tests being significant in the bottom rows of the figures in Section 6.5.3, the legs distribution only makes up approximately 20% of the mixture distribution. Hence, a strong indication of significance (a majority of the tests being significant) does not occur until $n_1 \geq 100$ and $n_2 \approx 500$. Hence,

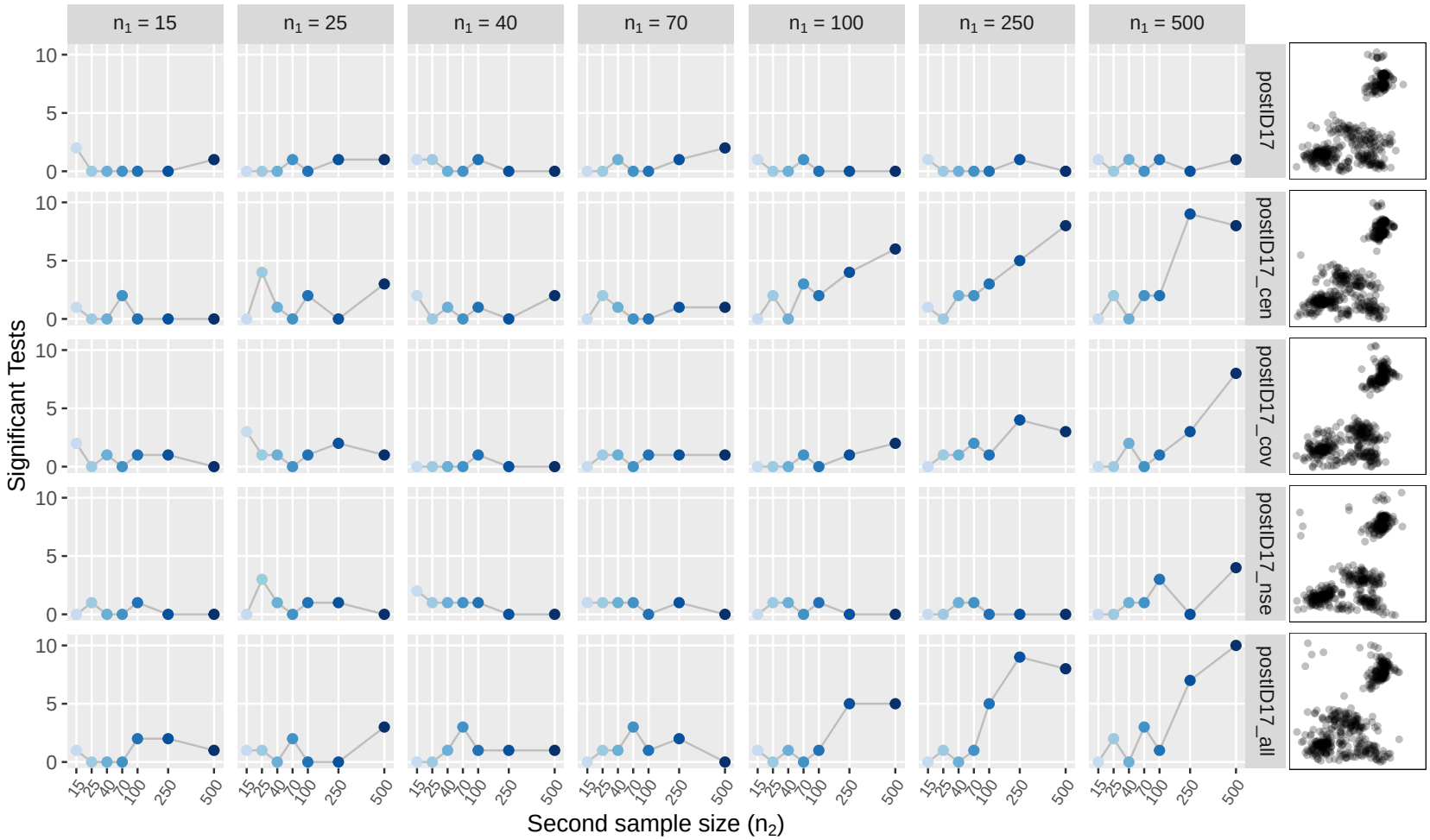


Fig. 51: A grid of line graphs showing the performance of the modified Syrjala test (using 0.1 proportion of points as origins of toroidal shifts, 8 rotations, and the CWS statistic) on simulated eye-tracking data which is similar to the aggregated subject contributions to Posture ID 17 (see Appendix C). The grid column name indicates the first sample size (n_1), and the grid row indicates the shape of the second sample. The shape of the first sample follows the same distribution exhibited in the first row. The horizontal axis indicates the second sample size (n_2). For example, the bottom left graph shows the number of significant tests (out of ten tests) between the first samples with 15 points and second samples with 15, 25, 40, 70, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

larger sample sizes or larger shifts in gaze point component distributions must occur in order to detect significance for these more complex mixture distributions which more closely pattern real-life data.

The third row analyzes a departure from the initial distribution through changes in the covariance structures for the head, neck, sur. shoulder, and sur. r. foot distributions. This is similar to the simulation studied in Section 6.5.4. However, here four of the component distributions have changing covariance structures (which make up roughly 40% of the overall mixture distribution) instead of only one main distribution. Hence, greater changes in the covariance structure or larger sample sizes are needed in order to make a majority of the tests significant. Indeed, this only occurs when $n_1 = 500$ and $n_2 = 500$.

The fourth row from the top compares the initial mixture distribution to a similar one except with a larger proportion of noise. The increase of the proportion of noise from approximately 0.004 to 0.0178 was chosen to reflect the differences in proportions of outliers exhibited between the treatment and control groups of the USU Posture Study. However, we see here that none of the combinations of sample sizes resulted in a majority of tests being significant. This result contributes to the foundation of evidence which suggests that the test is relatively robust in the presence of a small number of outliers (see also Sections 6.5.6, 6.5.7, and 6.5.8).

The bottom row combines all of the departures from previous three rows. The results are quite similar to the second row, which suggests that while there is some amount of contribution to significance from changes in the covariance structures and additional outliers, a majority of the significance is being contributed by the change in the center of gaze point clusters. This result is consistent with the previous results found in the simulations in Sections 6.5.3–6.5.8.

6.5.10 Simulating Differences in Fixation Allocation within the USU Posture Study Data

Similar to the simulation in Section 6.5.9, the simulation in this section also employs generated data which models more complicated real-life data such as data taken from the USU Posture Study (see Chapter 4 for more details).

However, while the simulation in Section 6.5.9 builds upon those discussed in Sections 6.5.3, 6.5.4, 6.5.7, and 6.5.8, this simulation aims at building upon the simulation carried out in Section 6.5.5 where the second subject's visual attention was split between two objects. This simulation takes a closer look at the affect of changes in proportions of points allocated between the various gaze point clusters.

The first sample is still modeled exactly as described in Section 6.5.9. However, instead of the four departures from the initial distribution exhibiting changes the cluster centers, cluster shape, or an increased presence of outliers, the four departures differ only in the multinomial probabilities assigned to the clusters.

Specifically, Figure 52 displays five bar charts of the multinomial event success probabilities (vertical axes) for each mixture distribution cluster (horizontal axes) for the original distribution (top-center) as well as each of the four departures from the initial distribution (lower four bar charts displayed in order of top-left, top-right, bottom-left, and bottom-right). The departures gradually change the multinomial probabilities from those approximated by the original data to a uniform probability across all of the clusters except for the between feet (abbreviated as b. feet) and noise distributions.

The multinomial probabilities of these latter two distributions were held constant in this simulation for several reasons. First, the b. feet distribution is unique in that it was likely created by subjects looking from the posture in the projected image down to a time bar and back up at the posture. Some form of the data can be

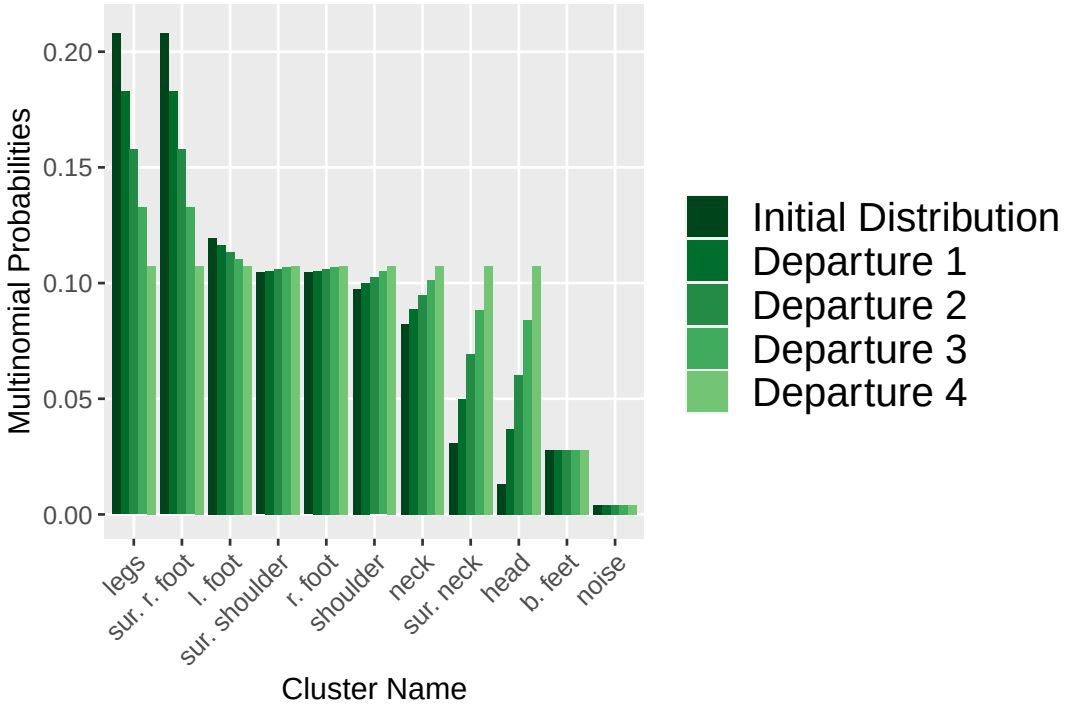


Fig. 52: A side-by-side bar chart displaying the multinomial event success probabilities (vertical axes) for each mixture distribution cluster (horizontal axes) for the initial distribution (left-most bar in each group) as well as each of the four departures from the initial distribution (four right-most bars in each group, respectively).

seen throughout several other aggregated collections of the gaze points between the treatment and control groups. Hence, since this cluster of gaze points is more related to the learning curve of the experiment and less related to what subjects are looking for when trying to assess the postural stability of a depicted actor, it is omitted from the uniform flattening of the multinomial probabilities.

Similarly, the amount of noise in the data is much more likely related to interference with the ability of the eye-tracking device accurately assigning coordinates to gaze points (due to a variety of unknown circumstances, e.g., the frequency of subject blinking during the experiment), and less likely to be related to what subjects are looking for when trying to assess the postural stability of a depicted actor. Hence, the multinomial event success probability associated with noise is held constant for

this simulation.

Thus, for the first sample, the cluster probabilities are assigned as labeled in Table 12. When assuming that the null hypothesis is true, the second sample is also generated using the same multinomial probabilities. However, when assuming the null hypothesis is false, and some departure from the initial distribution has occurred, the probabilities for the second sample are taken from one of the four lower bar charts in Figure 52.

These individual probabilities were computed in the following way. First, compute the average probability (μ_c) while omitting the b. feet and noise probabilities (0.0265 and 0.004, respectively).

$$\mu_c = \frac{1 - 0.0265 - 0.004}{11}$$

Then, $f_i(p_c)$ the new probability for the i^{th} departure from the original distribution can be computed.

$$f_i(p_c) = \begin{cases} p_c - i \cdot \frac{|p_c - \mu_c|}{11}, & \text{if } p_c \geq \mu_c \\ p_c + i \cdot \frac{|p_c - \mu_c|}{11}, & \text{if } p_c < \mu_c \end{cases},$$

where p_c is the original multinomial event success probability for the given cluster.

Results

Figure 53 displays a grid of line graphs as described in Section 6.5.2. Here, the first row shows 39 out of the 490 test results are significant. This is roughly 0.08 or 8%. This result is greater than the 5% significance level, which does not agree with the conservative nature of the test as seen in previous simulations (see Sections 6.4.1 and 6.4.2). However, this is likely due to chance error as a repetition of the simulation using different random number seeds resulted in a false positive rate

below the significance level.

In the remaining four rows, the departures from the null distribution are exhibited by reassigning the multinomial probabilities for the various clusters according to the method discussed in the previous section. A few of these differences in proportion of gaze points being allocated to another object of interest are similar to those exhibited in the previous simulation in Section 6.5.5. Indeed, the results of this simulation are close to those displayed in Figure 38. Hence, similar larger sample sizes or larger differences in proportions of gaze points in component distributions (or both) must occur in order to detect significance for a majority of these more complex mixture distributions which more closely pattern real-life data.

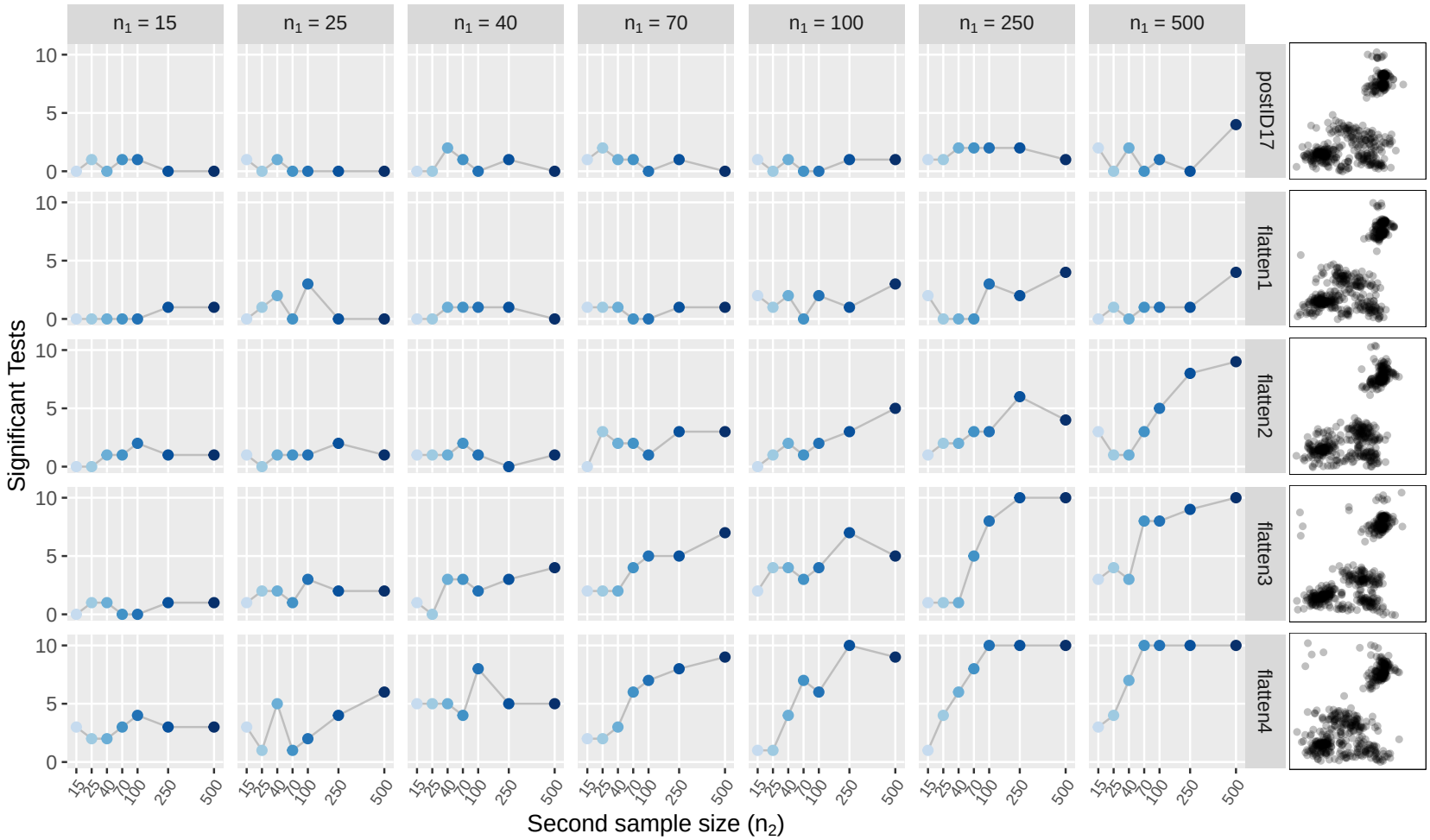


Fig. 53: A grid of line graphs showing the performance of the modified Syrjala test (using 0.1 proportion of points as origins of toroidal shifts, 8 rotations, and the CWS statistic) on simulated eye-tracking data which is similar to the aggregated subject contributions to Posture ID 17 (see Appendix C). The grid column name indicates the first sample size (n_1), and the grid row indicates the shape of the second sample. The shape of the first sample follows the same distribution exhibited in the first row. The horizontal axis indicates the second sample size (n_2). For example, the bottom left graph shows the number of significant tests (out of ten tests) between the first samples with 15 points and second samples with 15, 25, 40, 70, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

6.5.11 Eye-Tracking Inspired Simulation Results

Overall, the simulations carried out in Sections 6.5.3–6.5.10 provide a more accurate view of the performance of the modified Syrjala test (when using 0.1 proportion of points as toroidal shift origins, eight rotations, and the CWS statistic) when applied to data which more closely represent that taken from eye-tracking analysis.

Specifically, although outliers are commonly found in eye-tracking data (as mentioned in Chapter 3), the test has been shown to be robust against a small number of outliers (see Sections 6.5.6, 6.5.7, 6.5.8, and 6.5.9). Furthermore, the test can detect a variety of differences between samples, including differences in gaze point cluster centers (Sections 6.5.3 and 6.5.9), cluster shapes (Sections 6.5.4 and 6.5.9), proportions allocated to different objects being viewed by subjects (Sections 6.5.5 and 6.5.10), and proportions of noise (Sections 6.5.7, 6.5.8 and 6.5.9).

However, for some of the simulated the differences which are subtle and/or when the sample sizes are relatively small, the test is unable to label a majority of the comparisons as significant. Still, the overall performance of the test on simulated eye-tracking data is sufficient to say that the test is well suited for eye-tracking analysis. The test is applied to the USU Posture Study data in Chapter 7.

6.6 Conclusions from the Simulation Results

From the series of simulations that have been discussed throughout Sections 6.2–6.5, the following conclusions can be made:

- The Syrjala (1996) test has been shown to depend upon data aggregation techniques such as regular and random binning. It is recommended to use another bivariate two sample test of distributional equality which does not assume identical sampling locations. Such tests include the Energy test by Székely and Rizzo (2004), the kernel maximum mean discrepancy by Gretton et al. (2012),

the Friedman and Rafsky (1979) generalization to the Kolmogorov (1933) test, or one of the modified Syrjala tests proposed in this dissertation.

- The modified Syrjala tests have been shown to be insensitive to differences in the weightings of the tests statistics, and only a marginal gain in power was found when using squared differences in the ECDFs as compared to absolute differences. Additionally, all of the tests were shown to produce relatively stable results regardless of the number of rotations or toroidal shifts explored within the simulations.

- While the modified Syrjala tests which employ toroidal shifts achieve roughly the same power as the tests which employ both rotational and toroidal shifts, the latter tests achieve an average false positive rate (0.03 vs. 0.045, respectively) closer to the significance level (0.05). Thus, the combined modifications produce conservative tests which are more powerful in the face of all departures from the null than the tests which employ toroidal shifts alone. However, this balance comes at the cost of increased computational complexity.

- The modified Syrjala test which uses eight rotations, 0.1 proportion of points as origins for toroidal shifts, and the CWS statistic has been shown by simulation to achieve a higher number of significant tests (when the null is false) than several other competing methods including the Energy test by Székely and Rizzo (2004), the kernel maximum mean discrepancy by Gretton et al. (2012), the Friedman and Rafsky (1979) generalization to the Kolmogorov (1933) test, and the original Syrjala (1996) test (when preliminary data binning is employed).

- The modified Syrjala test which uses a threshold of 25 randomly chosen points from the pooled sample as origins for toroidal shifts (see Section 6.3.4) achieves comparable results as the tests which employ proportions of points as origins

for toroidal shifts (see Section 6.3.3). This provides motivation to use a default threshold value for the tests in the `R` package (see Chapter 8), which guides new users of the package toward relatively reasonable parameter values.

- The modified Syrjala tests which employ the CWS statistic, eight rotations, and 0.1 proportions of points as toroidal shifts, or thresholds of 25 toroidal shifts have been shown to be well suited to certain types of eye-tracking data by simulation. Specifically, the tests are robust to a reasonable number of outliers, and can detect a variety of differences in distributional structure. Consequently, these tests are used when analyzing data from the USU Posture Study in Chapter 7.

6.7 Simulation Computational Performances

While the modified Syrjala tests have been shown to be more powerful than alternative tests in the literature (Section 6.4), the modified Syrjala tests are more computationally expensive, especially when employing both the rotational and toroidal shift modifications. Consequently, Section 6.7.1 provides a benchmarking study to assess the computational expense of the modified Syrjala test (which employs 0.1 proportion of points as toroidal shift origins, eight rotations, and the CWS statistic). Section 6.7.2 details the specifications of computational resources used at the University of Utah's Center for High Performance Computing (CHPC, `https://www.chpc.utah.edu/`).

6.7.1 A Benchmarking Study of the Modified Syrjala Test

To provide a relative understanding of how computationally expensive the modified Syrjala test (using 0.1 proportion of points as toroidal shift origins, eight rotations, and the CWS statistic) is on different machines, the test was applied to iterations of

the null distribution of the eye-tracking inspired simulation data (see Section 6.5.9 across a series of sample sizes. Figure 54 shows the mean computational time in minutes for ten replications (vertical axis) for the modified Syrjala test across the following sample sizes (horizontal axis): 25, 50, 75, 100, 125, 150, 175, 200, 250, 300, 350, 400, 500, and 600. Each of these sample sizes were used for both the first and second samples. Hence, the total number of data values used is simply two times the horizontal axis values. Additionally, the test was applied ten times for each of the sample sizes, except for the sample sizes 500 and 600 for Intel i5 where the test was only applied once due to the computational cost.

While side-by-side boxplots were considered for this figure, it turned out that the maximum standard deviation of computational times between the machines and groups of replications was 0.117 minutes (7.02 seconds). Hence, the small variability compared to the large range in the vertical axis made plotting boxplots less meaningful than a simple mean for each replication group.

To prevent overplotting, the computational times from the various machines were slightly offset horizontally. The specifications of each of the computational environments for each of the machines displayed in Figure 54 are listed in Table 16.

Many factors are at play when timing the clock speed of the modified Syrjala test on these different machines. Specifically, while the AMD Server consisted of nodes with many more cores and available RAM than the Intel i7 Desktop, the computation of the tests were run in serial (using only a single CPU core at a time). The AMD CPUs on the (Linux) Server are also throttled down to 2.7GHz when many users are running jobs simultaneously on the remaining cores, which is often the case. Consequently, with a higher maximum turbo frequency of 3.90, and more available cache memory than the Intel i5 Laptop (8MB versus 3MB, respectively), the Intel i7 Desktop demonstrated the fastest times among the three machines.

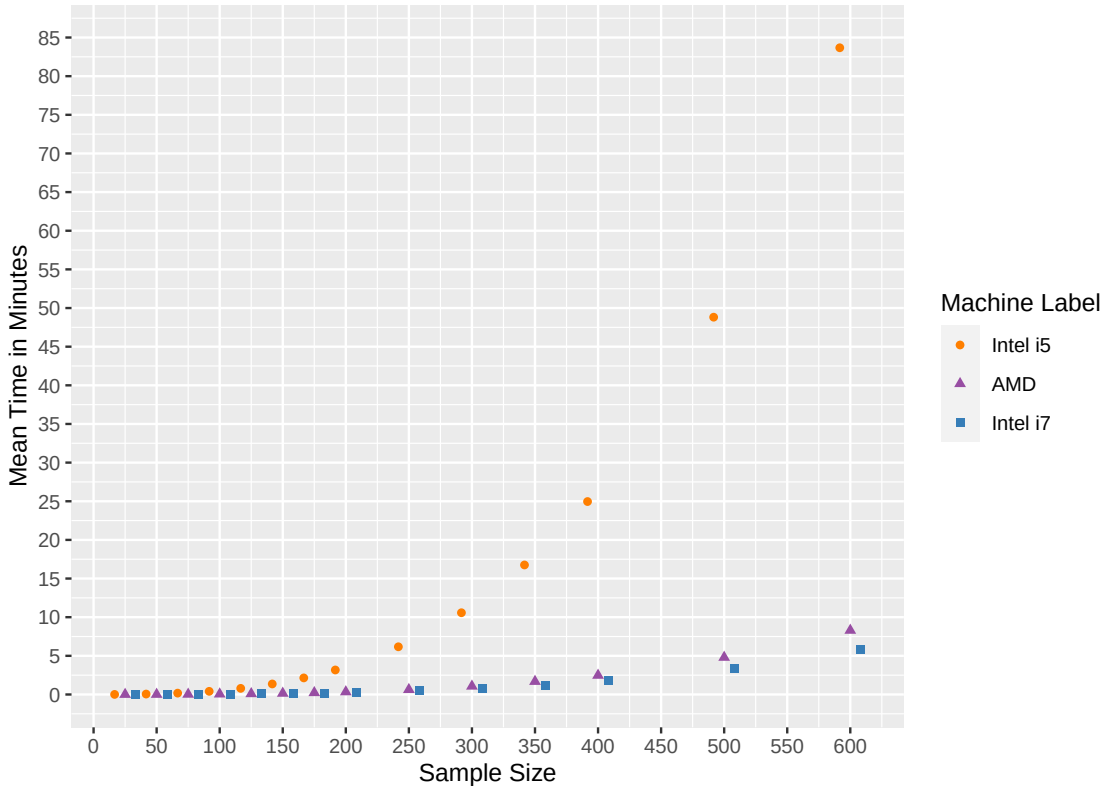


Fig. 54: A graph of mean computational times (in minutes) for the modified Syrjala test (using 0.1 proportion of points as toroidal shift origins, eight rotations, and the CWS statistic) when applied to eye-tracking inspired generated data (see Section 6.5.9 on various machines.

6.7.2 Computational Resource Specifications

All of the simulation studies carried out in this dissertation (see Sections 6.2–6.5) were computed using the University of Utah's CHPC (`https://www.chpc.utah.edu/`) resources. Table 17 provides the CHPC cluster used, the total number of cores, the total RAM, and computational times (in minutes) for each test of interest within that simulation. The section number of the simulation is also provided for direction to further details on each simulation. "OnDemand" in the cluster column indicates that the Open OnDemand (`https://openondemand.org/`) interface was used to queue jobs on the Notchpeak-shared-short nodes. The CHPC documentation provides more details into the specifications of the computational re-

sources (see `https://www.chpc.utah.edu/documentation/guides/notchpeak.php` and `https://www.chpc.utah.edu/documentation/guides/kingspeak.php`).

Notice that while individual tests run faster on other machines (see Section 6.7.1), many more tests were able to be computed simultaneously using the CHPC resources than would have been feasible on a common personal machine.

Table 16: A table of each of the computational environment specifications for each of the machines displayed in Figure 54.

Machine Label	Processor	Base Frequency	Max Turbo Frequency	Total RAM	Operating System
Intel i5	Intel(R) Core(TM) i5-7200U	2.50GHz	3.10GHz	8GB	Windows 10 Enterprise
AMD	AMD EPYC 7601	2.20GHz	3.20GHz	32GB	CentOS Linux 3.10.0
Intel i7	Intel(R) Core(TM) i7-1065G7	1.30GHz	3.90GHz	16GB	Windows 10 Home

Table 17: A table of each of the environment specifications and computational times for each of the simulation studies carried out in Sections 6.2–6.5.

Section Number	Test	Cluster	Cores	Total RAM	Comp. Time (in mins)
6.2.2	Syrjala	OnDemand	4	128GB	0.6
6.3.2	Rot. DWS	Notchpeak	64	512GB	10.3
6.3.2	Rot. UWS	Notchpeak	64	512GB	9.8
6.3.2	Rot. CWS	Notchpeak	64	512GB	9.8
6.3.2	Rot. DWA	Notchpeak	64	512GB	9.8
6.3.2	Rot. CWA	Notchpeak	64	512GB	9.5
6.3.2	Rot. UWA	Notchpeak	64	512GB	10.0
6.3.2	Toro. DWS	Notchpeak	64	512GB	121.9
6.3.2	Toro. CWS	Notchpeak	64	512GB	124.2
6.3.2	Toro. UWS	Notchpeak	64	512GB	111.3
6.3.2	Toro. DWA	Notchpeak	64	512GB	118.3
6.3.2	Toro. CWA	Notchpeak	64	512GB	115.6
6.3.2	Toro. UWA	Notchpeak	64	512GB	115.2
6.3.3	0.1 Toro. Rot. DWS	Lonepeak	80	1TB	511.0
6.3.3	0.1 Toro. Rot. UWS	Lonepeak	80	1TB	505.6
6.3.3	0.1 Toro. Rot. CWS	Kingspeak	48	192GB	1002.9
6.3.3	0.2 Toro. Rot. DWS	Lonepeak	80	1TB	999.3
6.3.3	0.2 Toro. Rot. UWS	Lonepeak	80	1TB	994.1
6.3.3	0.2 Toro. Rot. CWS	Kingspeak	48	192GB	2059.3
6.3.3	0.3 Toro. Rot. DWS	Lonepeak	80	1TB	1480.9
6.3.3	0.3 Toro. Rot. UWS	Lonepeak	80	1TB	1489.5
6.3.3	0.3 Toro. Rot. CWS	Kingspeak	48	192GB	3087.5
6.3.4	25 Toro. Thrshld. DWS	Lonepeak	80	1TB	190.0
6.3.4	25 Toro. Thrshld. UWS	Lonepeak	80	1TB	187.2
6.3.4	25 Toro. Thrshld. CWS	Lonepeak	80	1TB	188.6
6.4.1	Energy	OnDemand	4	128GB	0.2
6.4.1	Kmmd	OnDemand	4	128GB	2.2
6.4.1	FR-KS	OnDemand	4	128GB	11.7
6.4.1	0.1 Toro. 8 Rot.	OnDemand	4	128GB	466.8
6.5.3	0.1 Toro. 8 Rot.	OnDemand	8	128GB	244.2
6.5.4	0.1 Toro. 8 Rot.	OnDemand	8	128GB	243.6
6.5.5	0.1 Toro. 8 Rot.	OnDemand	16	128GB	87.0
6.5.6	0.1 Toro. 8 Rot.	OnDemand	8	128GB	262.8
6.5.7	0.1 Toro. 8 Rot.	OnDemand	8	128GB	244.2
6.5.8	0.1 Toro. 8 Rot.	OnDemand	8	128GB	247.2
6.5.9	0.1 Toro. 8 Rot.	OnDemand	16	128GB	67.8
6.5.10	0.1 Toro. 8 Rot.	OnDemand	16	128GB	87.6

CHAPTER 7

Applications to the Utah State University Posture Study

An application of one of the modified Syrjala tests to the USU Posture Study is provided within this chapter. After a brief overview in Section 7.1, Section 7.1.1 provides the analyses of group-wise comparisons within the study followed by an example in Section 7.1.2. Section 7.1.3 details within-group comparisons, and Section 7.1.4 provides some within-group comparison examples. Concluding remarks are also provided in Section 7.1.5.

7.1 USU Posture Study Analyses

In answering the question, "Does judging the action capabilities of another person depend on one's own experiences?" two-sample tests of distributional difference were applied to the USU Posture Study data (described in greater detail in Chapter 4) in order to make group-wise comparisons (Section 7.1.1) as well as within-group comparisons (Section 7.1.3).

Due to the stability of results exhibited in the simulations in Chapter 6, the modified Syrjala test with the following test statistic properties has been applied: (1) eight rotations of the data, (2) 0.1 proportion of the gaze points as origins of toroidal shifts, (3) the CWS statistic. The number of permutations of the data is 99 for both the group-wise test and the pair-wise tests when conducting the within-group comparisons. While more permutations are typically used in this scenario in order to achieve significance in the most extreme cases even after a correction for multiple comparisons is made, only 99 are used here due to the rich number of significant tests discovered in some preliminary analyses of the data. Hence, the

smallest possible p-value is 0.01 for a permutation test where the original test statistic is more extreme than the 99 test statistics computed on permutated versions of the data. See Section 7.1.3 for further details.

7.1.1 Group-wise Comparisons

Group-wise comparisons were made between the treatment (yoga) and control (non-yoga) groups for all 22 postures. In other words, all of the treatment subject's data were aggregated and compared to all of the control subject's data for the same posture. However, since each subject typically spent a different amount of time viewing a particular posture, differing amounts of gaze points are contributed by each subject when combining all subject data together. To resolve this issue, each subject's contribution is given the same weight in the group-wise comparisons, rather than allowing subjects with more contributions to go unweighted and thus have a greater impact on the test statistic than subjects who contributed less. The hypotheses for this test are as follows:

H_0: The grouped distributions of the treatment and control populations which weight each subject's contributions equally are the same across the viewing region.

H_a: There is some unspecified difference between the treatment and control group population distributions (which weight each subject's contributions equally).

This weighting is accomplished by modifying the test statistic in the following manner. Each subject's bivariate ECDF is computed separately. The ECDFs are grouped by treatment and control, and differences in mean ECDF values are compared between the two groups of ECDFs. This new statistic is referred to as ξ_g^{CWS} (the

subscript g refers to "grouped"), and can be computed as

$$\xi_g^{CWS} = \frac{n_2}{(n_1+n_2)} \sum_{i=1}^{n_1} \left[\frac{1}{K} \sum_{k=1}^{K} \Gamma^*_{1,k}(x_{1,i}, y_{1,i}) - \frac{1}{K} \sum_{k=1}^{K} \Gamma^*_{2,k}(x_{1,i}, y_{1,i}) \right]^2 + \frac{n_1}{(n_1+n_2)} \sum_{j=1}^{n_2} \left[\frac{1}{K} \sum_{k=1}^{K} \Gamma^*_{1,k}(x_{2,j}, y_{2,j}) - \frac{1}{K} \sum_{k=1}^{K} \Gamma^*_{2,k}(x_{2,j}, y_{2,j}) \right]^2$$

where there are $K = 20$ subjects within each group, and n_1 and n_2 are the total sample sizes within each group (treatment and control, respectively). This is equivalent to

$$\xi_g^{CWS} = \frac{1}{20(n_1+n_2)} \left(n_2 \sum_{i=1}^{n_1} \left[\sum_{k=1}^{20} \Gamma^*_{1,k}(x_{1,i}, y_{1,i}) - \sum_{k=1}^{20} \Gamma^*_{2,k}(x_{1,i}, y_{1,i}) \right]^2 + n_1 \sum_{j=1}^{n_2} \left[\sum_{k=1}^{20} \Gamma^*_{1,k}(x_{2,j}, y_{2,j}) - \sum_{k=1}^{20} \Gamma^*_{2,k}(x_{2,j}, y_{2,j}) \right]^2 \right) \tag{10}$$

where k is the subject number within group i ($i = 1$ for the treatment group, and $i = 2$ for the control group).

Alternatively, an equivalent way exists to treating each subject's contributions equally. This method involves computing the lowest common multiple (LCM) between all of the gaze point contributions across all of the subjects within each group. Then by duplicating each subject's gaze points by the product of the missing factors the subject's gaze point number lacks to be equal to the LCM, each subject's contribution will be treated equally. However, this method is more computationally intensive than the former. A more detailed description of this method as well as a mathematical proof demonstrating the equivalence between the two methods is provided in Appendix A.

Hence, by using Equation 10, group-wise comparisons were made between the treatment and control groups for each of the 22 postures. The results of the tests as well as the computational time (wall time in hours) are displayed in Table 18. As seen in the table, all of the 22 tests resulted in statistically significant p-values equal

to 0.01. Hence, all the null hypotheses are rejected in favor of the alternative. The associated conclusion is that there is some unspecified difference in the distributions of gaze points between the treatment and control groups for every posture observed. This implies that there is a difference between what subjects look at when assessing postural stability between the two groups in at least one part of an actors posture. However, this does not indicate how much similarity there is between subjects within each group. Hence, an additional analysis was carried out in Section 7.1.3 to assess whether there is any similarity between individual subject's gaze point distributions within each group.

Typically, a correction is made to account for the multiple comparisons. However, this is omitted here due to the abundance of significant results. This is further discussed and justified in Section 7.1.3.

Table 18 also shows the computational times (in hours) of the modified Syrjala tests which use a threshold of 25 toroidal shifts (along with the CWS statistic, and eight rotations) applied to the all of the posture comparisons where each groups data was aggregated. The p-values for both of the tests in Table 18 were 0.01 except for posture ID 6 where the threshold test achieved a p-value of 0.02 as indicated by the asterisk. These results are included here to not only show the agreement in the test results, but to also demonstrate the computational speed of using the toroidal shift thresholds. When assessing the ratios of computational times between the proportional test and the threshold test, the proportional test took on average approximately 8.3 times longer than the threshold test. The associated standard deviation of these ratios of computational times was approximately 1.3.

7.1.2 Group-wise Comparison Example

An example of a group-wise comparison of the aggregated gaze point distributions

Table 18: A table of results and computational times (in hours) from applying the modified Syrjala tests (using the CWS statistic, eight rotations, and either 0.1 proportion of toroidal shifts or a threshold of 25 toroidal shifts) to all of the postures where each groups data was aggregated. The p-values for both of the tests were 0.01 except for posture ID 6 where the threshold test achieved a p-value of 0.02 as indicated by the asterisk *. All of the computations were carried out on the University of Utah's Center for High Performance Computing (`https://chpc.utah.edu/`) on the Notchpeak cluster using 4 cores (see AMD in Table 17) and 128GB of RAM.

Posture ID	n_1	n_2	p-values	Comp. Time (in hours) of Proportional Test	Comp. Time (in hours) of Threshold Test
1	1091	1297	0.01	2.10	0.31
2	1289	1281	0.01	2.49	0.37
3	1356	1346	0.01	2.87	0.39
4	1452	1479	0.01	3.46	0.42
5	1900	1523	0.01	5.00	0.52
6	1716	1360	0.01^*	3.89	0.45
7	1359	1246	0.01	2.63	0.34
8	1089	1186	0.01	1.87	0.32
9	1313	1242	0.01	2.54	0.35
10	1624	1215	0.01	3.26	0.41
11	1184	1443	0.01	2.66	0.35
12	1436	1337	0.01	3.64	0.38
13	1869	1768	0.01	6.45	0.63
14	1421	1202	0.01	3.08	0.48
15	1396	1343	0.01	3.64	0.46
16	1541	1352	0.01	3.97	0.42
17	1661	1508	0.01	4.80	0.59
18	1746	1367	0.01	4.88	0.46
19	1126	1108	0.01	2.34	0.32
20	1348	1374	0.01	3.53	0.37
21	1476	1597	0.01	4.33	0.44
22	1501	1517	0.01	4.04	0.43

for Posture ID 2 is shown in Figure 55. The test results, which employed the ξ_g^{CWS} statistic detailed in Equation 10, were significant (p-value = 0.01) for this comparison. Some of the differences in the aggregated gaze point distributions can be seen in the control group's (right image in Figure 55) emphasis on the top of the head and left arm, and the treatment group's (left image in Figure 55) emphasis on the left side of the torso, right hand, and left foot.

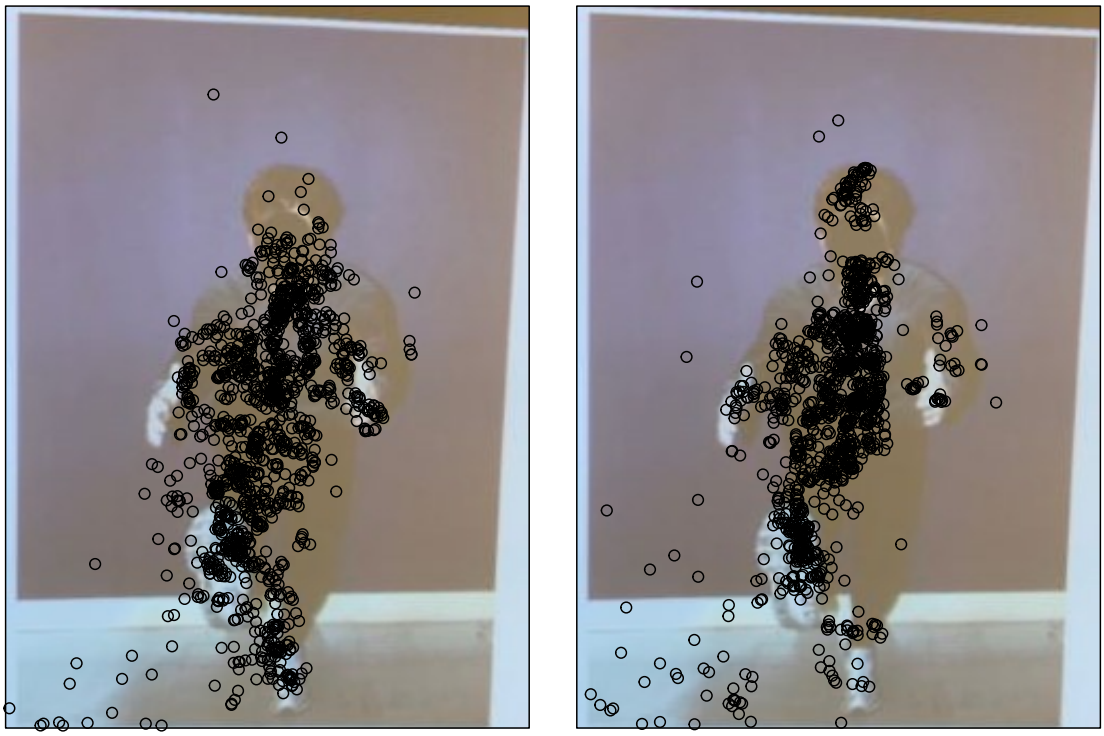


Fig. 55: Scatterplots of the gaze points for treatment (left) and control (right) groups for posture ID 2. The test result was significant (p-value = 0.01) for this comparison.

7.1.3 Within-group Comparisons

Furthermore, within-group comparisons were made to determine whether the viewing patterns in a group would be similar. Specifically, each of the 20 subjects within a group were compared to all of the other subjects within that group. Hence,

for each posture there are $\binom{20}{2} = 190$ comparisons within each subject group (treatment or control). Overall, 190 comparisons $\times$ 22 postures $\times$ 2 groups = 8,360 individual comparisons.

With so many statistical tests being conducted, it is highly likely that the results will produce significant p-values when in fact the null hypothesis (that there is no difference) is true. This is well known within the broader statistical community as the multiple comparisons problem (Miller, 1981). Depending on the type of error measure that is desired to be controlled, a variety of methods have been proposed to account for the few false alarms that arise just by chance, e.g., Bonferroni (1936) proposed a widely used conservative p-value adjustment which controls the family-wise error rate (FWER). Given a set of m independent comparisons, the FWER is defined as

$$\alpha_{FWER} = 1 - (1 - \alpha_{\text{per comparison}})^m.$$

Another commonly used, yet slightly less conservative method for controlling the FWER was proposed by Šidák (1967).

However, for large-scale multiple testing within exploratory studies, controlling the false-discovery rate (FDR) is often preferred over the FWER. The FDR is loosely defined as the expected number of false positives among all of the significant tests. Benjamini and Hochberg (1995) developed a sequential or step-down approach for controlling the FDR. Additionally, Storey (2003) defined a modified version of the FDR, called the positive-FDR (pFDR), and proposed a method for controlling the pFDR by converting p-values into Bayesian posterior p-values called q-values.

Nonetheless, these multiple testing correction methods are of most use when there are only a handful of significant results among many tests with some of the significant tests being due to chance error. By applying the multiple testing correction, the

error rate of interest can be controlled, and some of the significant results due to multiple testing can be attributed to random variation. However, in the case where an overwhelming majority or all of the test results are significant, a multiple testing correction method becomes less informative due to the effect size overshadowing the chance variation.

For example, suppose we want to conduct 100 independent two-sample t-tests for a difference in means between two populations at the 5% significance level. Even if the two populations are identical, and the effect size (or difference in population means) is zero, we would expect to see roughly five significant tests due to chance variation. However, if the two population distributions have little overlap due to considerably different means and small variances, then the effect size would be overwhelmingly large (compared to the chance variation), and a majority or all of the tests would be significant. In this latter case, the conclusion from the 100 independent tests that the two populations have different means would hold regardless of any kind of multiple correction method being applied.

Such is the case with many of the tests conducted in the USU Posture Study. Figures 56–57 show strongly positively skewed histograms of p-values for both the treatment and control group pairwise tests, respectively. Additionally, out of the 4,180 individual tests within each group, only 35 were non-significant within the treatment group, and only 54 were non-significant in the control group. Hence, approximately 99.16% and 98.71% of the pairwise tests were significant in the treatment and control groups, respectively. Furthermore, the largest non-significant p-values within the treatment and control groups are only 0.25 and 0.37, respectively, and many of the remaining non-significant p-values tend to be close to the significance level of 0.05. Due to the richness of significant tests within the respective groups, even if the usual application of a multiple testing correction method is conducted,

the results will be essentially the same, and a conclusion is made that the subjects within each respective group exhibit mostly heterogeneous gaze point patterns. The non-significant test results for the treatment and control groups are listed in Tables 19 and 20, respectively.

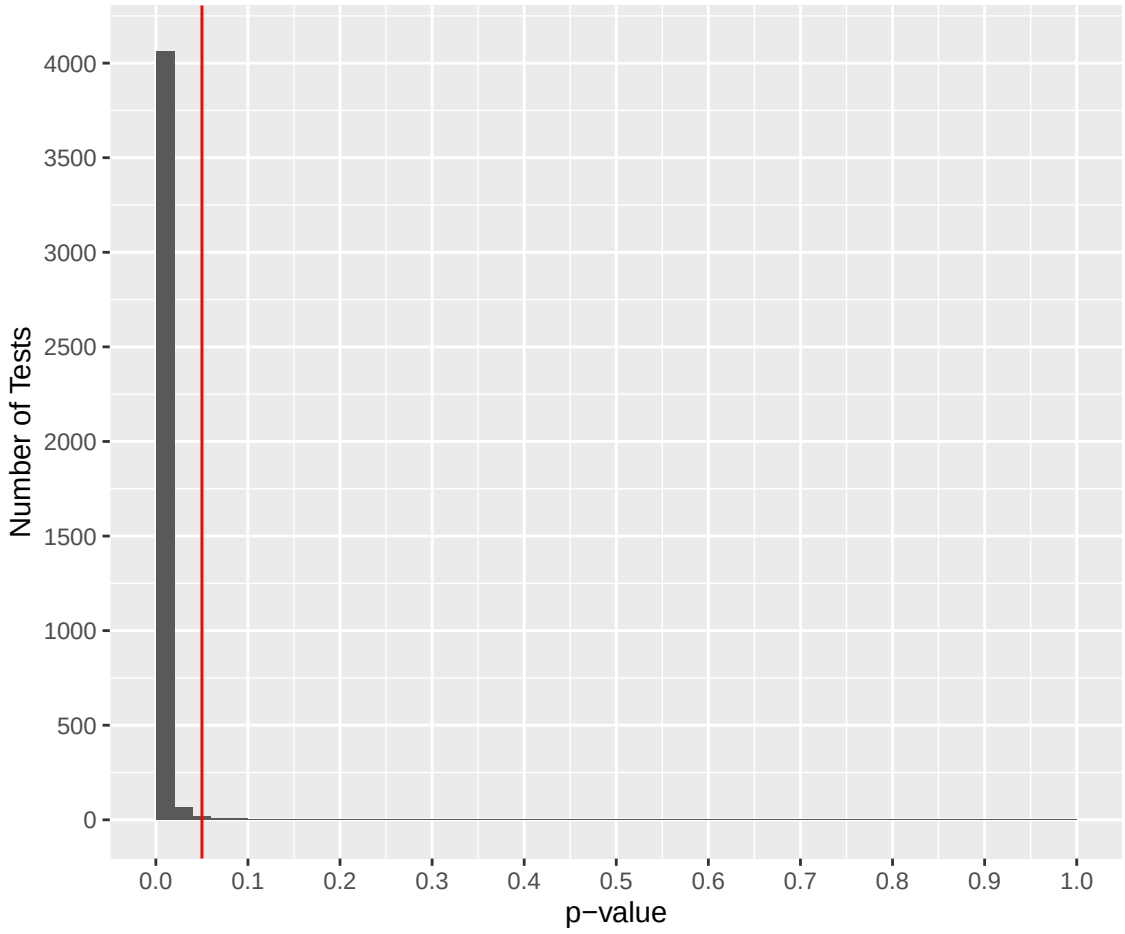


Fig. 56: A frequency histogram of p-values for the 4,180 pairwise modified Sryjala tests across all of the subjects within the USU Posture Study treatment group. The vertical red line at 0.05 indicates the significance level of the tests.

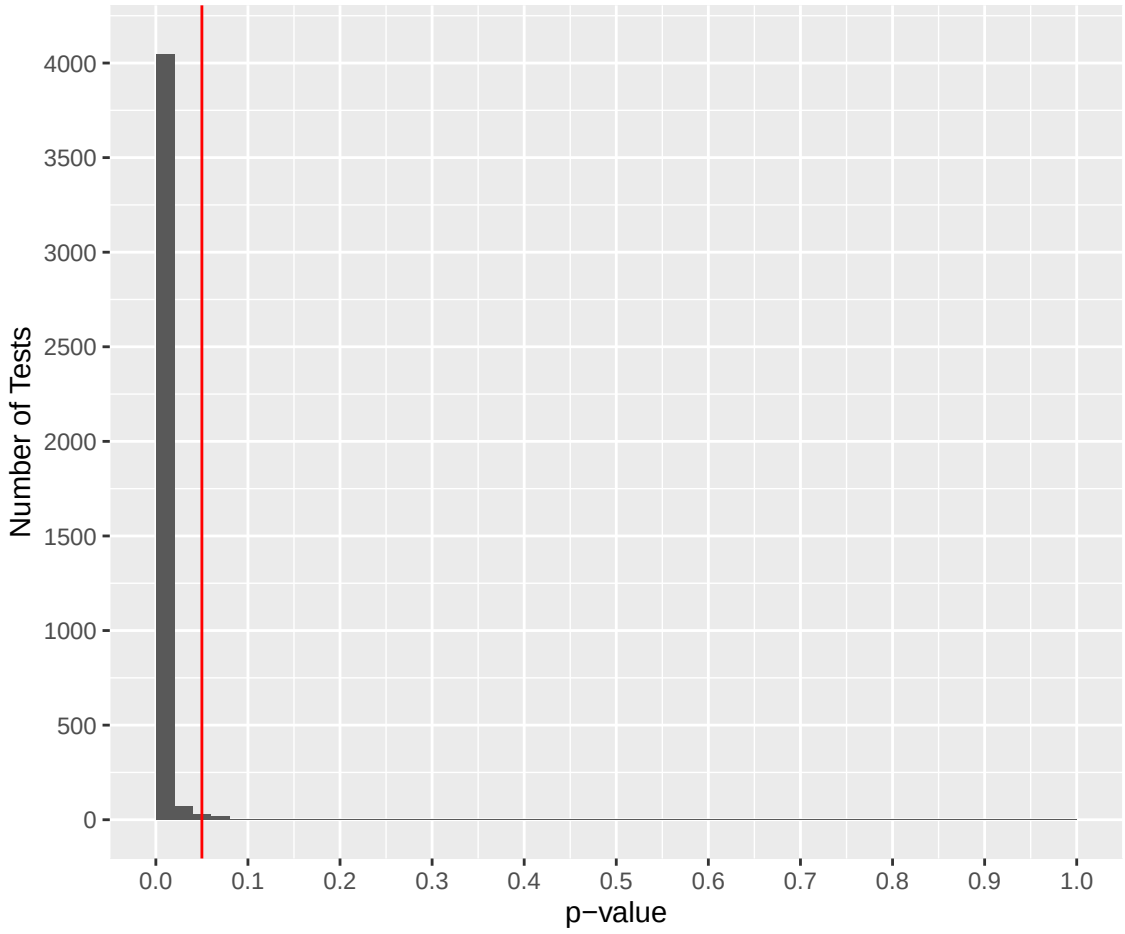


Fig. 57: A frequency histogram of p-values for the 4,180 pairwise modified Sryjala tests across all of the subjects within the USU Posture Study control group. The vertical red line at 0.05 indicates the significance level of the tests.

Table 19: All 35 non-significant results from the modified Syrjala test applied to all subject comparisons within the treatment group of the USU Posture Study. The remaining 4,145 tests for the treatment group yielded significant results with p-values of 0.05 or less. All computations were carried out on the University of Utah's Center for High Performance Computing (`https://chpc.utah.edu/`) on the Notchpeak cluster using 4 cores (see AMD in Table 17) and 128GB of RAM.

Subject 1 ID	Subject 2 ID	Posture ID	n_1	n_2	p-values	Comp. Time (in secs)
1	15	19	32	29	0.08	0.20
2	14	10	56	77	0.08	1.07
2	18	10	56	94	0.13	1.53
3	5	1	41	49	0.08	0.55
3	20	2	53	135	0.23	2.90
3	7	8	56	39	0.12	0.51
4	10	19	21	28	0.07	0.15
4	18	19	21	49	0.08	0.26
5	9	2	42	39	0.11	0.34
5	18	3	46	28	0.10	0.27
5	10	9	45	26	0.10	0.28
5	11	17	56	71	0.07	1.00
5	10	19	57	28	0.06	0.67
5	13	21	44	32	0.10	0.52
6	7	17	82	36	0.06	0.79
7	20	9	63	94	0.19	1.77
7	8	15	56	57	0.10	1.25
7	17	17	36	60	0.10	0.53
7	18	17	36	53	0.11	0.70
8	12	19	72	49	0.15	0.87
8	17	19	72	56	0.07	1.01
9	17	11	80	30	0.10	0.68
9	12	19	51	49	0.25	0.56
9	13	21	53	32	0.06	0.40
11	12	9	44	77	0.25	0.90
11	13	9	44	30	0.08	0.35
11	12	10	61	76	0.06	1.33
11	18	21	43	62	0.13	0.67
12	13	9	77	30	0.06	0.68
13	17	21	32	53	0.06	0.42
14	18	10	77	94	0.20	2.05
15	20	3	23	85	0.08	1.36
15	18	17	35	53	0.06	0.42
16	18	1	49	37	0.06	0.53
17	20	15	43	128	0.09	2.07

Table 20: All 54 non-significant results from the modified Syrjala test applied to all subject comparisons within the control group of the USU Posture Study. The remaining 4,126 tests for the control group yielded significant results with p-values of 0.05 or less. All computations were carried out on the University of Utah's Center for High Performance Computing (`https://chpc.utah.edu/`) on the Notchpeak cluster using 4 cores (see AMD in Table 17) and 128GB of RAM.

Subject 1 ID	Subject 2 ID	Posture ID	n_1	n_2	p-values	Comp. Time (in secs)
1	8	16	31	49	0.06	0.33
1	10	16	31	55	0.17	0.41
1	14	16	31	55	0.07	0.69
1	16	16	31	63	0.08	0.46
2	16	13	127	46	0.14	2.23
2	4	16	114	40	0.06	1.56
2	7	21	104	40	0.06	1.32
4	15	9	34	71	0.06	0.68
4	7	21	46	40	0.07	0.44
5	11	9	49	58	0.06	0.70
5	7	21	84	40	0.09	0.87
6	7	3	47	28	0.10	0.33
6	7	19	48	39	0.08	0.41
6	8	19	48	21	0.07	0.26
6	14	22	51	103	0.08	1.59
7	16	3	28	70	0.06	0.56
7	13	10	40	19	0.07	0.20
7	17	15	28	157	0.24	2.68
7	16	19	39	45	0.08	0.37
7	15	21	40	70	0.07	0.71
8	9	14	36	60	0.06	0.58
8	13	14	36	71	0.10	0.69
8	18	14	36	58	0.10	0.49
8	15	15	43	63	0.08	0.71
8	9	16	49	64	0.16	0.77
8	12	16	49	51	0.18	0.93
8	14	16	49	55	0.37	0.61
8	20	16	49	88	0.26	1.22
8	11	19	21	31	0.14	0.15
8	12	19	21	21	0.08	0.10
9	16	1	59	54	0.06	1.02
9	20	1	59	74	0.09	1.99
9	19	3	34	57	0.06	0.42
9	12	4	61	23	0.07	0.34
9	10	6	59	40	0.06	0.52
9	13	14	60	71	0.11	1.12
9	14	16	64	55	0.35	0.86
9	18	21	60	63	0.11	0.91
10	14	16	55	55	0.08	0.71
10	16	16	55	63	0.06	0.88
10	19	19	39	43	0.08	0.59
11	12	8	40	42	0.06	0.35
11	14	9	58	43	0.08	0.59
11	15	20	44	58	0.08	0.59
12	16	2	56	34	0.06	0.74
12	18	14	22	58	0.23	0.37
12	20	16	51	88	0.06	1.26
13	15	19	55	49	0.07	0.62
14	17	9	43	107	0.20	1.56
14	16	13	56	46	0.12	0.59
15	16	19	49	45	0.06	0.63
16	20	9	50	87	0.22	1.30
16	20	17	35	179	0.06	3.34
17	20	9	107	87	0.06	3.08

7.1.4 Within-group Comparison Examples

Sections 7.1.1 and 7.1.2 revealed an overwhelming majority of comparisons between subjects in the USU Posture Study were statistically significant. In this section, a closer look is made at a series of subject gaze point scatterplots. The results of the respective tests are justified using connections to the previous simulations studies detailed in Sections 6.5.3–6.5.10. Furthermore, these connections to previous simulations which generated data inspired by eye-tracking scenarios demonstrates the utility of the modified Syrjala tests in eye-tracking data analyses. Within each of the treatment and control groups test results, one comparison was randomly selected from the set of significant results and one from the non-significant results (four in total). Table 21 provides the details of the randomly chosen tests. For each comparison, the gaze point scatterplots are displayed in Figures 58–61.

Table 21: Four randomly chosen test results (one significant and one non-significant from the treatment and control groups, respectively) along with their respective sample sizes and computational times (in seconds). All computations were carried out on the University of Utah's Center for High Performance Computing (`https://chpc.utah.edu/`) on the Notchpeak cluster using 4 cores (see AMD in Table 17) and 128GB of RAM.

Group	Subj. 1 ID	Subj. 2 ID	Post. ID	n_1	n_2	p-values	Comp. Time (in secs)
Treatment	8	18	20	56	53	0.01	0.67
Control	3	4	22	61	44	0.01	0.67
Treatment	9	17	11	80	30	0.10	0.68
Control	8	12	19	21	21	0.08	0.10

Figure 58 compares the gaze point scatterplots for subject ID 8 (left) and subject ID 18 (right) within the treatment group for posture ID 20. The sample sizes were roughly similar (56 and 53, respectively), and the test result was significant (p-value = 0.01) for this comparison. Hence, the conclusion drawn from the test is there is

some unspecified difference in the gaze point distributions of these two subjects. Indeed, while both subjects spent time gathering visual information about the postural stability of the actor by looking at the right bicep, subject ID 8 focused on the crown / right shoulder, right forearm / left thigh, and right foot, whereas subject ID 18 looked at the left side of the head, core, between the legs, and left foot more.

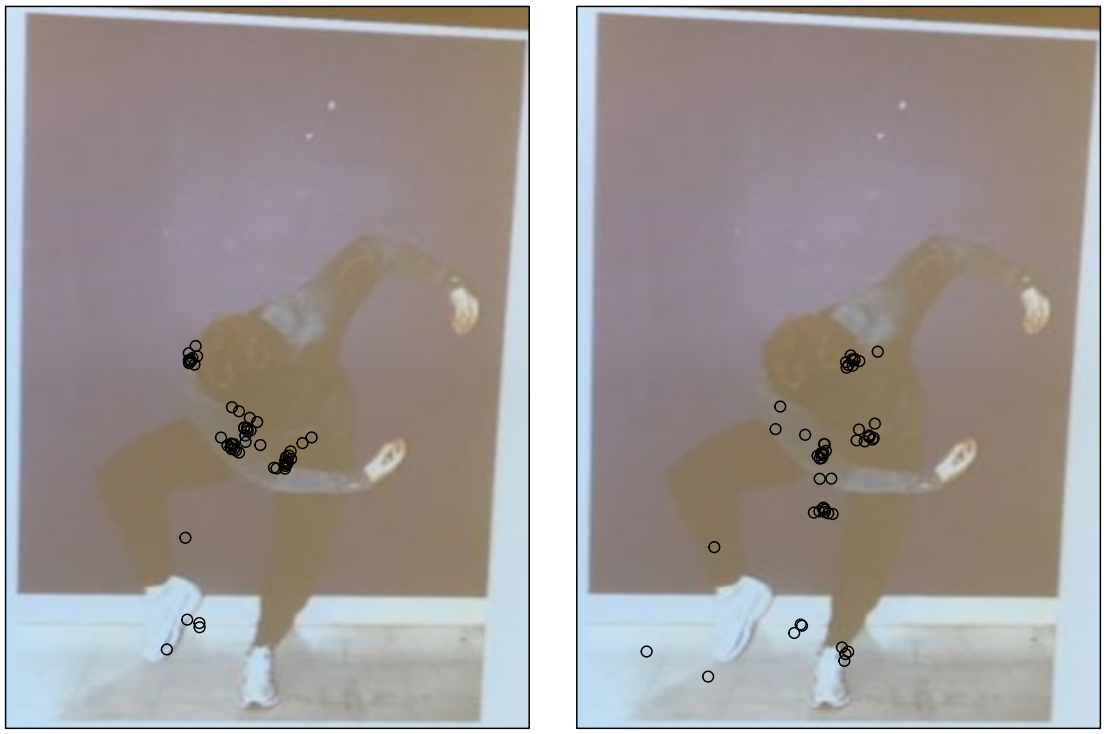


Fig. 58: Scatterplots of the gaze points for subject ID 8 (left) and subject ID 18 (right) within the treatment group for posture ID 20. The test result was significant (p-value = 0.01) for this comparison.

These differences are similar to several of the simulations carried out in Section 6.5. Particularly, the difference in where subjects looked at the head resembles a shift in bivariate mean of the clusters as studied in Section 6.5.3. The difference between right forearm and core also resembles a small difference in bivariate means. Hence, while the test is relatively robust against some of the noise seen around the

right foot of subject ID 18 (simulated in Section 6.5.7), there is enough difference between subject scatterplots to conclude a significant difference.

Furthermore, while 35 (out of 4,180) tests were non-significant for all of the pairwise comparisons between subjects within the treatment group, none of the remaining comparisons between subjects for Posture ID 20 (such as the comparison in Figure 58) were non-significant. Hence, all of the treatment subject's gaze point distributions exhibited some level of significant difference from all of the remaining distributions for Posture ID 20. The test results for Posture ID 20, along with the other overwhelming majority of significant test results, contribute to the conclusion of general heterogeneity among subjects within the treatment and control groups. This also implies that there simply may not be enough of an impact that practicing yoga actively for at least two times a week for at least three months may have on human subject's visual behavior when assessing postural stability.

Similarly, Figure 59 compares the gaze point scatterplots for subject ID 3 (left) and subject ID 4 (right) within the control group for posture ID 22. The sample sizes were 61 and 44, respectively. The test result was significant (p-value = 0.01) for this comparison. Hence, the conclusion drawn from the test is there is some unspecified difference in the gaze point distributions of these two subjects. This is visually confirmed by the gaze point clusters in the left subplot around the left forearm, inside of the right thigh, and between the feet, as compared to the clusters in the right subplot around the upper torso, right waist, and outside of the right thigh.

Several comparisons can be made between differences seen in Figure 59 and the simulated differences in Sections 6.5.3–6.5.10. For example, the shift of visual attention from the upper torso (right subplot) to the left forearm (left subplot) shows a change in the center and shape of the fixation distribution even more dramatic than

the shifts in bivariate mean coordinates or variance-covariance structure simulated in Sections 6.5.3–6.5.4. Hence, it is no surprise that these two subject's gaze point distributions resulted in a highly significant p-value.

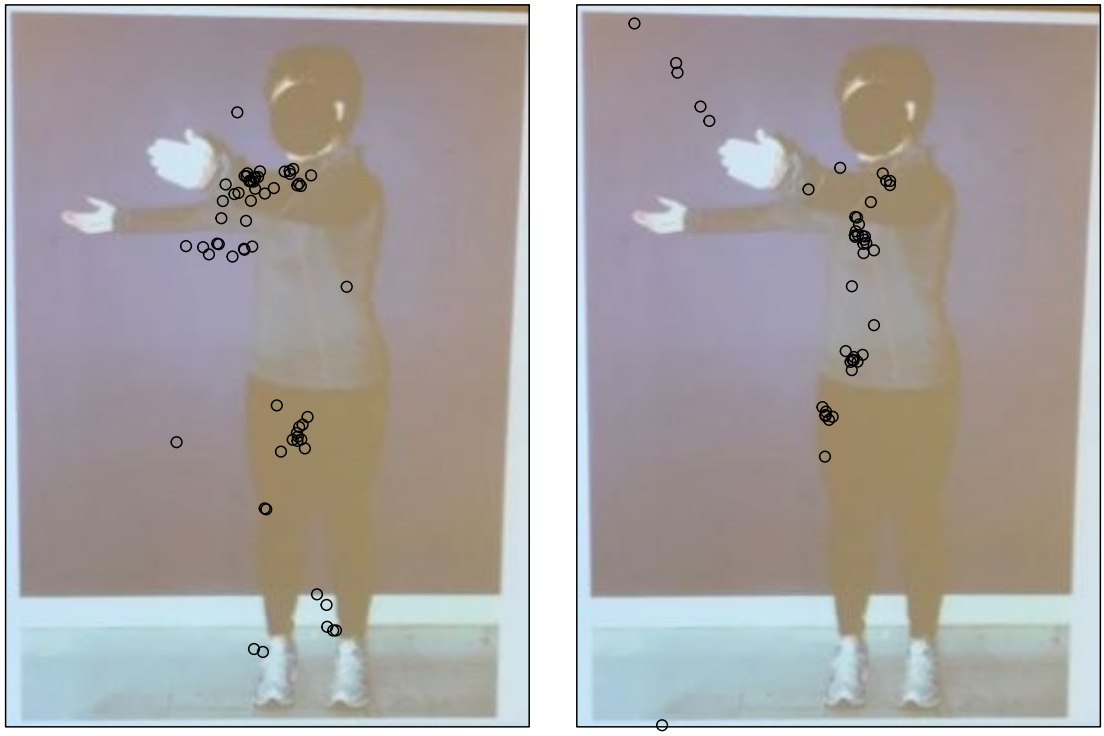


Fig. 59: Scatterplots of the gaze points for subject ID 3 (left) and subject ID 4 (right) within the control group for posture ID 22. The test result was significant (p-value = 0.01) for this comparison.

However, in rare instances, some of the subjects exhibited similarities in their gaze point scatterplots. For example, Figure 60 compares the gaze point scatterplots for subject ID 9 (left) and subject ID 17 (right) within the treatment group for posture ID 11. The test result was non-significant (p-value = 0.10) for this comparison. Hence, the null hypothesis cannot be rejected. There is not enough evidence to conclude that these subjects have different gaze point distributions.

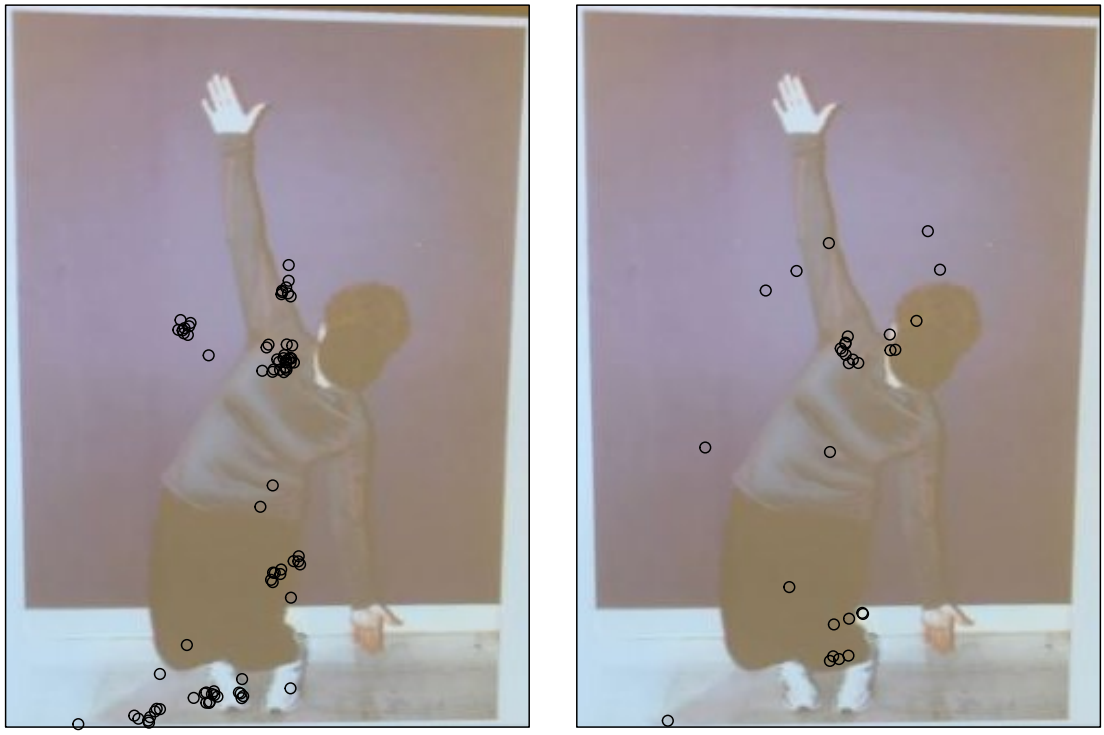


Fig. 60: Scatterplots of the gaze points for subject ID 9 (left) and subject ID 17 (right) within the treatment group for posture ID 11. The test result was non-significant (p-value = 0.10) for this comparison.

Initially, it may seem that this conclusion is counter-intuitive given the visual emphases made by the dark clusters in subject ID 9's scatterplot as compared to the lighter sparse clusters made by subject ID 17. However, notice that the sample sizes are relatively different. Subject ID 9 contributed 80 gaze points while subject ID 17 contributed only 30. While the test has been shown to be relatively robust to differences in sample sizes, the step sizes in bivariate cumulative distribution functions are smaller for large sample sizes and vice versa. Consequently, only a few gaze points are needed in posture ID 11's scatterplot near the locations of clusters in posture ID 9's scatterplot in order for the distribution functions to overlap and create small differences in the final statistic. Hence, this is a good example for emphasizing that

differences in sample sizes must be taken into consideration when visually comparing differences and similarities between the two gaze point scatterplots. This is confirmed by the simulations conducted in Section 6.5.10 where many non-significant test results are observed between two subject's simulated data which exhibit similar clusters across widely varying sample sizes. For example, there are many cases where one sample size is 25 and the other sample size is 100, and most of the test results are non-significant even for moderately different proportions of the sample sizes being allocated to different simulated fixation clusters.

Additionally, the data for a non-significant test result (p-value = 0.08) can be seen in Figure 61. Here, the gaze point scatterplots for subject ID 8 (left) and subject ID 12 (right) within the control group are compared for posture ID 19. The sample sizes were 21 for both of the subjects.

The conclusion drawn from this test is there is not enough evidence to conclude that these subjects have different gaze point distributions. Indeed, both subject's concentrated briefly on the right hand, stomach, and right knee. While the clusters around the stomach seem to exhibit slight differences in center and shape, simulations of varying cluster covariance structures in Section 6.5.4 indicate that the test is relatively tolerant of these changes, especially for small sample sizes. Additionally, the differences in these two clusters at the stomach are minimalized when considered as a part of the overall mixture distributions which form the entire gaze point distributions (see Sections 6.5.9 and 6.5.10). The two points between the feet in the right subject's scatterplot can be considered as outliers, which the test has demonstrated robustness against within the simulation carried out in Sections 6.5.6–6.5.8.

7.1.5 Conclusions from the USU Posture Study Analyses

In this chapter, the modified Syrjala test (which employs eight rotations, 0.1

Fig. 61: Scatterplots of the gaze points for subject ID 8 (left) and subject ID 12 (right) within the control group for posture ID 19. The test result was non-significant (p-value = 0.08) for this comparison.

proportion of toroidal shifts, and the CWS statistic) was used to make group-wise and within-group comparisons of the treatment and control subject gaze point data within the USU Posture Study. For the group-wise comparisons, an additional modification to the test (see Equation 10) enabled the comparisons to be made while treating each subject's contributions equally within their respective groups.

However, all of the group-wise test results are statistically significant. Furthermore, overwhelming majorities of the within-group tests for each group are also statistically significant. Specifically, approximately 99.16% and 98.71% of the pairwise tests were significant in the treatment and control groups, respectively. Due to the richness of significant tests within the respective groups, the usual application of a

multiple testing correction method is omitted, and a conclusion is made that the subjects within each respective group exhibit mostly heterogeneous gaze point patterns.

There are likely many reasons behind why so many significant differences were found between subject gaze distributions even within the treatment and control groups. Some of the likely sources of added gaze point variability could be the lack of precision in the ETMOBILE (`http://www.argusscience.com/ETMobile.html`) eye-tracking device, the algorithm (Li, 2017) which maps the gaze points from individual video frames to a master image, differences in subject eye physiology and presentation, and natural variation between human subjects. While a respectable piece of research equipment in its own time, the ETMOBILE (`http://www.argusscience.com/ETMobile.html`) eye-tracking device captures eye movement using only the subject's right eye, while other more modern eye trackers record both eyes and employ multiple forward facing cameras for additional gaze point location prediction accuracy. Additionally, eyeliner and mascara make-up along with contacts and certain types of glasses have been shown to negatively impact eye tracking device algorithms (Duchowski, 2007). Furthermore, during some of the preliminary data analyses of the subject data, many of the initial and final calibration gaze point distributions were shown to be noisy enough to conclude significant differences were present even though the subjects seemed to have focused their visual attention on the same four dots as instructed. Additional research should be conducted in each of these areas of additional variation to rule out any confounding effects.

CHAPTER 8

The `distdiffR` R Package

8.1 Overview

This chapter overviews the `distdiffR` package for the `R` computational environment (R Core Team, 2019), which was created to aid in distribution, reproducibility of results, and the ability to extend the functionality of the software through open-source availability. The package is publicly available on GitHub (`https://github.com/EricMcKinney77/distdiffR`).

Section 8.2 provides the vignette taken from the package which details the usage and functionality of the package. Documentation for all of the functions provided by the package is found in Section 8.3. While the package was developed using `R` version 4.1.1, the package only requires `R` version 3.5.0 or greater to be installed. To install the `distdiffR` package from GitHub the user has to install the `devtools` package (if not already installed), and then the user has to run the `install_github()` function in an R console as follows:

```
if (!require(devtools)) install.packages("devtools")
devtools::install_github("https://github.com/EricMcKinney77/distdiffR")
library(distdiffR)
```

8.2 Vignette for the `distdiffR` R Package

The distdiffR vignette

Eric McKinney

2022-04-14

Overview

`distdiffR` is an R package for bivariate two-sample tests of distributional equality.

The package provides a collection of nonparametric permutation tests for distributional equality. The tests make use of statistics between empirical cumulative distribution functions averaged across a series of rotations and / or toroidal shifts of the pooled samples. The variety of tests with their respective parameters can be called from the main function `distdiffr()`. It takes as input two bivariate samples (not necessarily the same size) in the form of two-column matrices.

Application (when no difference exists)

Below is an example using Anderson's Iris data[1] to show test results when the null hypotheses are true (and both distributions are equivalent). This is done by randomly assigning all three of the species of within the Iris data to two samples. Since `distdiffr()` employs bivariate tests of distributional equality, only the first two independent variables from the Iris data are used.

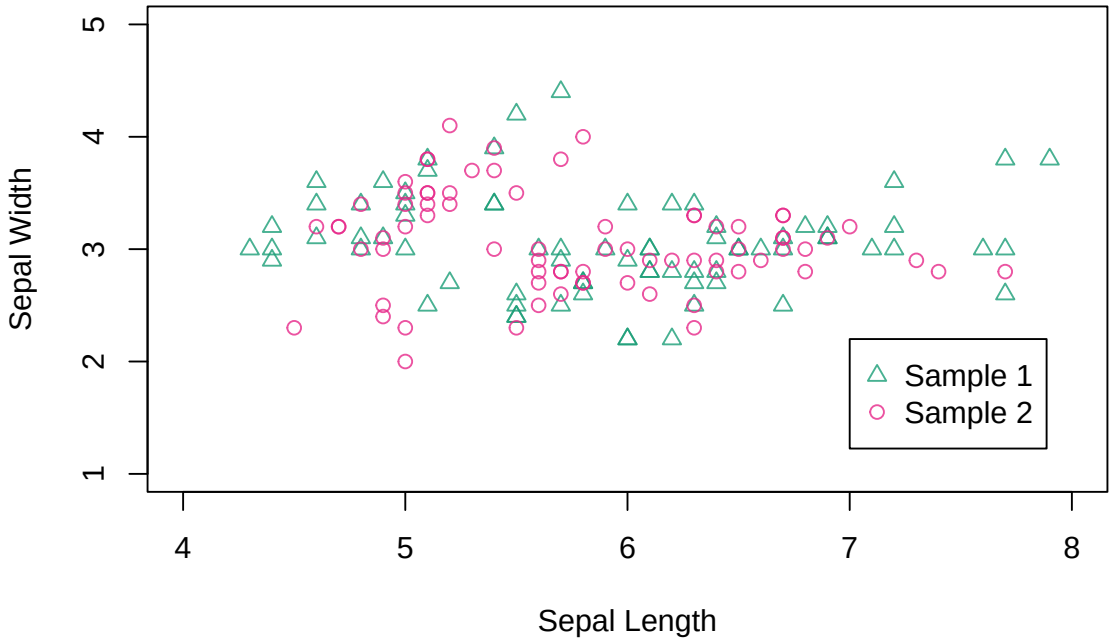
```
library(distdiffR)

seedNum <- 123
set.seed(seedNum)

data(iris)
# Randomly assign all three species to two samples
irisPermuted <- iris[sample.int(nrow(iris)), ]
sample1 <- as.matrix(irisPermuted[1:75, 1:2])
sample2 <- as.matrix(irisPermuted[76:150, 1:2])
pooled_data <- rbind(cbind(sample1, 1), cbind(sample2, 2))

plot(pooled_data[, 1],
     pooled_data[, 2],
     xlim = c(4, 8),
     ylim = c(1, 5),
     col = c("#1b9e77cc", "#e7298acc")[pooled_data[, 3]],
     pch = c(2, 1)[pooled_data[, 3]],
     pty = "s",
     xlab = "Sepal Length",
     ylab = "Sepal Width")
legend(7, 2.2,
       legend = c("Sample 1", "Sample 2"),
       pch = c(2, 1),
       col = c("#1b9e77cc", "#e7298acc"))
```

[1] Fisher, R. A., 1936. The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7, Part II, pp. 179–188.




```
# Rotational test
output <- distdiffr(sample1, # Note: Data inputs must be matrices
                    sample2,
                    testType = "rotational",
                    numRot = 8, # Default value
                    seedNum = seedNum)
output$pval
```

```
#> [1] 0.58
```

When `testType = "rotational"`, McKinney and Symanzik's rotational modified Syrjala test[2] is being employed. Since the p-value (`output$pval`) is much larger than any acceptable significance level, the null hypothesis is not rejected, and the conclusion is made that these two samples have been drawn from the same distribution.

Although simulations have suggested that the default number of rotations of eight (every 45 degrees) is sufficient, a different number may be passed to the `numRot` argument. For example, 48 rotations divides the statistic into a computation once every 7.5 degrees.

```
# Rotational test with 50 rotations
output <- distdiffr(sample1, # Note: Data inputs must be matrices
                    sample2,
                    testType = "rotational",
                    numRot = 48,
                    seedNum = seedNum)
output$pval
```

[2]McKinney, E., Symanzik, J., 2019. Modifications of the Syrjala Test for Testing Spatial Distribution Differences Between Two Populations, In: 2019 JSM Proceedings. American Statistical Association, Alexandria, VA. pp. 2518–2530.

```
#> [1] 0.566
```

Furthermore, while the default number of permutations (`numPerms`) is 999, a different number of permutations within any `testType` may be specified.

```
# Rotational test
output <- distdiffr(sample1, # Note: Data inputs must be matrices
                    sample2,
                    testType = "rotational",
                    numRot = 8, # Default value
                    numPerms = 9999,
                    seedNum = seedNum)
output$pval
```

```
#> [1] 0.5966
```

Test modifications

Similar results are also shown for the more powerful toroidal and combined (rotational and toroidal) modified Syrjala tests[3]:

```
# Toroidal shift test with proportions of points
output <- distdiffr(sample1,
                    sample2,
                    testType = "toroidal",
                    propPnts = 0.1,
                    seedNum = seedNum)
```

```
#> [1] "Using propPnts instead of default shiftThrshld."
```

```
output$pval
```

```
#> [1] 0.449
```

```
# Toroidal shift test with a threshold below pooled sample size
output <- distdiffr(sample1,
                    sample2,
                    testType = "toroidal",
                    shiftThrshld = 25, # Default
                    seedNum = seedNum)
output$pval
```

```
#> [1] 0.5
```

```
# Toroidal shift test with a threshold above pooled sample size
output <- distdiffr(sample1,
                    sample2,
                    testType = "toroidal",
                    shiftThrshld = 200,
                    seedNum = seedNum)
```

```
#> Warning in distdiffr(sample1, sample2, testType = "toroidal", shiftThrshld = 200,  :
#>   n_pooled is smaller than shiftThrshld. Can only compute n_pooled toroidal shifts.
```

```
output$pval
```

```
#> [1] 0.526
```

[3]McKinney, E., Symanzik, J., 2021. Extensions to the Syrjala Test with Eye-Tracking Analysis Applications, In: 2021 JSM Proceedings. American Statistical Association, Alexandria, VA. pp. 853–889.

```
# Toroidal shift test with a number of shifts
output <- distdiffr(sample1,
                    sample2,
                    testType = "toroidal",
                    numShifts = 8,
                    seedNum = seedNum)
```

```
#> Warning in distdiffr(sample1, sample2, testType = "toroidal", numShifts = 8, :
#> Using numShifts instead of default shiftThrshld.
```

```
output$pval
```

```
#> [1] 0.531
```

When employing the test which uses the combined rotational and toroidal shift modifications (which is the default argument for `testType`) the default behavior of the test is to use the 999 permutations, the CWS statistic (explained in more detail later in this vignette), eight rotations, and threshold the number of toroidal shifts to 25. If the combined sample size is less than the threshold, then the test will compute one toroidal shift per point (for each rotation). This is the default behavior when no other parameters are passed to the `distdiffr()` function other than the required sample matrices.

```
# Combined rotational and toroidal shift test
output <- distdiffr(sample1,
                    sample2,
                    testType = "combined", # Default
                    numRot = 8,            # Default
                    shiftThrshld = 25,     # Default
                    numPerms = 999,        # Default
                    psiFun = CalcPsiCWS,   # Default
                    seedNum = seedNum)
output$pval
```

```
#> [1] 0.331
```

```
# Same as above
output <- distdiffr(sample1,
                    sample2,
                    seedNum = seedNum)
output$pval
```

```
#> [1] 0.331
```

Alternatively, a proportion of the combined sample size may be specified for the test to determine the number of toroidal shifts (similar to the non-rotational toroidal shift test). If the proportion times the combined sample size is not an integer, the ceiling is taken to specify the number of toroidal shifts per rotation. Here, the proportion 0.1 multiplied to the combined sample size of 150 will result in 15 toroidal shifts per rotation. This will override the default behavior to use the `shiftThrshld` argument to limit the number of toroidal shifts.

```
# Combined rotational and toroidal shift test
output <- distdiffr(sample1,
                    sample2,
                    testType = "combined", # Default
                    numRot = 8,            # Default
                    propPnts = 0.1,
                    seedNum = seedNum)
```

```
#> [1] "Using propPnts instead of default shiftThrshld."
```

```
output$pval
```

```
#> [1] 0.321
```

Or, a specific number of toroidal shifts my be passed to `numShifts`. This will also override the default behavior to use the `shiftThrshld` argument to limit the number of toroidal shifts.

```
# Combined rotational and toroidal shift test
output <- distdiffr(sample1,
                    sample2,
                    testType = "combined", # Default
                    numRot = 8,            # Default
                    numShifts = 10,
                    seedNum = seedNum)
```

```
#> Warning in distdiffr(sample1, sample2, testType = "combined", numRot = 8, :
#> Using numShifts instead of default shiftThrshld.
```

```
output$pval
```

```
#> [1] 0.336
```

However, the number of toroidal shifts must be less than the combined sample size.

```
# Error: number of shifts larger than the combined sample sizes!
output <- distdiffr(sample1,
                    sample2,
                    testType = "combined", # Default
                    numRot = 8,            # Default
                    numShifts = 151,
                    seedNum = seedNum)
```

```
#> Warning in distdiffr(sample1, sample2, testType = "combined", numRot = 8,  :
#> Using numShifts instead of default shiftThrshld.
```

```
#> Error in distdiffr(sample1, sample2, testType = "combined", numRot = 8,  :
#>   n_pooled is smaller than shiftThrshld. Can only compute n_pooled toroidal shifts.
```

Also, `distdiffr()` will not allow arguments to be passed to more than one of `propPnts` or `numShifts`.

```
# Error: Must provide either propPnts or numShifts, but not both.
output <- distdiffr(sample1,
                    sample2,
                    testType = "combined", # Default
                    numRot = 8,            # Default
                    propPnts = 0.1,
                    numShifts = 10,
                    seedNum = seedNum)
```

```
#> Error in distdiffr(sample1, sample2, testType = "combined", numRot = 8, :
#>   Must provide either propPnts or numShifts, but not both.
```

Specifying a different number of rotations may also be combined with the above options for toroidal shifts.

```
output <- distdiffr(sample1,
                    sample2,
                    testType = "combined",
                    numRot = 10,
                    shiftThrshld = 25,
```

```
                    seedNum = seedNum)
output$pval
```

```
#> [1] 0.532
```

```
output <- distdiffr(sample1,
                    sample2,
                    testType = "combined",
                    numRot = 10,
                    propPnts = 0.1,
                    seedNum = seedNum)
```

```
#> [1] "Using propPnts instead of default shiftThrshld."
```

```
output$pval
```

```
#> [1] 0.548
```

```
output <- distdiffr(sample1,
                    sample2,
                    testType = "combined",
                    numRot = 10,
                    numShifts = 10,
                    seedNum = seedNum)
```

```
#> Warning in distdiffr(sample1, sample2, testType = "combined", numRot = 10, :
#> Using numShifts instead of default shiftThrshld.
```

```
output$pval
```

```
#> [1] 0.479
```

Alternative test statistics

Six alternative statistics are available for each of the previously discussed types of tests (rotational, toroidal, or combined). The six statistics can be accessed by passing one of CalcPsiDWS, CalcPsiUWS, CalcPsiCWS, CalcPsiDWA, CalcPsiUWA, or CalcPsiCWA to the `psiFun` argument. The abbreviations DWS, UWS, CWS, DWA, UWA, and CWA refer to the different computations taking place on the differences between the two sample's bivariate empirical cumulative distribution functions. The DW, UW, and CW mean that the differences are being double weighted, uniformly weighted, or complimentary weighted, respectively, and the appended S and A refer to the squared exponent or absolute value being applied to the differences. More details can be found in Section 5.1 of McKinney (2022)[4]. The default statistic is CWS. However, as seen here, the choice among these statistics has been shown to make little difference on the test results[5]. Consequently, the p-values within the S or A series are identical for the same random number seed.

```
output <- distdiffr(sample1,
                    sample2,
                    testType = "combined",
                    psiFun = CalcPsiDWS,
                    seedNum = seedNum)
output$psiStat
```

```
#> [1] 0.3131213
```

[4]McKinney, E., 2022. Extensions to the Syrjala Test with Eye-Tracking Data Analysis Applications in R. Ph.D. dissertation, Department of Mathematics and Statistics, Utah State University (Forthcoming)

[5]McKinney, E., Symanzik, J., 2021. Extensions to the Syrjala Test with Eye-Tracking Analysis Applications, In: 2021 JSM Proceedings. American Statistical Association, Alexandria, VA. pp. 853–889.

```
output$pval
```

```
#> [1] 0.331
```

```
output <- distdiffr(sample1,
                    sample2,
                    testType = "combined",
                    psiFun = CalcPsiUWS,
                    seedNum = seedNum)
output$psiStat
```

```
#> [1] 0.6262427
```

```
output$pval
```

```
#> [1] 0.331
```

```
output <- distdiffr(sample1,
                    sample2,
                    testType = "combined",
                    psiFun = CalcPsiCWS, # Default
                    seedNum = seedNum)
output$psiStat
```

```
#> [1] 0.3131213
```

```
output$pval
```

```
#> [1] 0.331
```

```
output <- distdiffr(sample1,
                    sample2,
                    testType = "combined",
                    psiFun = CalcPsiDWA,
                    seedNum = seedNum)
output$psiStat
```

```
#> [1] 3.7387
```

```
output$pval
```

```
#> [1] 0.333
```

```
output <- distdiffr(sample1,
                    sample2,
                    testType = "combined",
                    psiFun = CalcPsiUWA,
                    seedNum = seedNum)
output$psiStat
```

```
#> [1] 7.4774
```

```
output$pval
```

```
#> [1] 0.333
```

```
output <- distdiffr(sample1,
                    sample2,
                    testType = "combined",
                    psiFun = CalcPsiCWA,
```

```
                     seedNum = seedNum)
output$psiStat
```

```
#> [1] 3.7387
```

```
output$pval
```

```
#> [1] 0.333
```

Input order does not matter

Additionally, for any of the above tests, the data input order is arbitrary, e.g.,

```
output1 <- distdiffr(sample1,
                     sample2,
                     seedNum = seedNum)

output2 <- distdiffr(sample2,
                     sample1,
                     seedNum = seedNum)

output1$pval == output2$pval
```

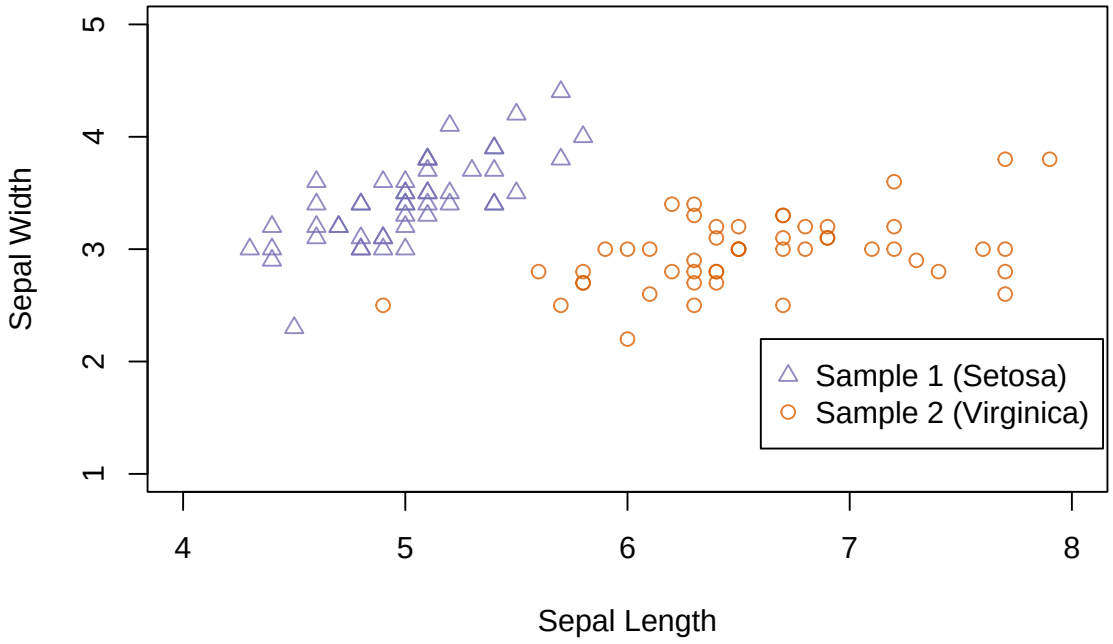
```
#> [1] TRUE
```

Application (when a difference exists)

The following examples demonstrate test results when the null hypothesis is false (i.e., there exists some difference between the two distributions). This is shown by separating the two samples by the species Setosa and Virginica, respectively.

```
data(iris)
sample1 <- as.matrix(iris[iris[5] == "setosa", -(3:5)])
sample2 <- as.matrix(iris[iris[5] == "virginica", -(3:5)])
pooled_data <- rbind(cbind(sample1, 1), cbind(sample2, 2))

plot(pooled_data[, 1],
     pooled_data[, 2],
     xlim = c(4, 8),
     ylim = c(1, 5),
     col = c("#7570b3cc", "#d95f02cc")[pooled_data[, 3]],
     pch = c(2, 1)[pooled_data[, 3]],
     pty = "s",
     xlab = "Sepal Length",
     ylab = "Sepal Width")
legend(6.6, 2.2,
       legend = c("Sample 1 (Setosa)", "Sample 2 (Virginica)"),
       pch = c(2, 1),
       col = c("#7570b3cc", "#d95f02cc"))
```



Indeed, the difference between the two samples results in minimal p-values among all of the test types.

```
# Rotational test
output <- distdiffr(sample1,
                    sample2,
                    testType = "rotational",
                    seedNum = seedNum)
output$pval
```

```
#> [1] 0.001
```

```
# Toroidal shift test with proportions of points
output <- distdiffr(sample1,
                    sample2,
                    testType = "toroidal",
                    propPnts = 0.1,
                    seedNum = seedNum)
```

```
#> [1] "Using propPnts instead of default shiftThrshld."
```

```
output$pval
```

```
#> [1] 0.001
```

```
# Toroidal shift test with thresholds below pooled sample size
output <- distdiffr(sample1,
                    sample2,
                    testType = "toroidal",
                    shiftThrshld = 25, # Default
```

```
                    seedNum = seedNum)
output$pval
```

```
#> [1] 0.001
```

```
# Toroidal shift test with thresholds above pooled sample size
output <- distdiffr(sample1,
                    sample2,
                    testType = "toroidal",
                    shiftThrshld = 200,
                    seedNum = seedNum)
```

```
#> Warning in distdiffr(sample1, sample2, testType = "toroidal", shiftThrshld = 200,  :
#>   n_pooled is smaller than shiftThrshld. Can only compute n_pooled toroidal shifts.
```

```
output$pval
```

```
#> [1] 0.001
```

```
# Toroidal shift test with a number of shifts
output <- distdiffr(sample1,
                    sample2,
                    testType = "toroidal",
                    numShifts = 8,
                    seedNum = seedNum)
```

```
#> Warning in distdiffr(sample1, sample2, testType = "toroidal", numShifts = 8, :
#> Using numShifts instead of default shiftThrshld.
```

```
output$pval
```

```
#> [1] 0.001
```

Again, the default test type is the combined rotational and toroidal shift test, with eight rotations and a threshold of 25 toroidal shifts as default values. These default settings are usually adequate to obtain meaningful results for both cases of when the null hypothesis is true (equal distributions) and when the null hypothesis is false (unequal distributions).

```
# Combined rotational and toroidal shift test
output <- distdiffr(sample1,
                    sample2,
                    testType = "combined", # Default
                    numRot = 8,            # Default
                    shiftThrshld = 25,     # Default
                    seedNum = seedNum)
output$pval
```

```
#> [1] 0.001
```

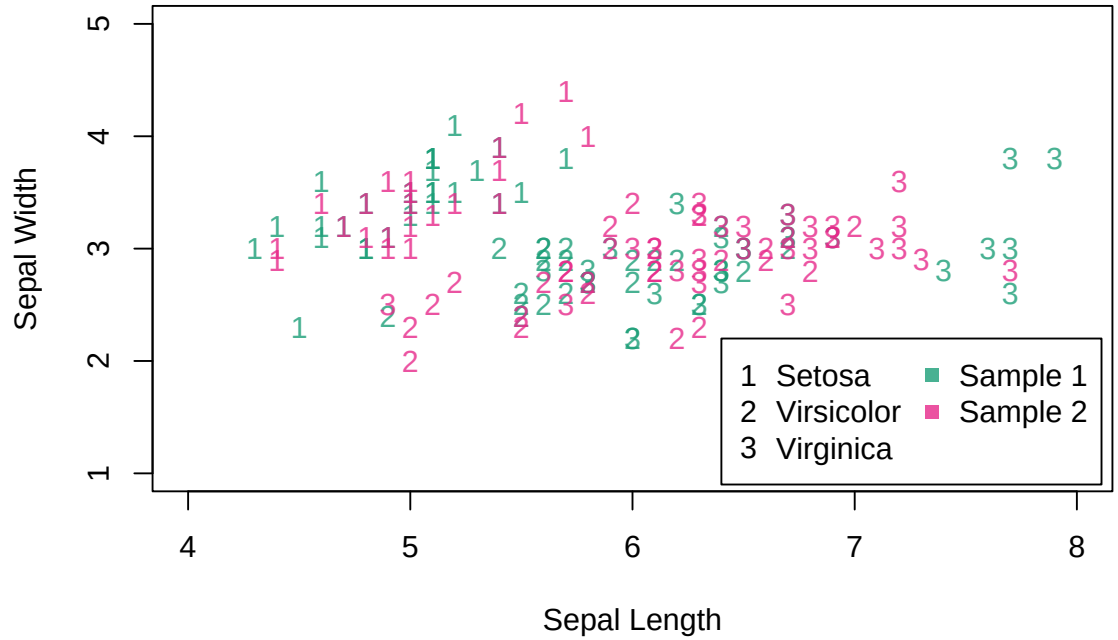
The grouped_distdiffr() test

Another version of the test also exists when combining bivariate data from multiple sources (e.g., subjects) into each of the two-samples, respectively. This test treats each subject's contribution equally. It can be called via the `grouped_distdiffr()` function. The plot below labels the species Setosa, Virsicolor, and Virginica as the integers 1, 2, and 3, respectively. Since the species is treated as a subject labeling, then each subject's contributions to the respective samples can be grouped and weighted such that each contribution is

treated equally. Section 7.1.1 of McKinney (2022)[6] describes the underlying mathematics in greater detail.

```
# Randomly assign all three species to two samples
iris$Species <- rep(1:3, each = 50)
irisPermuted <- iris[sample.int(nrow(iris)), ]
sample1 <- as.matrix(irisPermuted[1:75, c(1:2, 5)])
sample2 <- as.matrix(irisPermuted[76:150, c(1:2, 5)])
pooled_data <- rbind(cbind(sample1, 1), cbind(sample2, 2))

plot(pooled_data[, 1],
     pooled_data[, 2],
     xlim = c(4, 8),
     ylim = c(1, 5),
     col = c("#1b9e77cc", "#e7298acc")[pooled_data[, 4]],
     pch = c("1", "2", "3")[pooled_data[, 3]],
     pty = "s",
     xlab = "Sepal Length",
     ylab = "Sepal Width")
legend(6.4, 2.2,
       legend = c("Setosa", "Virsicolor", "Virginica", "Sample 1", "Sample 2"),
       pch = c(49, 50, 51, 15, 15),
       col = c("black", "black", "black", "#1b9e77cc", "#e7298acc"),
       ncol = 2)
```



[6]McKinney, E., 2022. Extensions to the Syrjala Test with Eye-Tracking Data Analysis Applications in R. Ph.D. dissertation, Department of Mathematics and Statistics, Utah State University (Forthcoming)

```
table(sample1[, "Species"])
```

```
#>
#>  1  2  3
#> 28 23 24
```

```
table(sample2[, "Species"])
```

```
#>
#>  1  2  3
#> 22 27 26
```

For example, although `sample1` has 28 Setosas, 23 Virsicolors, and 24 Virginicas, the contributions of each to the overall test statistic will be weighted equally. This is also true of `sample2` which has 22 Setosas, 27 Virsicolors, and 26 Virginicas as seen in the above table outputs.

```
output <- grouped_distdiffr(sample1,
                            sample2,
                            seedNum = seedNum)
output$pval
```

```
#> [1] 0.268
```

8.3 Documentation for the `distdiffR` `R` Package

This section provides the documentation for all of the functions within the `distdiffR` package.

Package 'distdiffR'

April 13, 2022

Title Two-Sample Tests of Distributional Equality for Bivariate Data

Version 0.1.0

Author Eric McKinney

Description A collection of bivariate two-sample tests for distributional equality are provided. The tests make use of statistics between empirical distribution functions averaged across a series of rotations and or toroidal shifts of the pooled samples. Another version of the test also exists when combining bivariate data from multiple sources into each of the two-samples, respectively, which treats each contribution equally.

License MIT + file LICENSE

Encoding UTF-8

LazyData true

Roxygen list(markdown = TRUE)

RoxygenNote 7.1.2

LinkingTo Rcpp

Imports Rcpp,
stats,
Rdpack

RdMacros Rdpack

SystemRequirements GNU make

Suggests rmarkdown,
knitr

VignetteBuilder knitr

R topics documented:

`bcdf` *Construct and evaluate a bivariate empirical cumulative distribution function*

Description

Construct a bivariate empirical cumulative distribution function (BECDF) using `data` and pass each of the `eval` points through the BECDF.

Usage

```
bcdf(data, eval)
```

Arguments

`data`	A two column matrix for constructing the BECDF
`eval`	A two column matrix for input into the BECDF

Value

A numeric vector of output values from the BECDF

Examples

```
data(iris)
sample1 <- as.matrix(iris[iris$Species == "virginica", 1:2])
sample2 <- as.matrix(iris[iris$Species == "versicolor", 1:2])

bcdf(sample1, sample2)
```

`CalcGroupPsiCWS` *The Psi CWS statistic for aggregated group data*

Description

This statistic computes the complementary weighted squared (CWS) differences between the averaged subject empirical cumulative distribution functions for the two samples. For more information, see McKinney (2022) and McKinney and Symanzik (2021).

Usage

```
CalcGroupPsiCWS(data, groups, subjects)
```

Arguments

data	A two column matrix of the bivariate pooled samples
groups	A numeric vector of sample (or group) labels (use either 1 or 2)
subjects	A numeric vector of subject labels

Value

The Psi CWS statistic for aggregated group data

References

McKinney E (2022). *Extensions to the Syrjala Test with Eye-Tracking Data Analysis Applications in R*. Ph.D. dissertation, Department of Mathematics and Statistics, Utah State University. (Forthcoming).

McKinney E, Symanzik J (2021). "Extensions to the Syrjala Test with Eye-Tracking Analysis Applications." In *2021 JSM Proceedings.*, 853–889. American Statistical Association, Alexandria, VA.

Examples

```
# Randomly assign all three species to two samples
data(iris)
iris$Species <- rep(1:3, each = 50) # Species will serve as the subject label
irisPermuted <- iris[sample.int(nrow(iris)), ]
sample1 <- as.matrix(irisPermuted[1:75, c(1:2, 5)])
sample2 <- as.matrix(irisPermuted[76:150, c(1:2, 5)])
pooled_data <- rbind(cbind(sample1, 1), cbind(sample2, 2))

CalcGroupPsiCWS(pooled_data[, 1:2], pooled_data[, 4], pooled_data[, 3])
```

CalcPsiCWA *The Psi CWA Statistic*

Description

This statistic computes the complementary weighted absolute (CWA) differences between the empirical cumulative distribution functions for the two samples. For more information, see McKinney (2022) and McKinney and Symanzik (2021).

Usage

```
CalcPsiCWA(data, subjects)
```

Arguments

data	A two column matrix of the bivariate pooled samples
subjects	A numerical vector of sample labels (use either 1 or 2)

Value

the Psi CWA statistic

References

McKinney E (2022). *Extensions to the Syrjala Test with Eye-Tracking Data Analysis Applications in R*. Ph.D. dissertation, Department of Mathematics and Statistics, Utah State University. (Forthcoming).

McKinney E, Symanzik J (2021). "Extensions to the Syrjala Test with Eye-Tracking Analysis Applications." In *2021 JSM Proceedings.*, 853–889. American Statistical Association, Alexandria, VA.

Examples

```
data(iris)
pooled_data <- iris[iris$Species %in% c("setosa", "virginica"), 1:2]
sample_labels <- rep(1:2, c(sum(iris$Species == "setosa"),
                            sum(iris$Species == "virginica")))

CalcPsiCWA(as.matrix(pooled_data), sample_labels)
```

`CalcPsiCWS` *The Psi CWS Statistic*

Description

This statistic computes the complementary weighted squared (CWS) differences between the empirical cumulative distribution functions for the two samples. For more information, see McKinney (2022) and McKinney and Symanzik (2021).

Usage

```
CalcPsiCWS(data, subjects)
```

Arguments

`data`	A two column matrix of the bivariate pooled samples
`subjects`	A numerical vector of sample labels (use either 1 or 2)

Value

The Psi CWS statistic

References

McKinney E (2022). *Extensions to the Syrjala Test with Eye-Tracking Data Analysis Applications in R*. Ph.D. dissertation, Department of Mathematics and Statistics, Utah State University. (Forthcoming).

McKinney E, Symanzik J (2021). "Extensions to the Syrjala Test with Eye-Tracking Analysis Applications." In *2021 JSM Proceedings.*, 853–889. American Statistical Association, Alexandria, VA.

Examples

```
data(iris)
pooled_data <- iris[iris$Species %in% c("setosa", "virginica"), 1:2]
sample_labels <- rep(1:2, c(sum(iris$Species == "setosa"),
                            sum(iris$Species == "virginica")))

CalcPsiCWS(as.matrix(pooled_data), sample_labels)
```

`CalcPsiDWA` *The Psi DWA Statistic*

Description

This statistic computes the double weighted absolute (DWA) differences between the empirical cumulative distribution functions for the two samples. For more information, see McKinney (2022) and McKinney and Symanzik (2021).

Usage

```
CalcPsiDWA(data, subjects)
```

Arguments

`data`	A two column matrix of the bivariate pooled samples
`subjects`	A numerical vector of sample labels (use either 1 or 2)

Value

The Psi DWA statistic

References

McKinney E (2022). *Extensions to the Syrjala Test with Eye-Tracking Data Analysis Applications in R*. Ph.D. dissertation, Department of Mathematics and Statistics, Utah State University. (Forthcoming).

McKinney E, Symanzik J (2021). "Extensions to the Syrjala Test with Eye-Tracking Analysis Applications." In *2021 JSM Proceedings.*, 853–889. American Statistical Association, Alexandria, VA.

Examples

```
data(iris)
pooled_data <- iris[iris$Species %in% c("setosa", "virginica"), 1:2]
sample_labels <- rep(1:2, c(sum(iris$Species == "setosa"),
                            sum(iris$Species == "virginica")))

CalcPsiDWA(as.matrix(pooled_data), sample_labels)
```

CalcPsiDWS *The Psi DWS Statistic*

Description

This statistic computes the double weighted squared (DWS) differences between the empirical cumulative distribution functions for the two samples. For more information, see McKinney (2022) and McKinney and Symanzik (2021).

Usage

```
CalcPsiDWS(data, subjects)
```

Arguments

data	A two column matrix of the bivariate pooled samples
subjects	A numerical vector of sample labels (use either 1 or 2)

Value

The Psi DWS statistic

References

McKinney E (2022). *Extensions to the Syrjala Test with Eye-Tracking Data Analysis Applications in R*. Ph.D. dissertation, Department of Mathematics and Statistics, Utah State University. (Forthcoming).

McKinney E, Symanzik J (2021). "Extensions to the Syrjala Test with Eye-Tracking Analysis Applications." In *2021 JSM Proceedings.*, 853–889. American Statistical Association, Alexandria, VA.

Examples

```
data(iris)
pooled_data <- iris[iris$Species %in% c("setosa", "virginica"), 1:2]
sample_labels <- rep(1:2, c(sum(iris$Species == "setosa"),
                            sum(iris$Species == "virginica")))

CalcPsiDWS(as.matrix(pooled_data), sample_labels)
```

CalcPsiUWA *The Psi UWA Statistic*

Description

This statistic computes the uniformly weighted absolute (UWA) differences between the empirical cumulative distribution functions for the two samples. For more information, see McKinney (2022) and McKinney and Symanzik (2021).

Usage

```
CalcPsiUWA(data, subjects)
```

Arguments

data	A two column matrix of the bivariate pooled samples
subjects	A numerical vector of sample labels (use either 1 or 2)

Value

The Psi UWA statistic

References

McKinney E (2022). *Extensions to the Syrjala Test with Eye-Tracking Data Analysis Applications in R*. Ph.D. dissertation, Department of Mathematics and Statistics, Utah State University. (Forthcoming).

McKinney E, Symanzik J (2021). "Extensions to the Syrjala Test with Eye-Tracking Analysis Applications." In *2021 JSM Proceedings.*, 853–889. American Statistical Association, Alexandria, VA.

Examples

```
data(iris)
pooled_data <- iris[iris$Species %in% c("setosa", "virginica"), 1:2]
sample_labels <- rep(1:2, c(sum(iris$Species == "setosa"),
                            sum(iris$Species == "virginica")))

CalcPsiUWA(as.matrix(pooled_data), sample_labels)
```

CalcPsiUWS *The Psi UWS Statistic*

Description

This statistic computes the uniformly weighted squared (UWS) differences between the empirical cumulative distribution functions for the two samples. For more information, see McKinney (2022) and McKinney and Symanzik (2021).

Usage

```
CalcPsiUWS(data, subjects)
```

Arguments

data	A two column matrix of the bivariate pooled samples
subjects	A numerical vector of sample labels (use either 1 or 2)

Value

The Psi UWS statistic

References

McKinney E (2022). *Extensions to the Syrjala Test with Eye-Tracking Data Analysis Applications in R*. Ph.D. dissertation, Department of Mathematics and Statistics, Utah State University. (Forthcoming).

McKinney E, Symanzik J (2021). "Extensions to the Syrjala Test with Eye-Tracking Analysis Applications." In *2021 JSM Proceedings.*, 853–889. American Statistical Association, Alexandria, VA.

Examples

```
data(iris)
pooled_data <- iris[iris$Species %in% c("setosa", "virginica"), 1:2]
sample_labels <- rep(1:2, c(sum(iris$Species == "setosa"),
                            sum(iris$Species == "virginica")))

CalcPsiUWS(as.matrix(pooled_data), sample_labels)
```

`distdiffr` *The distdiffR two-sample tests of bivariate distributional equality*

Description

The `distdiffr()` function conducts two-sample permutation tests of distributional equality based on differences in the bivariate empirical cumulative density functions (BECDFs). The differences in BECDFs are computed across a series of rotations, toroidal shifts, or both rotations and toroidal shifts of the combined data (specified via `testType`). The number of rotations and toroidal shifts may be specified (via `numRot` or `numShifts`, respectively). The number of toroidal shifts may also be determined by a proportion of the combined sample size (via `propPnts`). However, McKinney (2022) has shown that limiting the number of toroidal shifts to ease the computational load of the test will still provide stable results. Simulations have shown the combined rotational and toroidal shift test to be the most powerful yet appropriately conservative test. For more information, see McKinney (2022) and McKinney and Symanzik (2021).

Usage

```
distdiffr(
  data1,
  data2,
  testType = "combined",
  numRot = 8,
  propPnts = NULL,
  numShifts = NULL,
  shiftThrshld = 25,
  numPerms = 999,
  psiFun = CalcPsiCWS,
  seedNum = NULL
)
```

Arguments

data1	A two column matrix of bivariate observations from one sample.
data2	A two column matrix of bivariate observations from another sample.
testType	A string indicating the type of test to be used. Must be one of c("rotational", "toroidal", "combined").
numRot	An integer number of rotational shifts of the pooled samples.
propPnts	A numeric proportion of points to be used as toroidal shift origins. Cannot provide both propPnts and numShifts. If neither are provided, shiftThrshld is used.
numShifts	A numeric integer. The number of points to be used as toroidal shift origins. Must be less than the pooled sample size. Cannot provide both propPnts and numShifts. If neither are provided, shiftThrshld is used.
shiftThrshld	A numeric integer. Used if neither propPnts or numShifts are provided. If the pooled sample size is less than shiftThrshld, every point will be used as a toroidal shift origin. Otherwise, only a random sample of shiftThrshld points will be used.
numPerms	An integer number of permutations of the original data.
psiFun	A function specifying the Psi statistic calculation.
seedNum	An integer random seed value.

Value

A list including three objects: (1) the Psi statistic computed on the original data (2) a vector of Psi statistics computed on the permuted data (3) the p-value for the test

References

McKinney E (2022). *Extensions to the Syrjala Test with Eye-Tracking Data Analysis Applications in R*. Ph.D. dissertation, Department of Mathematics and Statistics, Utah State University. (Forthcoming).

McKinney E, Symanzik J (2021). "Extensions to the Syrjala Test with Eye-Tracking Analysis Applications." In *2021 JSM Proceedings.*, 853–889. American Statistical Association, Alexandria, VA.

Examples

```
# Randomly assign all three species to two samples
seedNum <- 123
set.seed(seedNum)

data(iris)
# Randomly assign all three species to two samples
irisPermuted <- iris[sample.int(nrow(iris)), ]
sample1 <- as.matrix(irisPermuted[1:75, 1:2])
sample2 <- as.matrix(irisPermuted[76:150, 1:2])
pooled_data <- rbind(cbind(sample1, 1), cbind(sample2, 2))

# Rotational test
output <- distdiffr(sample1, # Note: Data inputs must be matrices
                    sample2,
                    testType = "rotational",
```

```
                    numRot = 8, # Default value
                    seedNum = seedNum)
output$pval

# Toroidal shift test with proportions of points
output <- distdiffr(sample1,
                    sample2,
                    testType = "toroidal",
                    propPnts = 0.1,
                    seedNum = seedNum)
output$pval

# Toroidal shift test with a threshold below pooled sample size
output <- distdiffr(sample1,
                    sample2,
                    testType = "toroidal",
                    shiftThrshld = 25, # Default
                    seedNum = seedNum)
output$pval

# Toroidal shift test with a threshold above pooled sample size
output <- distdiffr(sample1,
                    sample2,
                    testType = "toroidal",
                    shiftThrshld = 200,
                    seedNum = seedNum)
output$pval

# Toroidal shift test with a number of shifts
output <- distdiffr(sample1,
                    sample2,
                    testType = "toroidal",
                    numShifts = 8,
                    seedNum = seedNum)
output$pval

# Combined rotational and toroidal shift test
output <- distdiffr(sample1,
                    sample2,
                    testType = "combined", # Default
                    numRot = 8,            # Default
                    shiftThrshld = 25,     # Default
                    seedNum = seedNum)
output$pval

# Also see browseVignettes(package = "distdiffR")
```

`grouped_distdiffr` *The combined rotational and toroidal shift distdiffR test for aggregated group data*

Description

The `grouped_distdiffr()` function conducts two-sample permutation tests of distributional equality based on differences in the bivariate empirical cumulative density functions (BECDFs). The

differences in BECDFs are computed across a series of rotations, toroidal shifts, or both rotations and toroidal shifts of the combined data (specified via `testType`). The number of rotations and toroidal shifts may be specified (via `numRot` or `numShifts`, respectively). The number of toroidal shifts may also be determined by a proportion of the combined sample size (via `propPnts`).

Usage

```
grouped_distdiffr(
  aggdata1,
  aggdata2,
  numRot = 8,
  propPnts = NULL,
  numShifts = NULL,
  shiftThrshld = 25,
  numPerms = 999,
  psiFun = CalcGroupPsiCWS,
  seedNum = NULL
)
```

Arguments

`aggdata1`	A three column matrix of bivariate observations from one sample with the third column being the numeric subject labels
`aggdata2`	A three column matrix of bivariate observations from another sample with the third column being the numeric subject labels
`numRot`	An integer number of rotational shifts of the pooled samples
`propPnts`	A numeric proportion of points to be used as toroidal shift origins. Cannot provide both propPnts and numShifts. If neither are provided, shiftThrshld is used.
`numShifts`	A numeric integer. The number of points to be used as toroidal shift origins. Must be less than the pooled sample size. Cannot provide both propPnts and numShifts. If neither are provided, shiftThrshld is used.
`shiftThrshld`	A numeric integer. Used if neither propPnts or numShifts are provided. If the pooled sample size is less than shiftThrshld, every point will be used as a toroidal shift origin. Otherwise, only a random sample of shiftThrshld points will be used.
`numPerms`	An integer number of permutations of the original data
`psiFun`	A function specifying the Psi statistic calculation. Default is the CalcGroupPsiCWS.
`seedNum`	An integer random seed value

Details

Additionally, `grouped_distdiffr()` assumes multiple sources (subjects) are contributing to each sample. As such, the function weights each sources contribution as to treat each equally within the samples, respectively. However, this test is only currently available with the grouped CWS statistic and employs both rotational and toroidal shifts. For more information, see McKinney (2022) and McKinney and Symanzik (2021).

Value

A list including three objects: (1) the Psi statistic computed on the original data (2) a vector of Psi statistics computed on the permuted data (3) the p-value for the test

References

McKinney E (2022). *Extensions to the Syrjala Test with Eye-Tracking Data Analysis Applications in R*. Ph.D. dissertation, Department of Mathematics and Statistics, Utah State University. (Forthcoming).

McKinney E, Symanzik J (2021). "Extensions to the Syrjala Test with Eye-Tracking Analysis Applications." In *2021 JSM Proceedings.*, 853–889. American Statistical Association, Alexandria, VA.

Examples

```
# Randomly assign all three species to two samples
# The species serve as subject labels within each sample
seedNum <- 123
set.seed(seedNum)

data(iris)
irisPermuted <- iris
irisPermuted$Species <- rep(1:3, each = 50)
irisPermuted <- irisPermuted[sample.int(nrow(irisPermuted)), ]
sample1 <- as.matrix(irisPermuted[1:75, c(1:2, 5)])
sample2 <- as.matrix(irisPermuted[76:150, c(1:2, 5)])

output <- grouped_distdiffr(sample1,
                            sample2,
                            seedNum = seedNum)
output$pval
```

`hashMat` *Assigns a hash value to a two-column matrix.*

Description

The `hashMat()` function assigns hash numbers to the two-column sample matrices for the purpose of providing identical test results regardless of the order in which the input data is passed to the `distdiffr()` or `grouped_distdiffr()` functions.

Usage

```
hashMat(mat)
```

Arguments

`mat`	A two column matrix of bivariate observations.

Value

A numeric hash value

Examples

```
data(iris)
sample1 <- as.matrix(iris[iris$Species == "virginica", 1:2])
sample2 <- as.matrix(iris[iris$Species == "versicolor", 1:2])

hashMat(sample1)
hashMat(sample2)
```

NumToroShiftData *Apply a toroidal shift to the pooled samples using a number of points*

Description

The NumToroShiftData() function produces a list of toroidal shifted versions of the two-column input matrix. The number of toroidal shifts is an integer passed to numShifts. The origins of the toroidal shifts are randomly selected from the combined samples. The pooled data is assumed to list all of the first sample of size n1 before the second sample (of size n2).

Usage

```
NumToroShiftData(data, n1, n2, numShifts)
```

Arguments

data	A two column matrix of the pooled samples
n1	An integer sample size for the first sample
n2	An integer sample size for the second sample
numShifts	A numeric number of points to be used as toroidal shift origins

Value

A list of toroidal shifted pooled sample matrices

Examples

```
data(iris)
sample1 <- as.matrix(iris[iris$Species == "setosa", 1:2])
sample2 <- as.matrix(iris[iris$Species == "virginica", 1:2])
pooled_data <- rbind(sample1, sample2)
n1 <- nrow(sample1)
n2 <- nrow(sample2)

# Create a list of five toroidal shifts of the pooled data
output <- NumToroShiftData(pooled_data, n1, n2, 25)
summary(output)
```

PropToroShiftData *Apply a toroidal shift to the pooled samples using a proportion of points*

Description

The PropToroShiftData() function produces a list of toroidal shifted versions of the two-column input matrix. The number of toroidal shifts is (the ceiling of) the proportion (propPnts) multiplied by the combined sample size. The origins of the toroidal shifts are randomly selected from the combined samples. The pooled data is assumed to list all of the first sample of size n1 before the second sample (of size n2).

Usage

```
PropToroShiftData(data, n1, n2, propPnts = 1)
```

Arguments

data	A two column matrix of the pooled samples
n1	An integer sample size for the first sample
n2	An integer sample size for the second sample
propPnts	A numeric proportion of points to be used as toroidal shift origins

Value

A list of toroidal shifted pooled sample matrices

Examples

```
data(iris)
sample1 <- as.matrix(iris[iris$Species == "setosa", 1:2])
sample2 <- as.matrix(iris[iris$Species == "virginica", 1:2])
pooled_data <- rbind(sample1, sample2)
n1 <- nrow(sample1)
n2 <- nrow(sample2)

# Creates a list of 0.1 times (n1 + n2) = 10 toroidal shifts of the pooled data
output <- PropToroShiftData(pooled_data, n1, n2, 0.1)
summary(output)
```

RotateData *Create rotated versions of the data*

Description

This function produces a list of rotated versions of the two-column input matrix. Specifically, the number of rotations (i.e., an integer passed to numRotations) divides a complete circle into numRotations equal angles, and numRotations rotated versions of the input data are output in a list.

Usage

```
RotateData(data, numRotations)
```

Arguments

data	A two column matrix of the bivariate combined samples
numRotations	A non-negative integer specifying the number of rotations to be applied to the data within 360 degrees.

Value

A list of matrices containing the coordinates for each version of the rotated data (including the original data, which is the first matrix of the list)

Examples

```
data(iris)
sample1 <- as.matrix(iris[iris$Species == "setosa", 1:2])

# Generate five rotated versions of sample1 (every 72 degrees) within 360 degrees.
RotateData(sample1, 5)
```

Index

CHAPTER 9

Discussion and Future Work

This chapter provides a conclusion to this dissertation by discussing the main insights of the research (in Section 9.1) in addition to providing directions for future work (in Section 9.2).

9.1 Concluding Discussion

This dissertation introduced a series of new two-sample tests of distributional equality. The new tests are a generalization of the Syrjala (1996) test and make use of both rotations and toroidal shifts of the data. The new tests also remove the requirement for identical sampling locations between the two samples as assumed in the original Syrjala test. While the inclusion of rotations of the data is more of a generalization to the original four rotations of the data in the Syrjala test (see Section 5.2 for additional details), the inclusion of toroidal shifts within the test, and combination of toroidal shifts within rotations, are novel extensions (see Sections 5.3 and 5.4). Furthermore, a version of the test was developed to treat each subject's contribution equally within the respective pooled samples (see Section 7.1.1 for more details).

From the series of simulations that have been discussed throughout Sections 6.2–6.5, the following conclusions can be made:

- The Syrjala (1996) test has been shown to depend upon data aggregation techniques such as regular and random binning. It is recommended to use another bivariate two sample test of distributional equality which does not assume identical sampling locations. Such tests include the Energy test by Székely and

Rizzo (2004), the kernel maximum mean discrepancy by Gretton et al. (2012), the Friedman and Rafsky (1979) generalization to the Kolmogorov (1933) test, or one of the modified Syrjala tests proposed in this dissertation.

- The modified Syrjala tests have been shown to be insensitive to differences in the weightings of the tests statistics, and only a marginal gain in power was found when using squared differences in the ECDFs as compared to absolute differences. Additionally, all of the tests were shown to produce relatively stable results regardless of the number of rotations or toroidal shifts explored within the simulations.

- While the modified Syrjala tests which employ toroidal shifts achieve roughly the same power as the tests which employ both rotational and toroidal shifts, the latter tests achieve an average false positive rate (0.03 vs. 0.045, respectively) closer to the significance level (0.05). Thus, the combined modifications produce conservative tests which are more powerful in the face of all departures from the null than the tests which employ toroidal shifts alone. However, this balance comes at the cost of increased computational complexity.

- The modified Syrjala tests which use eight rotations, 0.1 proportion of points as origins for toroidal shifts, and the CWS statistic has been shown by simulation to achieve a higher number of significant tests (when the null is false) than several other competing methods including the Energy test by Székely and Rizzo (2004), the kernel maximum mean discrepancy by Gretton et al. (2012), the Friedman and Rafsky (1979) generalization to the Kolmogorov (1933) test, and the original Syrjala (1996) test (when preliminary data binning is employed).

- The modified Syrjala test which uses a threshold of 25 randomly chosen points from the pooled sample as origins for toroidal shifts (see Section 6.3.4) achieves

comparable results as the tests which employ proportions of points as origins for toroidal shifts (see Section 6.3.3). This provides motivation to a default threshold value for the tests in the `R` package (see Chapter 8). These default values guide new users of the package toward parameter values for tests which compute reasonably fast (e.g., on the order of seconds for sample sizes less than 100, and on the order of minutes for sample sizes less than 500).

- The modified Syrjala tests have been shown to be well suited to certain types of eye-tracking data by simulation. Specifically, the versions of the test which use the CWS statistic, eight rotations, and either 0.1 proportion of toroidal shifts or a threshold of 25 toroidal shifts have been shown to be robust against a small number of outliers (see Sections 6.5.6, 6.5.7, 6.5.8, and 6.5.9). Additionally, these versions of the test can detect a variety of differences between samples, including differences in gaze point cluster centers (Sections 6.5.3 and 6.5.9), cluster shapes (Sections 6.5.4 and 6.5.9), proportions allocated to different objects being viewed by subjects (Sections 6.5.5 and 6.5.10), and proportions of noise (Sections 6.5.7, 6.5.8 and 6.5.9)

The test which employs the CWS statistic, eight rotations, and 0.1 proportions of toroidal shifts has also been applied to a new study in eye-tracking and postural stability assessment, called the Utah State University (USU) Posture Study (Symanzik et al., 2017, 2018; Studenka et al., 2020; Coltrin et al., 2020; McKinney and Symanzik, 2019, 2021). The setup, data collection, and data preprocessing of the USU Posture Study has been provided (Chapter 4). While additional analyses should be conducted, the study results in this dissertation suggest that there is a detectable difference between the treatment and control groups captured in the subject's eye-tracking data.

The new tests, called the modified Syrjala tests, have been made available via the `distdiffR` package for the `R` software environment for statistical computing and graphics. The `distdiffR` package can be downloaded from `https://github.com/EricMcKinney77/distdiffR`.

9.2 Future Work

In additional to the methods and results provided within this dissertation, additional areas of research have been discovered. These areas of future work are listed below.

- Application of the modified Syrjala test to previous studies where the Syrjala test has been applied for comparative purposes.
- Further refine the `distdiffR` algorithms and associated code for computational efficiency.
- Extend the methodology of the modified Syrjala tests to apply to higher dimensions of data beyond the bivariate case.
- Apply the methodology of the modified Syrjala tests in the setting of unsupervised learning.
- Analyze and fuse the other collected data with the eye-tracking data within the USU Posture Study.

APPENDIX A

Mathematical Proofs

This appendix provides the mathematical proofs of theorems and other mathematical properties stated in the main chapters of this dissertation. Theorem A.1.1, Corrolary A.1.1, and Theorem A.1.2 are well known and provided here to give further detail to the discussed material in Section 2.1. However, Theorem A.2.1 along with its associated proof is novel, and is provided to show equivalency between two approaches for treating each subject's contributions equally within the group-wise tests in the analysis of the Utah State University Posture Study data (provided in Section 7.1.1).

A.1 Mathematical Proofs

Theorem A.1.1. *The empirical cumulative distribution function (ECDF) is an unbiased estimator of cumulative distribution function (CDF).*

Proof. For n independent and identically distributed random variables, $X_1, X_2, \ldots, X_n$, let $S(x)$ and $F(x)$ be the ECDF and CDF, respectively. By definition, the ECDF can be written as $S(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_i \leq x}$, where $i = 1, \ldots, n$. The function $\mathbf{1}_{X_i \leq x}$ is one if $X_i \leq x$, and zero otherwise. $\mathbf{1}_{X_i \leq x}$ is also commonly known as the indicator function (Rice, 2006).

Notice for a fixed value x, the indicator function $\mathbf{1}_{X_i \leq x} = 1$ with probability $p = F(x)$. Hence, the indicator function is a Bernoulli random variable (Rice, 2006)

with a parameter p. If we multiply $S(x)$ by n, then

$$\begin{aligned} nS(x) &= n\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{X_i\leq x} \\ &= \sum_{i=1}^{n}\mathbf{1}_{X_i\leq x}. \end{aligned}$$

This implies that $nS(x)$ is a binomial random variable (Rice, 2006) with a mean of $nF(x)$ and variance of $nF(x)(1-F(x))$. Now observe that the expected value of the ECDF can be written as

$$E[S(x)] = \frac{n}{n}E[S(x)] = \frac{1}{n}E[nS(x)] = \frac{1}{n}\left(nF(x)\right) = F(x).$$

This implies that $S(x)$ is an unbiased estimator of $F(x)$. ■

Corollary A.1.1. *As a corollary to Theorem A.1.1, the variance of $S(x)$ can also be derived as* $\dfrac{F(x)(1-F(x))}{n}$.

Proof. For n independent and identically distributed random variables, $X_1, X_2, \ldots, X_n$, let $S(x)$ and $F(x)$ be the ECDF and CDF, respectively. Since $nS(x)$ is a binomial random variable (Rice, 2006) with a mean of $nF(x)$ and variance of $nF(x)(1-F(x))$, then

$$var(S(x)) = \frac{n^2}{n^2}var(S(x)) = \frac{1}{n^2}var(nS(x)) = \frac{1}{n^2}(nF(x)(1-F(x))) = \frac{F(x)(1-F(x))}{n}.$$

■

Theorem A.1.2. *The ECDF is a consistent estimator of the CDF.*

Proof. Let $X_1, X_2, \ldots, X_n$ be n independent and identically distributed random variables with $S(x)$ and $F(x)$ as the ECDF and CDF, respectively.

By definition, we need to show that for any arbitrarily small $\epsilon > 0$,

$$P[|S(x) - F(x)| \geq \epsilon] \to 0, \text{ as } n \to \infty.$$

By Chebyshev's inequality (Rice, 2006) we have

$$P[|S(x) - F(x)| \geq \epsilon] \leq \frac{var(S(x))}{\epsilon^2}.$$

Since $var(S(x)) = \frac{F(x)(1-F(x))}{n}$ (see Corrolary A.1.1), then

$$\begin{aligned} \frac{var(S(x))}{\epsilon^2} &= \frac{\frac{F(x)(1-F(x))}{n}}{\epsilon^2} \\ &= \frac{F(x)(1-F(x))}{n\epsilon^2}. \end{aligned}$$

Hence, $P[|S(x) - F(x)| \geq \epsilon] \leq \frac{F(x)(1-F(x))}{n\epsilon^2} \to 0$ as $n \to \infty$ for any $\epsilon > 0$. Therefore, $S(x)$ is a consistent estimator of $F(x)$.

■

A.2 Methods and Proof for Treating Each Subject's Contributions Equally

When computing the differences in ECDFs between the treatment and control groups, a difficulty arises in that each subject contributed a differing amount of gaze points depending on how long they observed the posture of interest. However, their contributions to a group-wise test need to be treated equally. Hence, two methods are proposed in Section 7.1.1 for treating each subject's contributions equally when conducting a between-group test. The methods are detailed here along with a proof of their equivalence.

Method 1: Consider all K subject's number of gaze point contributions when aggregating all group-wise gaze points. Since each subject's number of gaze points, $n_1, n_2, \ldots, n_K$, is an integer, then there exists a lowest common multiple (LCM) between them, say L. Hence, there exists an $a_i \in \mathbb{I}$ for each n_i such that $n_i a_i = L$. If we duplicate each subject's gaze points by a factor of a_i, then each subject will have contributed an equal number of L gaze points. From here, all of the K times L gaze points can be aggregated into a group sample, and an ECDF constructed, say Γ_c.

Method 2: Let Γ_i be the ECDFs for each of the subject's data within one group. An average ECDF value, Γ_{AVE} can be computed as $\Gamma_{AVE}(x, y) = \frac{1}{K} \sum_{i=1}^{K} \Gamma_i(x, y)$.

Theorem A.2.1. *Γ_c (as defined above in Method 1) is equal to Γ_{AVE} (as defined above in Method 2).*

Proof. Let $\tilde{X}_i = [(x_{i,1}, y_{i,1}), (x_{i,2}, y_{i,2}) \ldots, (x_{i,n_i}, y_{i,n_i})]$, $i = 1, 2, \ldots, K$, be bivariate sample vectors from our K subjects, each with sample size $n_1, n_2, \ldots, n_K$, respectively. Now, let L be the LCM of $n_1, n_2, \ldots, n_K$. Then each $n_1, n_2, \ldots, n_K$ has an $a_1, a_2, \ldots, a_K \in \mathbb{I}$, such that $n_i a_i = L$. (Hence, a_i are the product of missing factors which must be multiplied to n_i in order to equal L.)

Now duplicate each bivariate element of $\tilde{X}_i = [(x_{i,1}, y_{i,1}), (x_{i,2}, y_{i,2}) \ldots, (x_{i,n_i}, y_{i,n_i})] = [(x_{i,j_i}, y_{i,j_i})]_{j_i=1}^{n_i}$ by a_i. Call these new samples, $\tilde{X}_i^*$ such that

$$\begin{aligned}\tilde{X}_i^* = [&(x_{i,1,1}, y_{i,1,1}), \ldots, (x_{i,1,a_i}, y_{i,1,a_i}),\\ &(x_{i,2,1}, y_{i,2,1}), \ldots, (x_{i,2,a_i}, y_{i,2,a_i}),\\ &(x_{i,n_i,1}, y_{i,n_i,1}) \ldots, (x_{i,n_i,a_i}, y_{i,n_i,a_i})].\end{aligned}$$

Also, let $k_i = 1, \ldots, a_i$. (Note, all of the $\tilde{X}_i^*$ will now have an equal number of $n_i a_i = L$ elements.) Combine all of the $\tilde{X}_i^*$ samples into one large sample $\tilde{X}^*$. ($\tilde{X}^*$ will have $\sum_{i=1}^{K} a_i n_i = \sum_{i=1}^{K} L = KL$ elements.)

From here we can construct an ECDF for $\tilde{X}^*$, called Γ_c^*, where

$$\Gamma_c^*(x,y) = \frac{1}{KL}\sum_{i=1}^{K}\sum_{j_i=1}^{n_i}\sum_{k_i=1}^{a_i}\mathbf{I}_{x_{i,j_i,k_i}\leq x, y_{i,j_i,k_i}\leq y}.$$

Alternatively, we can construct an average ECDF value across all of the individual ECDFs, Γ_i, for the original (non-duplicated) subject data, say Γ_{AVE}, such that

$$\Gamma_{AVE}(x,y) = \frac{1}{K}\sum_{i=1}^{K}\Gamma_i(x,y).$$

Our goal is to show that $\Gamma_{AVE}(x,y) = \Gamma_c^*(x,y)$.

Notice that,

$$\Gamma_{AVE}(x,y) = \frac{1}{K}\sum_{i=1}^{K}\Gamma_i(x,y)$$

$$= \frac{1}{K}\left[\Gamma_1(x,y) + \Gamma_2(x,y) + \ldots + \Gamma_K(x,y)\right]$$

$$= \frac{1}{K}\left[\frac{1}{n_1}\sum_{j_1=1}^{n_1}\mathbf{I}_{x_{1,j_1}\leq x, y_{1,j_1}\leq y} + \frac{1}{n_2}\sum_{j_2=1}^{n_2}\mathbf{I}_{x_{1,j_2}\leq x, y_{1,j_2}\leq y} + \ldots + \frac{1}{n_K}\sum_{j_K=1}^{n_K}\mathbf{I}_{x_{1,j_K}\leq x, y_{1,j_K}\leq y}\right]$$

$$= \frac{1}{K}\left[\frac{a_1}{a_1 n_1}\sum_{j_1=1}^{n_1}\mathbf{I}_{x_{1,j_1}\leq x, y_{1,j_1}\leq y} + \right.$$
$$\frac{a_2}{a_2 n_2}\sum_{j_2=1}^{n_2}\mathbf{I}_{x_{1,j_2}\leq x, y_{1,j_2}\leq y} + \ldots +$$
$$\left.\frac{a_K}{a_K n_K}\sum_{j_K=1}^{n_K}\mathbf{I}_{x_{1,j_K}\leq x, y_{1,j_K}\leq y}\right]$$

$$= \frac{1}{K}\left[\frac{1}{a_1 n_1}\sum_{j_1=1}^{n_1}a_1\mathbf{I}_{x_{1,j_1}\leq x, y_{1,j_1}\leq y} + \right.$$
$$\frac{1}{a_2 n_2}\sum_{j_2=1}^{n_2}a_2\mathbf{I}_{x_{1,j_2}\leq x, y_{1,j_2}\leq y} + \ldots +$$
$$\left.\frac{1}{a_K n_K}\sum_{j_K=1}^{n_K}a_K\mathbf{I}_{x_{1,j_K}\leq x, y_{1,j_K}\leq y}\right]$$

$$= \frac{1}{K}\Bigg[\frac{1}{a_1 n_1}\sum_{j_1=1}^{n_1}\sum_{k_1=1}^{a_1}\mathbf{I}_{x_{1,j_1,k_1}\leq x, y_{1,j_1,k_1}\leq y} + \frac{1}{a_2 n_2}\sum_{j_2=1}^{n_2}\sum_{k_2=1}^{a_2}\mathbf{I}_{x_{2,j_2,k_2}\leq x, y_{2,j_2,k_2}\leq y} + \ldots + \frac{1}{a_K n_K}\sum_{j_K=1}^{n_k}\sum_{k_K=1}^{a_K}\mathbf{I}_{x_{K,j_K,k_K}\leq x, y_{K,j_K,k_K}\leq y}\Bigg] \quad (*)$$

$$= \frac{1}{K}\Bigg[\frac{1}{L}\sum_{j_1=1}^{n_1}\sum_{k_1=1}^{a_1}\mathbf{I}_{x_{1,j_1,k_1}\leq x, y_{1,j_1,k_1}\leq y} + \frac{1}{L}\sum_{j_2=1}^{n_2}\sum_{k_2=1}^{a_2}\mathbf{I}_{x_{2,j_2,k_2}\leq x, y_{2,j_2,k_2}\leq y} + \ldots + \frac{1}{L}\sum_{j_K=1}^{n_k}\sum_{k_K=1}^{a_K}\mathbf{I}_{x_{K,j_K,k_K}\leq x, y_{K,j_K,k_K}\leq y}\Bigg]$$

$$= \frac{1}{KL}\Bigg[\sum_{j_1=1}^{n_1}\sum_{k_1=1}^{a_1}\mathbf{I}_{x_{1,j_1,k_1}\leq x, y_{1,j_1,k_1}\leq y} + \sum_{j_2=1}^{n_2}\sum_{k_2=1}^{a_2}\mathbf{I}_{x_{2,j_2,k_2}\leq x, y_{2,j_2,k_2}\leq y} + \ldots + \sum_{j_K=1}^{n_k}\sum_{k_K=1}^{a_K}\mathbf{I}_{x_{K,j_K,k_K}\leq x, y_{K,j_K,k_K}\leq y}\Bigg]$$

$$= \frac{1}{KL}\sum_{i=1}^{K}\sum_{j_i=1}^{n_i}\sum_{k_i=1}^{a_i}\mathbf{I}_{x_{i,j_i,k_i}\leq x, y_{i,j_i,k_i}\leq y}$$

$$= \Gamma_c^*(x,y)$$

Note: the step $(*)$ holds because $\sum_{j_i=1}^{n_i} a_i \mathbf{I}_{x_{1,j_i}\leq x, y_{1,j_i}\leq y} = \sum_{j_i=1}^{n_i}\sum_{k_i=1}^{a_i}\mathbf{I}_{x_{i,j_i,k_i}\leq x, y_{i,j_i,k_i}\leq y}$ (i.e., multiplying each of the indicator functions by a_i has an equivalent effect as duplicating the subject's data). ■

APPENDIX B

Additional Simulation Results

Since Figures 18 and 62–66 show almost the same rotational modified test behavior aside from some chance variation (in Section 6.3.1), the latter figures (Figures 62–66) for the DWS, UWS, DWA, UWA, and CWA simulations (respectively) are provided in this appendix. Similarly, Figures 67–71 show almost the same toroidal shift modified test behavior aside from some chance variation as Figure 19 in Section 6.3.2. Hence, Figures 67–71 are also provided in this appendix. Also, Figures 20–22 in Section 6.3.4 are patterned closely to Figures 72–77, and Figure 23 in Section 6.3.4 is patterned closely to Figures 78 and 79. Hence, Figures 72–79 are also provided in this appendix. For similar reasons, additional simulation results which do not differ dramatically from those in Sections 6.5.3 and 6.5.9 are provided here as Figures 80–89

Definitions for the test statistic abbreviations in these plots can be found in Section 5.1. Additionally, explanations of the graph features in Figures 62–89 can be found in Section 6.3.

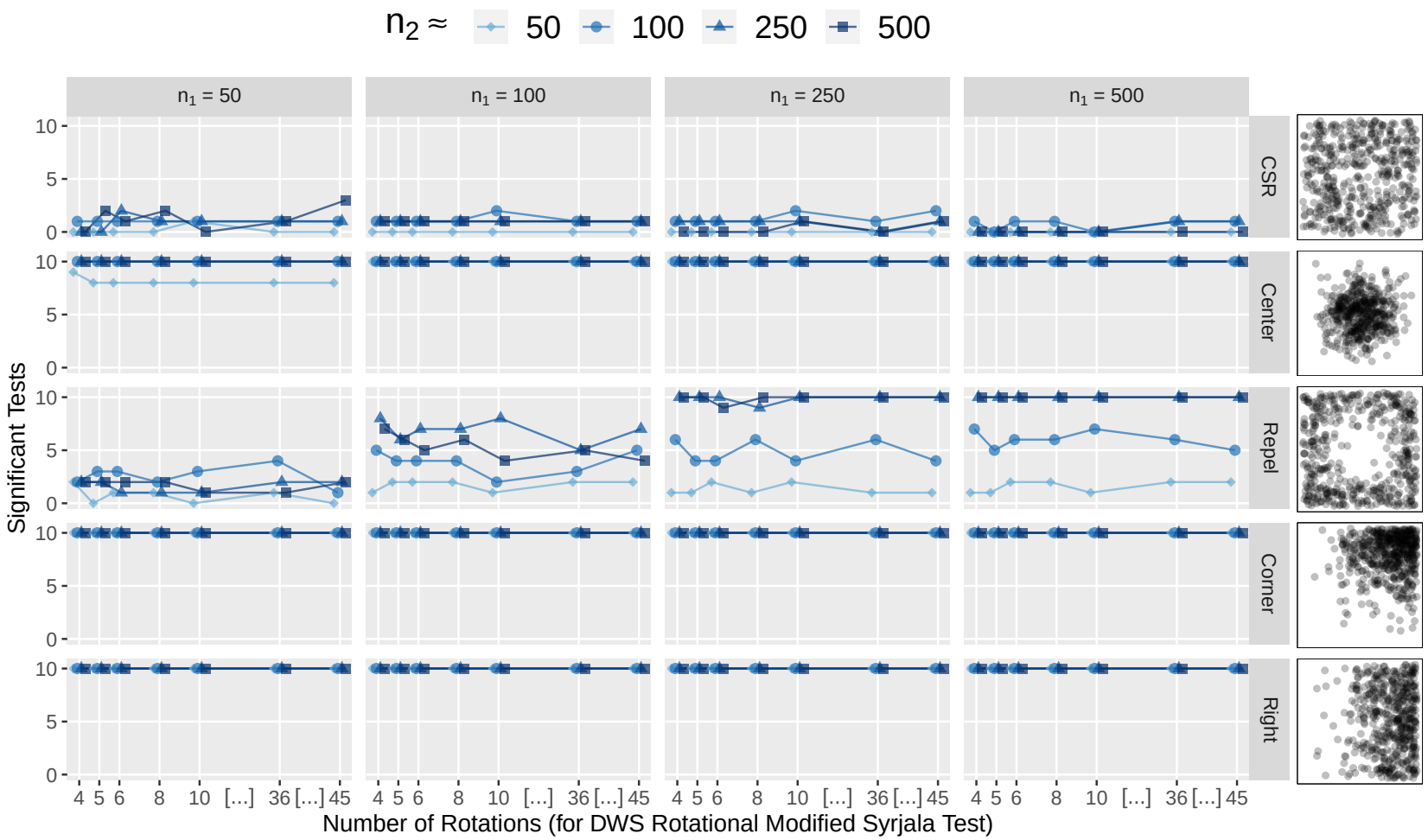


Fig. 62: A grid of line graphs showing the results of a simulation comparing multiple rotations of the modified Syrjala test using double weightings of the squared differences in the ECDFs (DWS). The grid column name indicates the CSR sample size (n_1), and the grid row indicates the shape of the second sample. The colors of the lines and symbols indicate the second sample size (approximate n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) on CSR realizations of 50 points with Right realizations of approximately 50, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

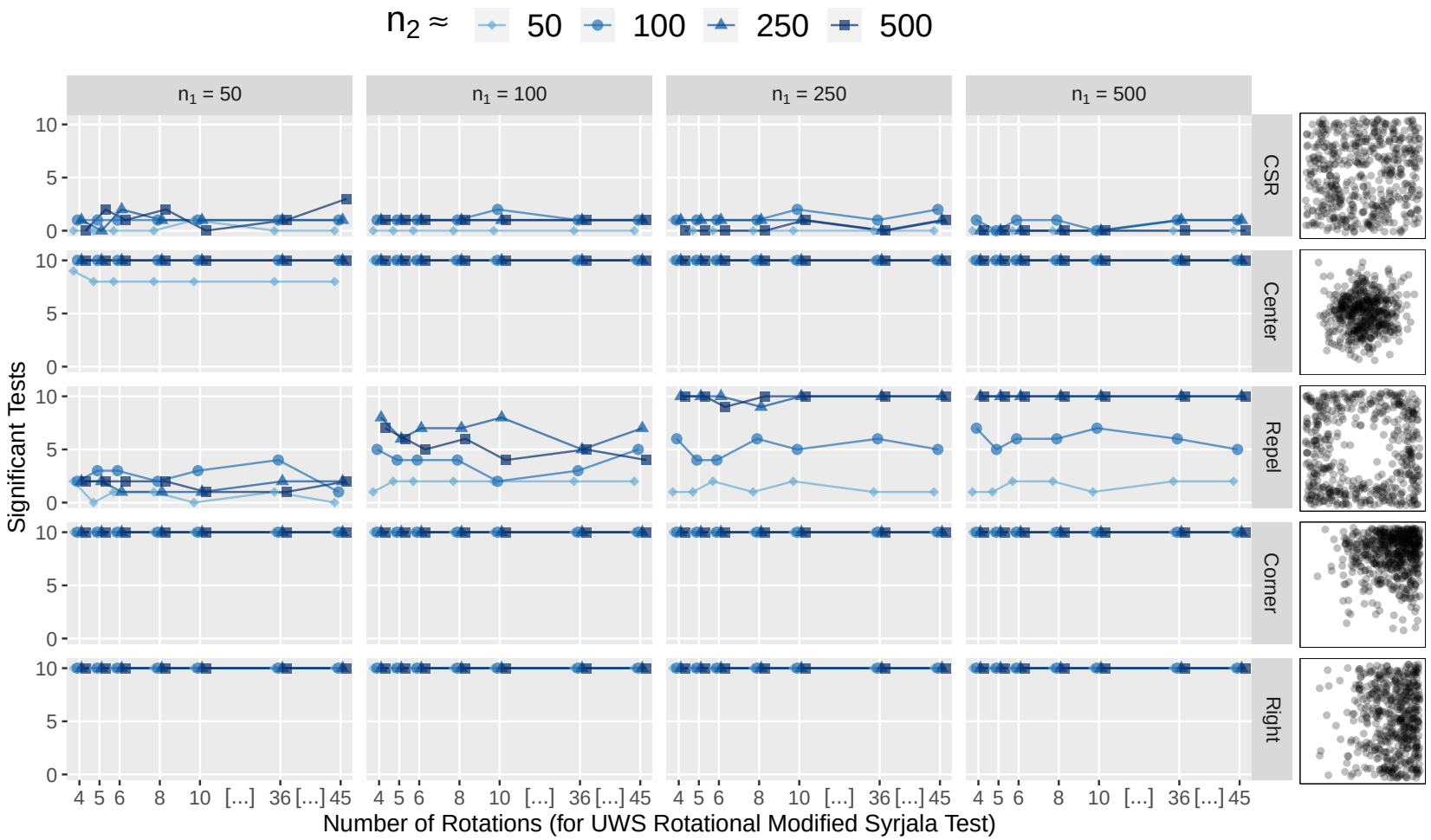


Fig. 63: A grid of line graphs showing the results of a simulation comparing multiple rotations of the modified Syrjala test using unweighted squared differences in the ECDFs (UWS). The grid column name indicates the CSR sample size (n_1), and the grid row indicates the shape of the second sample. The colors of the lines and symbols indicate the second sample size (approximate n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) on CSR realizations of 50 points with Right realizations of approximately 50, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

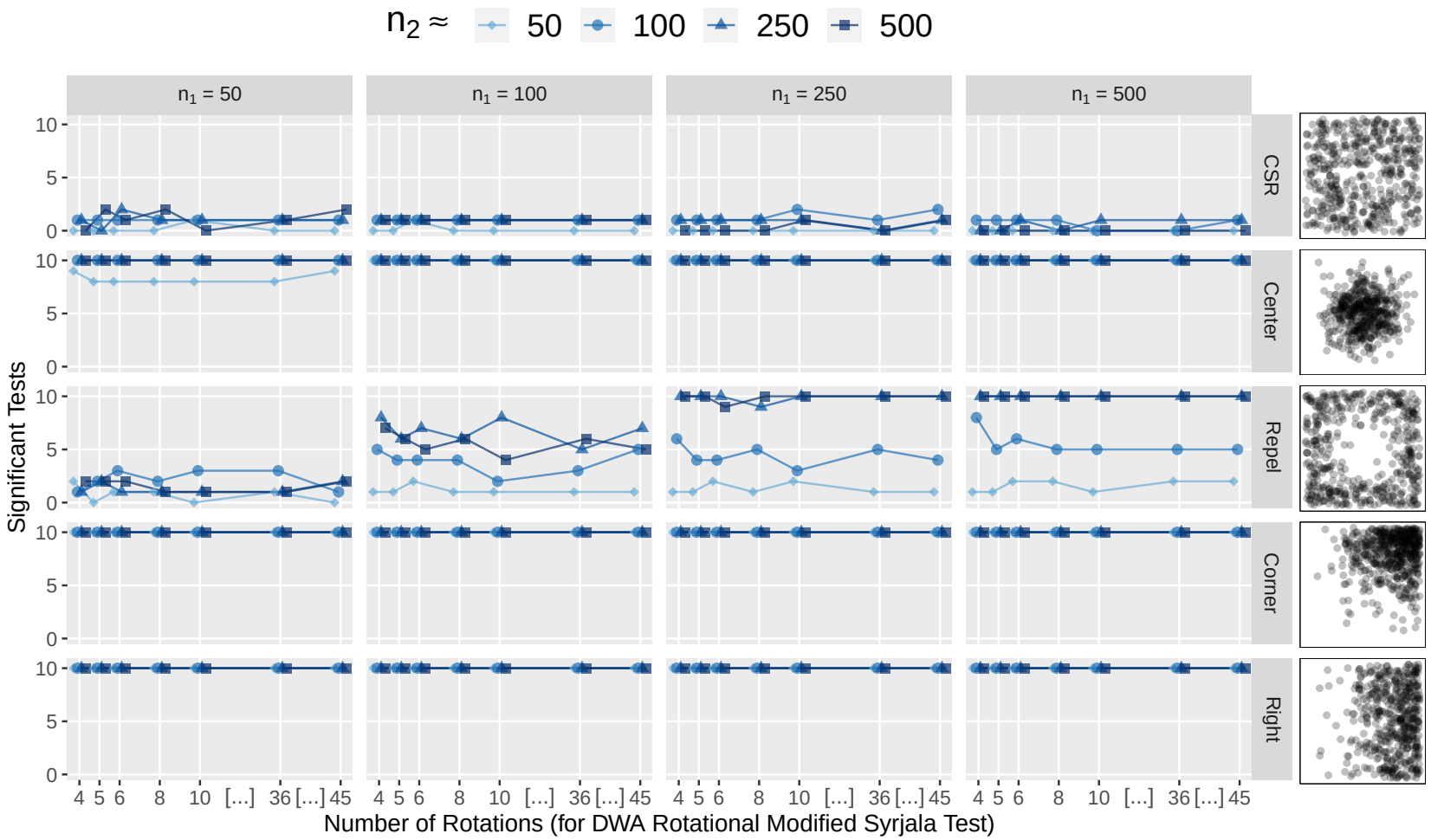


Fig. 64: A grid of line graphs showing the results of a simulation comparing multiple rotations of the modified Syrjala test using double weightings of the absolute differences in the ECDFs (DWA). The grid column name indicates the CSR sample size (n_1), and the grid row indicates the shape of the second sample. The colors of the lines and symbols indicate the second sample size (approximate n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) on CSR realizations of 50 points with Right realizations of approximately 50, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

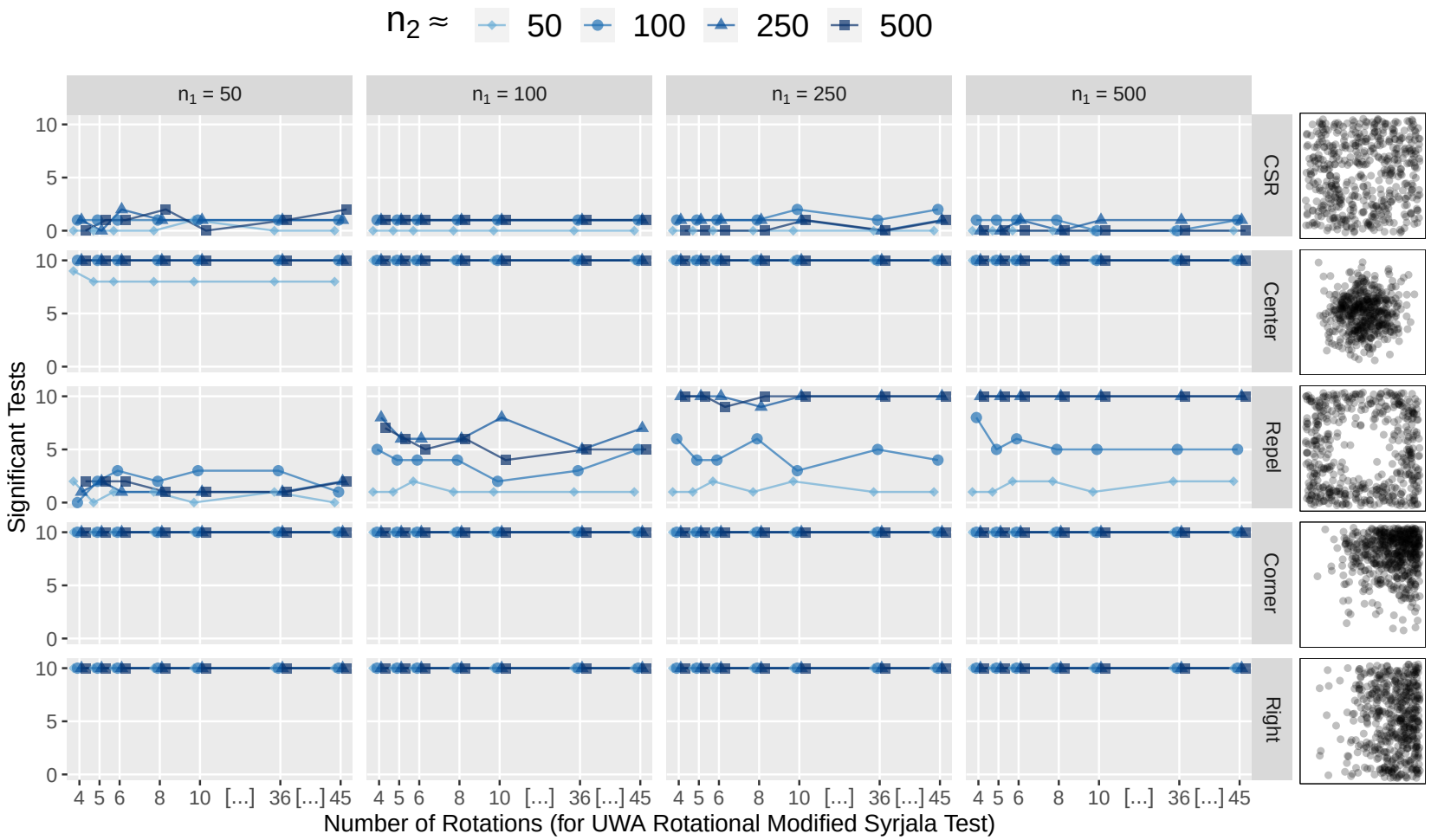


Fig. 65: A grid of line graphs showing the results of a simulation comparing multiple rotations of the modified Syrjala test using unweighted absolute differences in the ECDFs (UWA). The grid column name indicates the CSR sample size (n_1), and the grid row indicates the shape of the second sample. The colors of the lines and symbols indicate the second sample size (approximate n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) on CSR realizations of 50 points with Right realizations of approximately 50, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

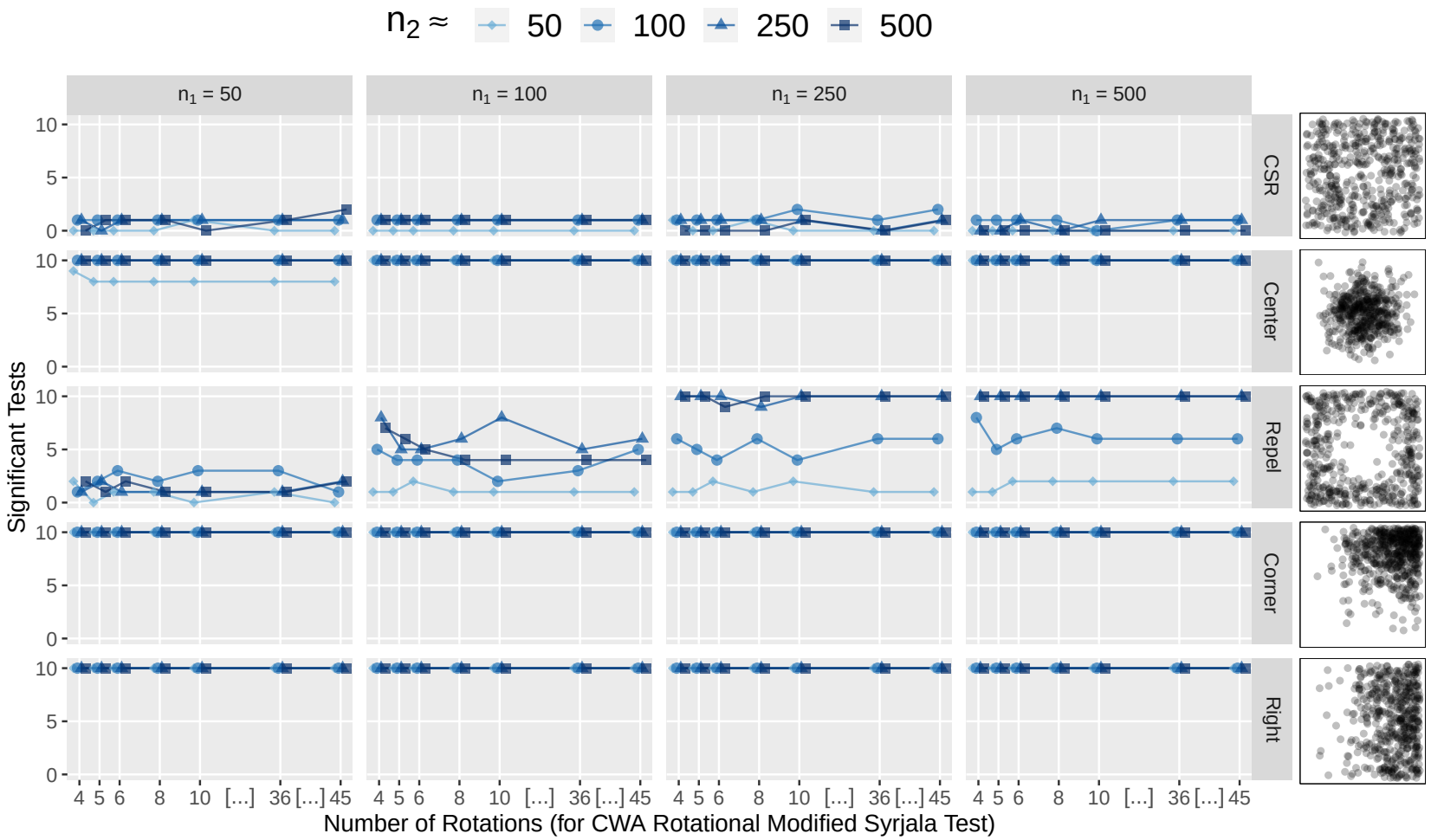


Fig. 66: A grid of line graphs showing the results of a simulation comparing multiple rotations of the modified Syrjala test using complementary weightings of the absolute differences in the ECDFs (CWA). The grid column name indicates the CSR sample size (n_1), and the grid row indicates the shape of the second sample. The colors of the lines and symbols indicate the second sample size (approximate n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) on CSR realizations of 50 points with Right realizations of approximately 50, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

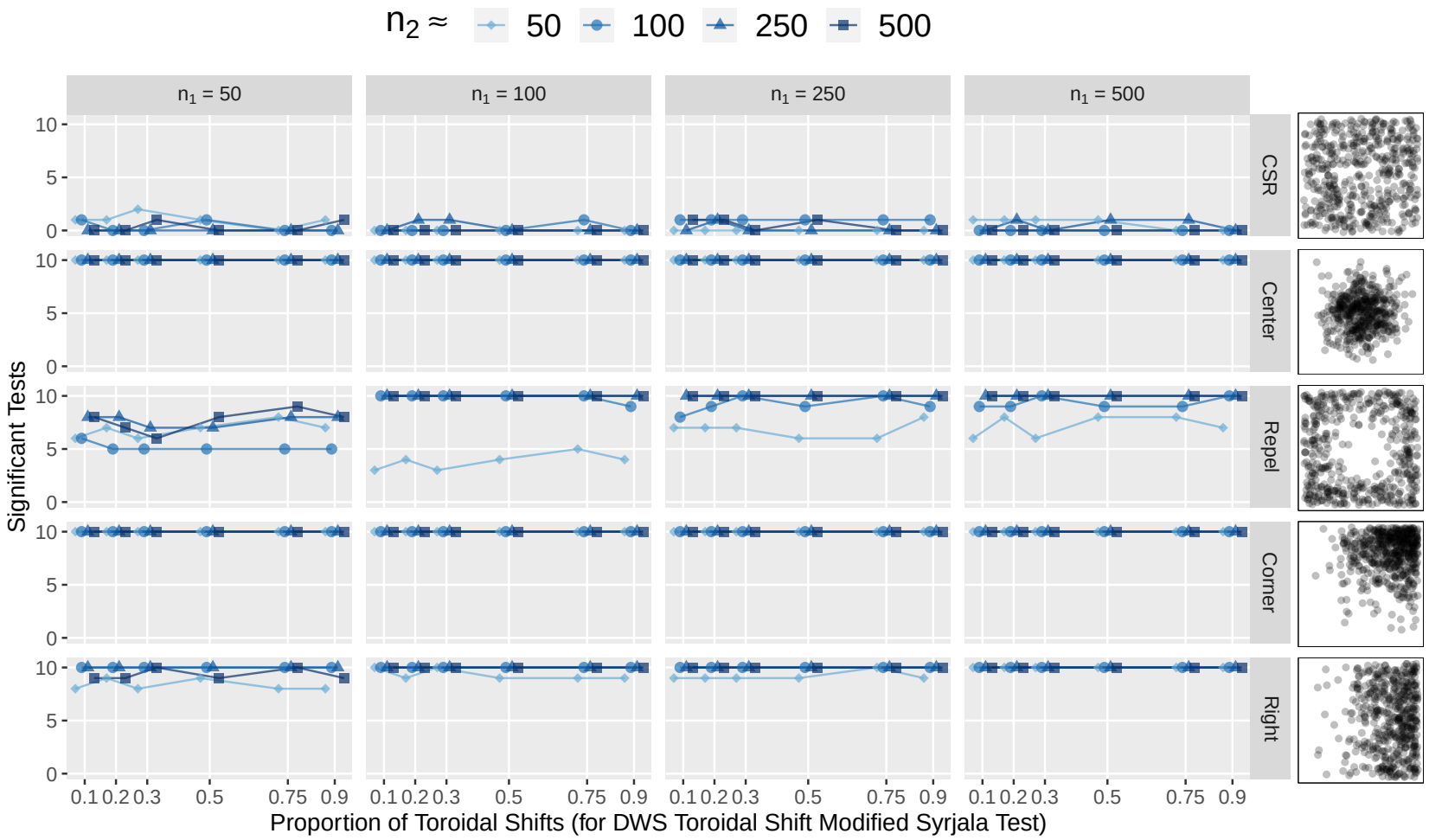


Fig. 67: A grid of line graphs showing the results of a simulation comparing multiple randomly selected proportions of points for the origins of the toroidal shifts of the modified Syrjala test using double weightings of the squared differences in the ECDFs (DWS). The grid column name indicates the CSR sample size (n_1), and the grid row indicates the shape of the second sample. The colors of the lines and symbols indicate the second sample size (approximate n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) on CSR realizations of 50 points with Right realizations of approximately 50, 100, 250, and 500 points.

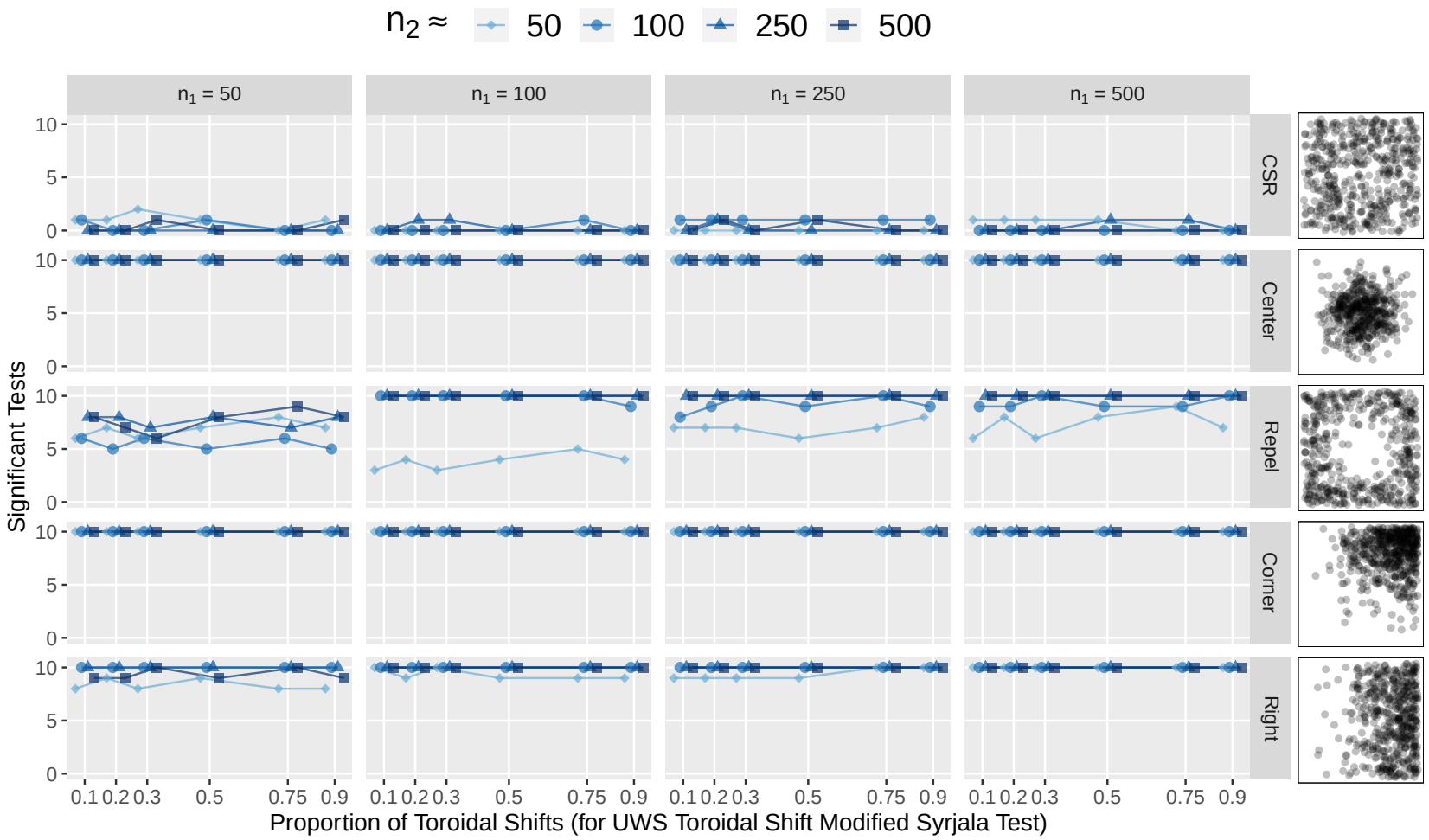


Fig. 68: A grid of line graphs showing the results of a simulation comparing multiple randomly selected proportions of points for the origins of the toroidal shifts of the modified Syrjala test using unweighted squared differences in the ECDFs (UWS). The grid column name indicates the CSR sample size (n_1), and the grid row indicates the shape of the second sample. The colors of the lines and symbols indicate the second sample size (approximate n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) on CSR realizations of 50 points with Right realizations of approximately 50, 100, 250, and 500 points.

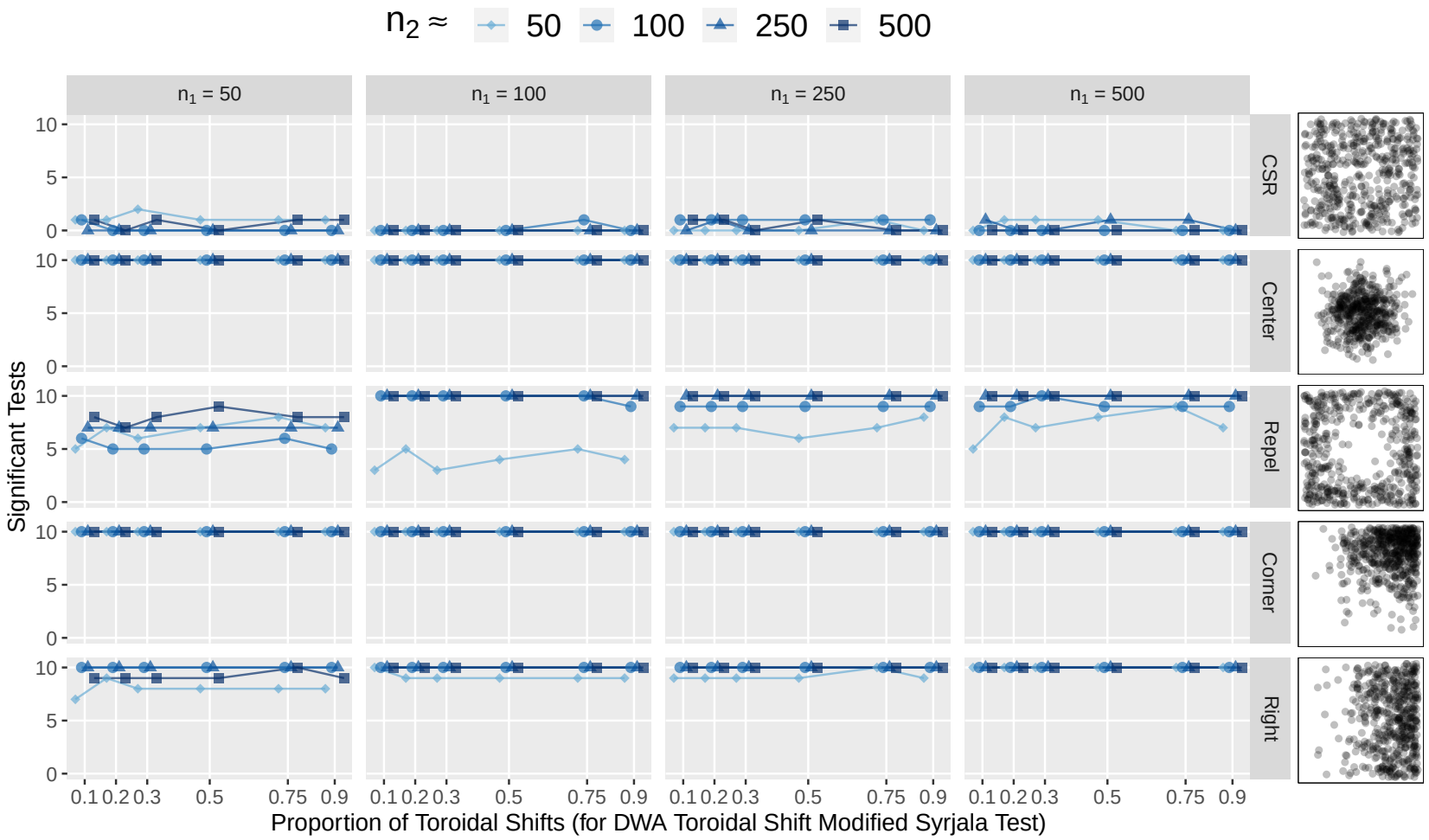


Fig. 69: A grid of line graphs showing the results of a simulation comparing multiple randomly selected proportions of points for the origins of the toroidal shifts of the modified Syrjala test using double weightings of the absolute differences in the ECDFs (DWA). The grid column name indicates the CSR sample size (n_1), and the grid row indicates the shape of the second sample. The colors of the lines and symbols indicate the second sample size (approximate n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) on CSR realizations of 50 points with Right realizations of approximately 50, 100, 250, and 500 points.

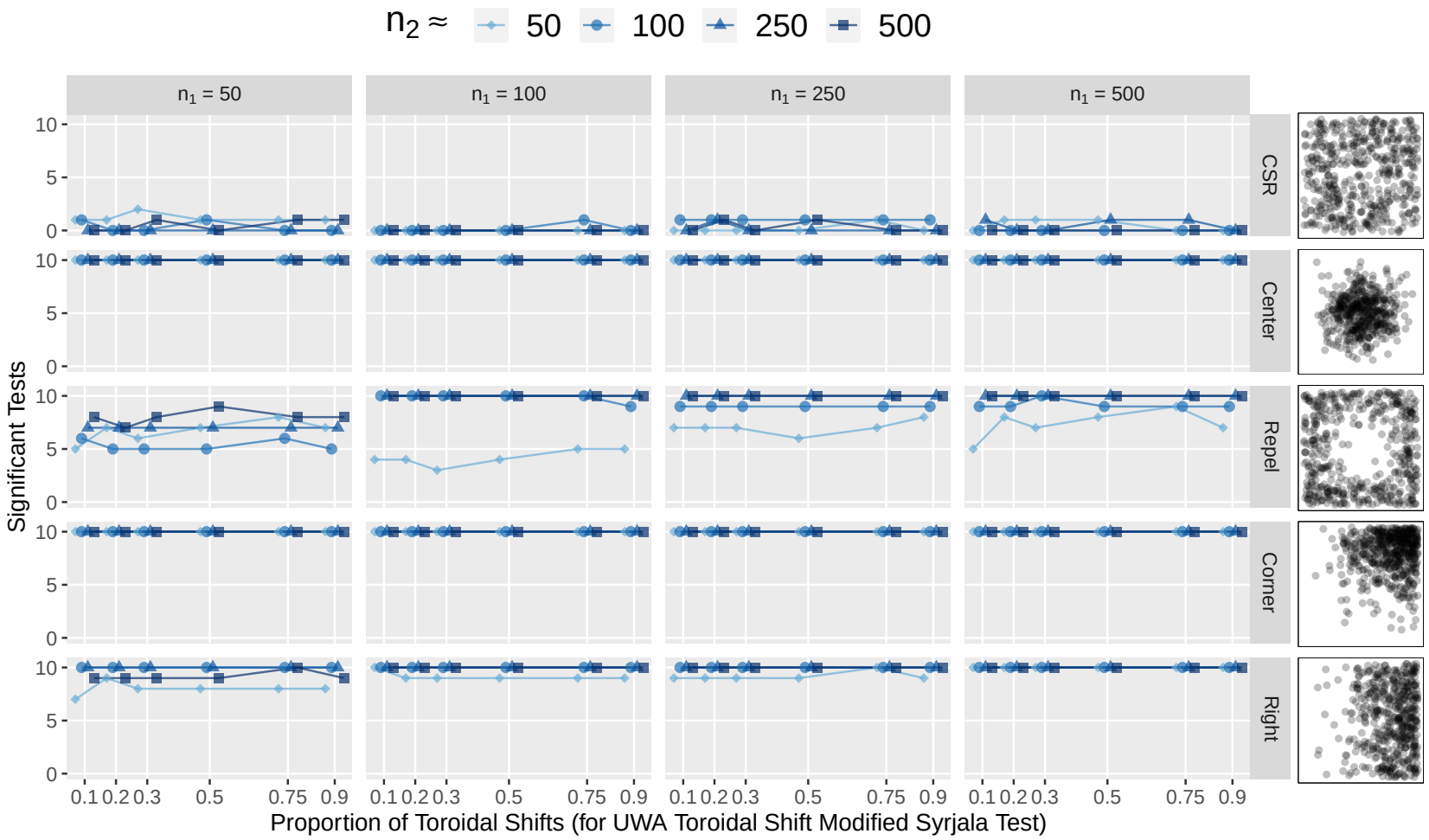


Fig. 70: A grid of line graphs showing the results of a simulation comparing multiple randomly selected proportions of points for the origins of the toroidal shifts of the modified Syrjala test using unweighted absolute differences in the ECDFs (UWA). The grid column name indicates the CSR sample size (n_1), and the grid row indicates the shape of the second sample. The colors of the lines and symbols indicate the second sample size (approximate n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) on CSR realizations of 50 points with Right realizations of approximately 50, 100, 250, and 500 points.

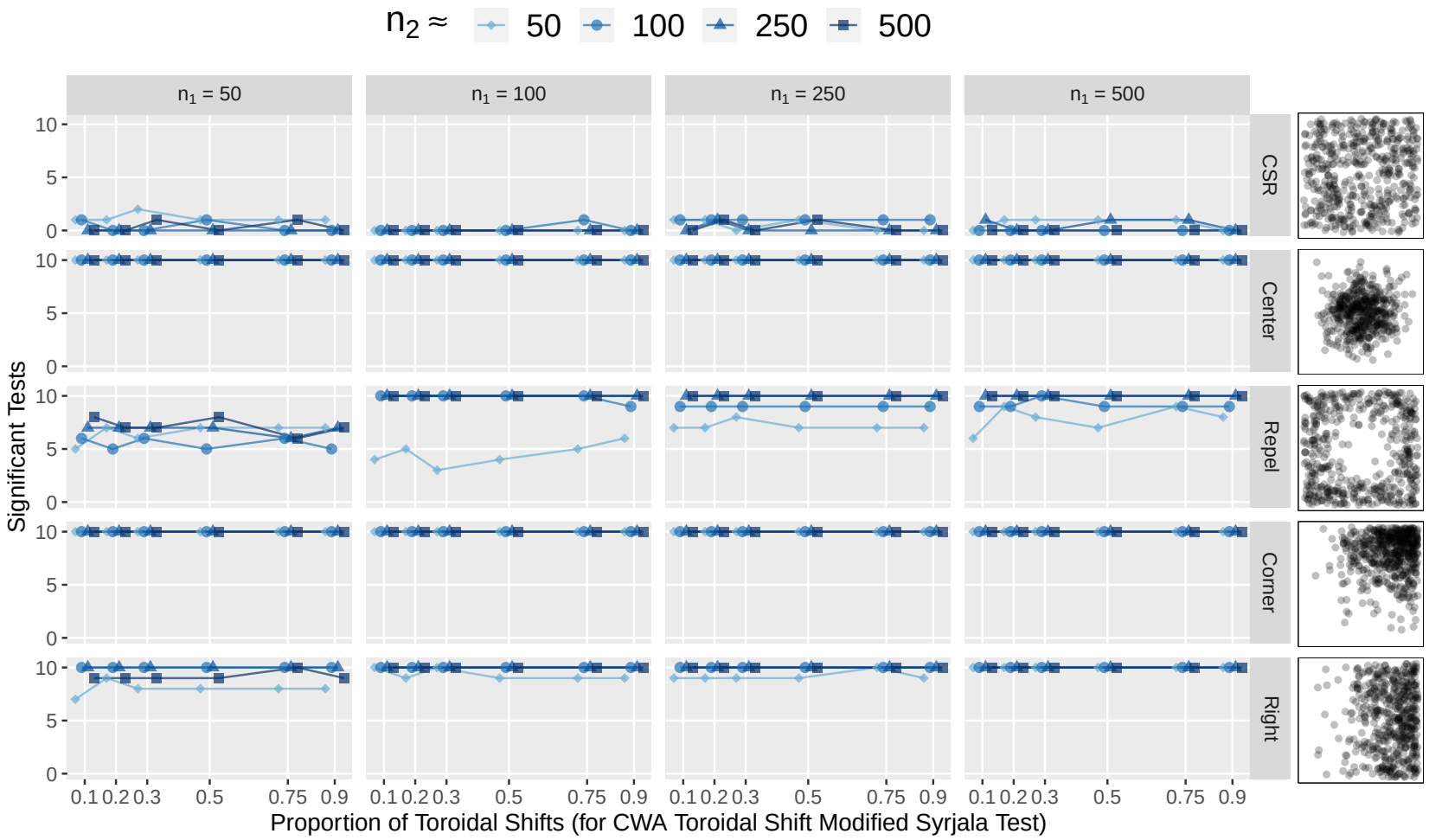


Fig. 71: A grid of line graphs showing the results of a simulation comparing multiple randomly selected proportions of points for the origins of the toroidal shifts of the modified Syrjala test using complementary weightings of the absolute differences in the ECDFs (CWA). The grid column name indicates the CSR sample size (n_1), and the grid row indicates the shape of the second sample. The colors of the lines and symbols indicate the second sample size (approximate n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) on CSR realizations of 50 points with Right realizations of approximately 50, 100, 250, and 500 points.

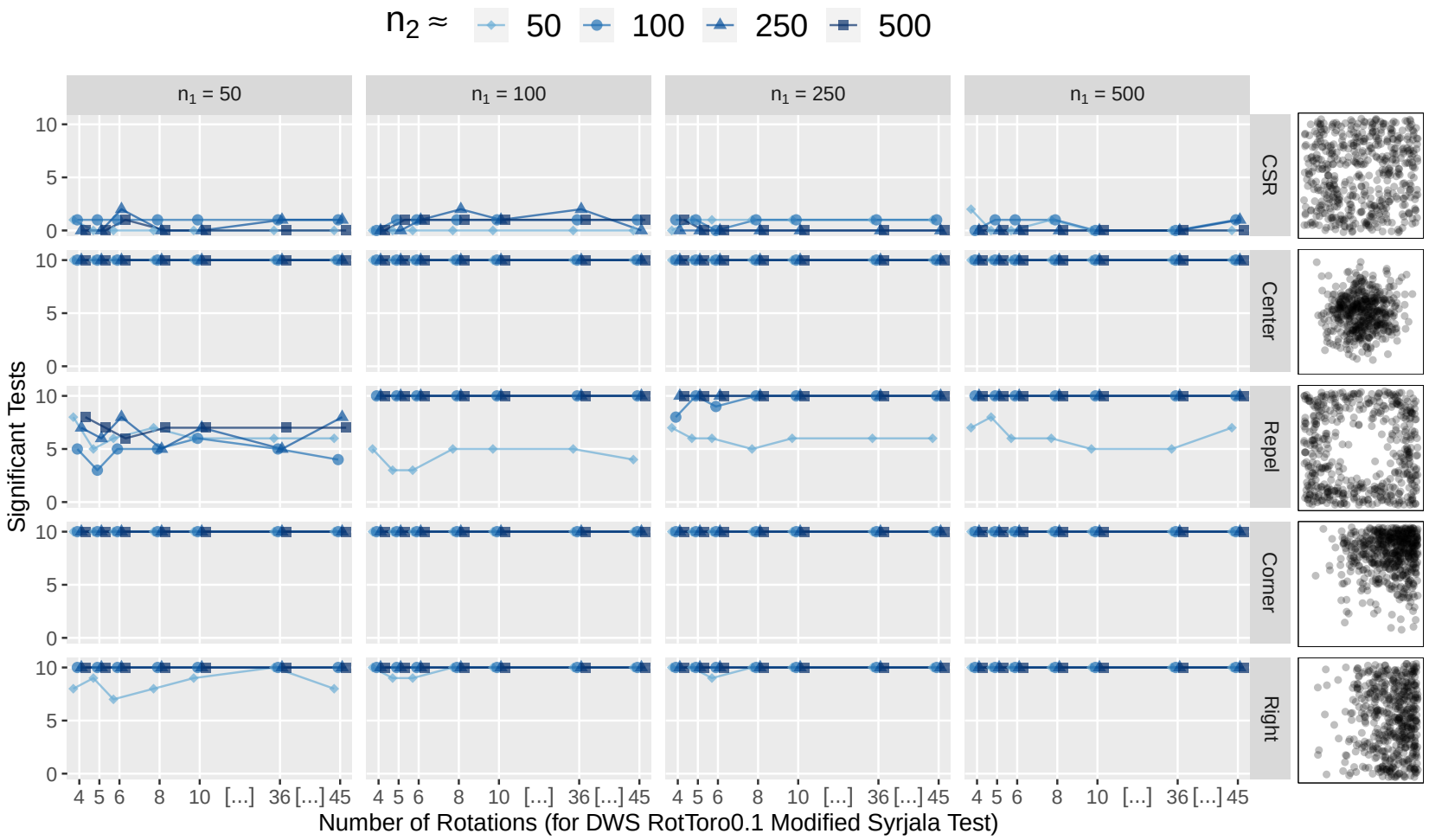


Fig. 72: A grid of line graphs showing the results of a simulation comparing multiple randomly selected proportions of points for the origins of the toroidal shifts of the modified Syrjala test across a number of rotations using double weightings of the squared differences in the ECDFs (DWS). The grid column name indicates the CSR sample size (n_1), and the grid row indicates the shape of the second sample. The colors of the lines and symbols indicate the second sample size (approximate n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) on CSR realizations of 50 points with Right realizations of approximately 50, 100, 250, and 500 points.

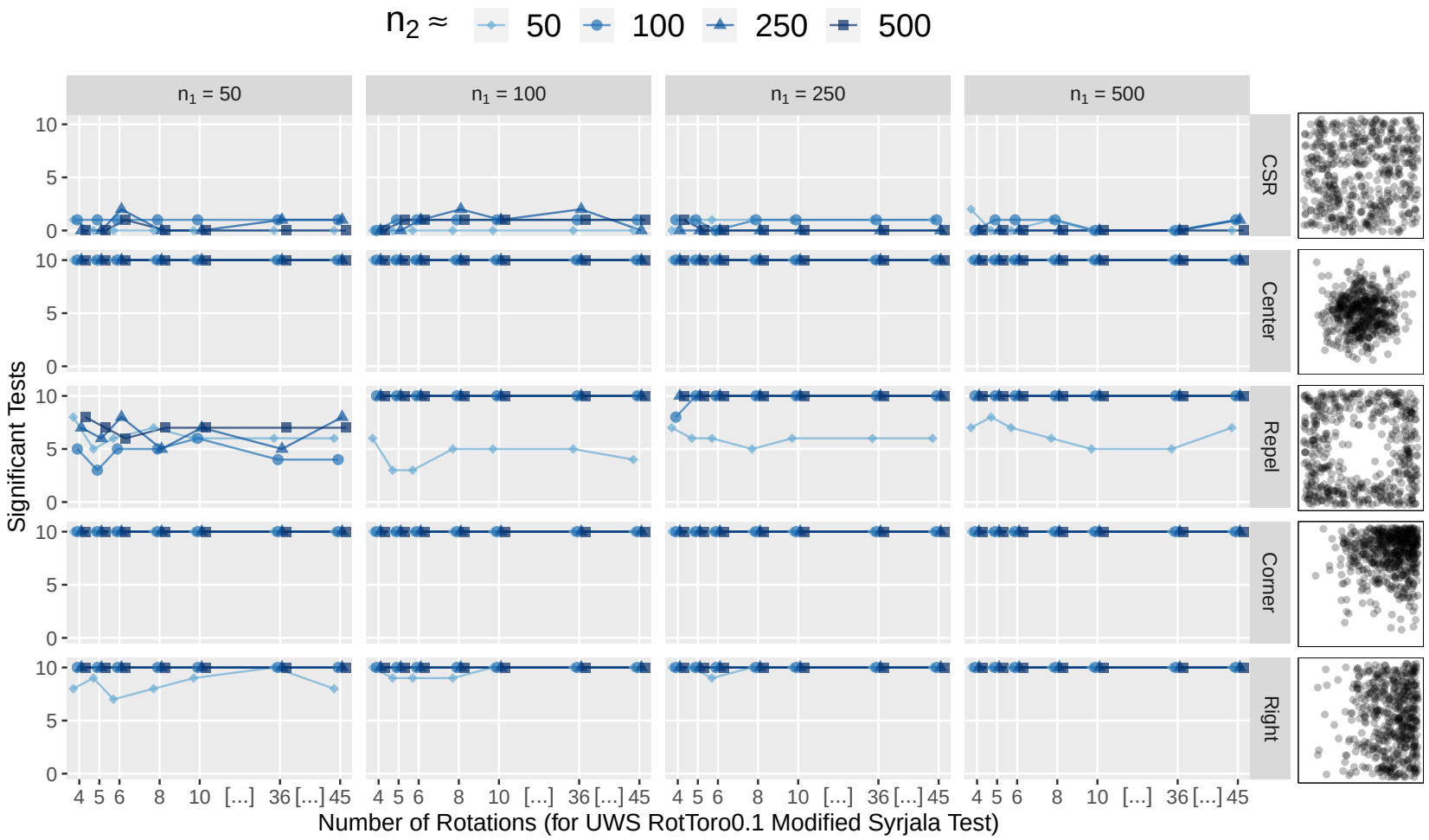


Fig. 73: A grid of line graphs showing the results of a simulation comparing multiple randomly selected proportions of points for the origins of the toroidal shifts of the modified Syrjala test across a number of rotations using uniform weightings of the squared differences in the ECDFs (UWS). The grid column name indicates the CSR sample size (n_1), and the grid row indicates the shape of the second sample. The colors of the lines and symbols indicate the second sample size (approximate n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) on CSR realizations of 50 points with Right realizations of approximately 50, 100, 250, and 500 points.

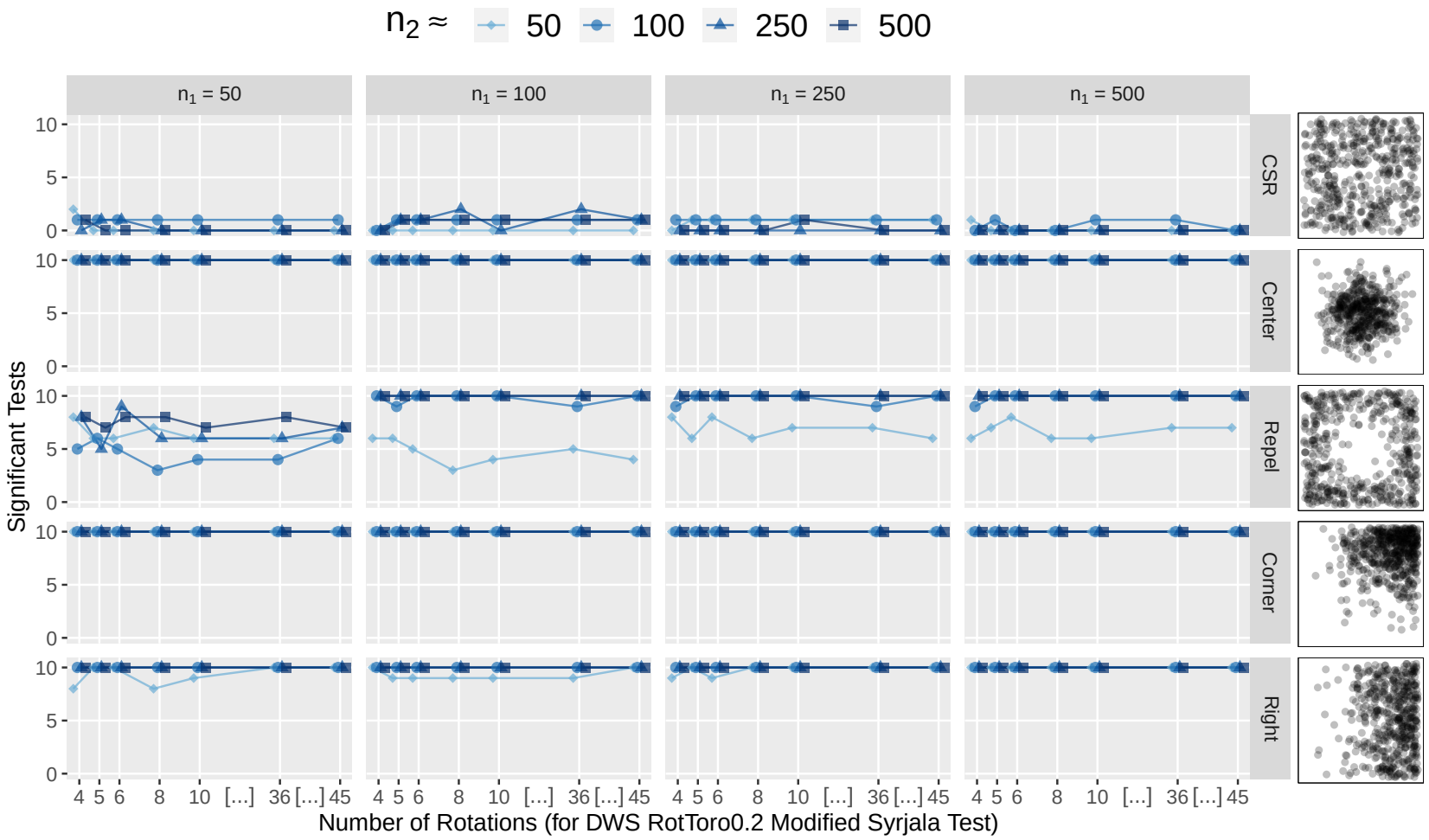


Fig. 74: A grid of line graphs showing the results of a simulation comparing multiple randomly selected proportions of points for the origins of the toroidal shifts of the modified Syrjala test across a number of rotations using double weightings of the squared differences in the ECDFs (DWS). The grid column name indicates the CSR sample size (n_1), and the grid row indicates the shape of the second sample. The colors of the lines and symbols indicate the second sample size (approximate n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) on CSR realizations of 50 points with Right realizations of approximately 50, 100, 250, and 500 points.

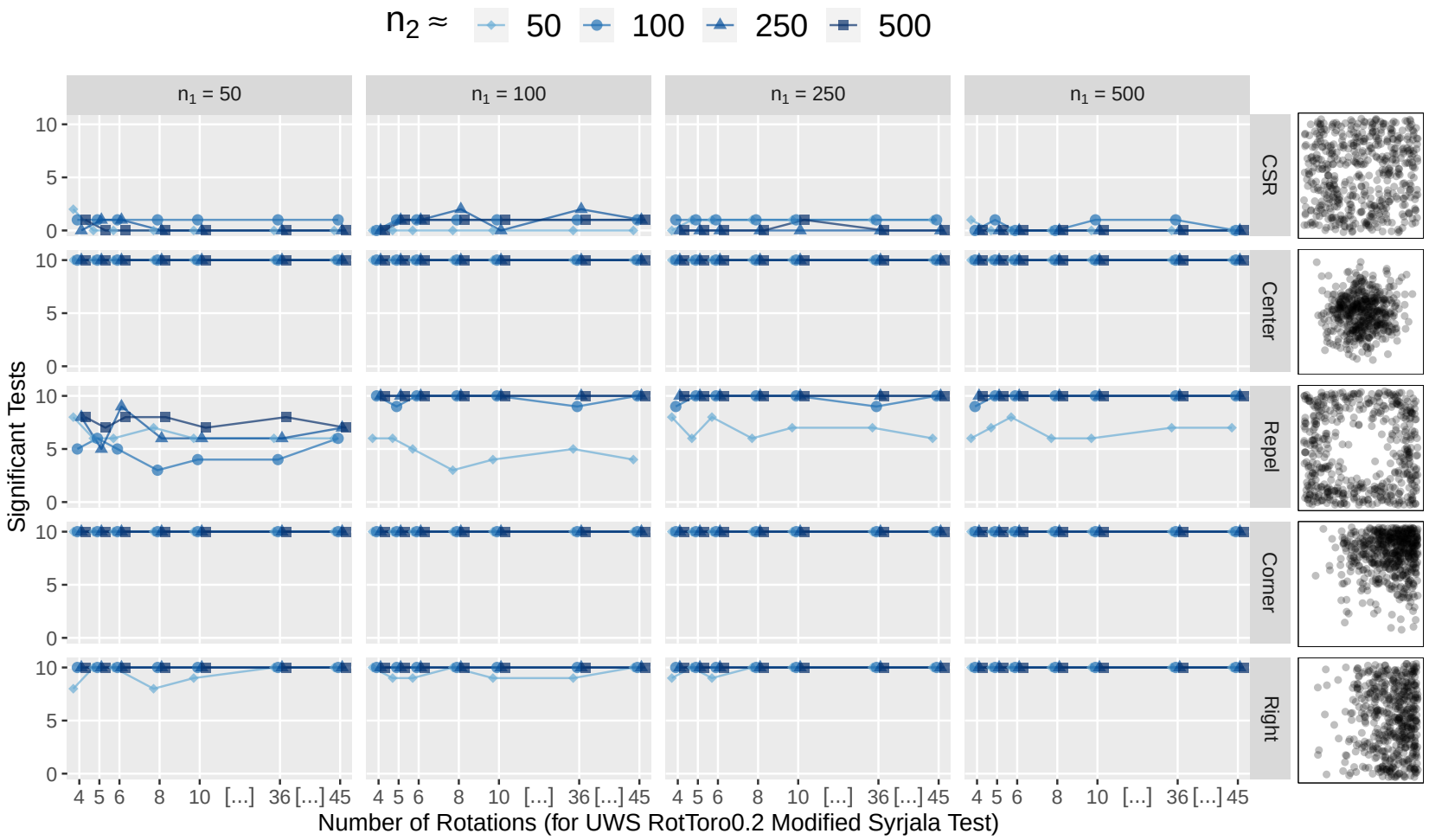


Fig. 75: A grid of line graphs showing the results of a simulation comparing multiple randomly selected proportions of points for the origins of the toroidal shifts of the modified Syrjala test across a number of rotations using uniform weightings of the squared differences in the ECDFs (UWS). The grid column name indicates the CSR sample size (n_1), and the grid row indicates the shape of the second sample. The colors of the lines and symbols indicate the second sample size (approximate n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) on CSR realizations of 50 points with Right realizations of approximately 50, 100, 250, and 500 points.

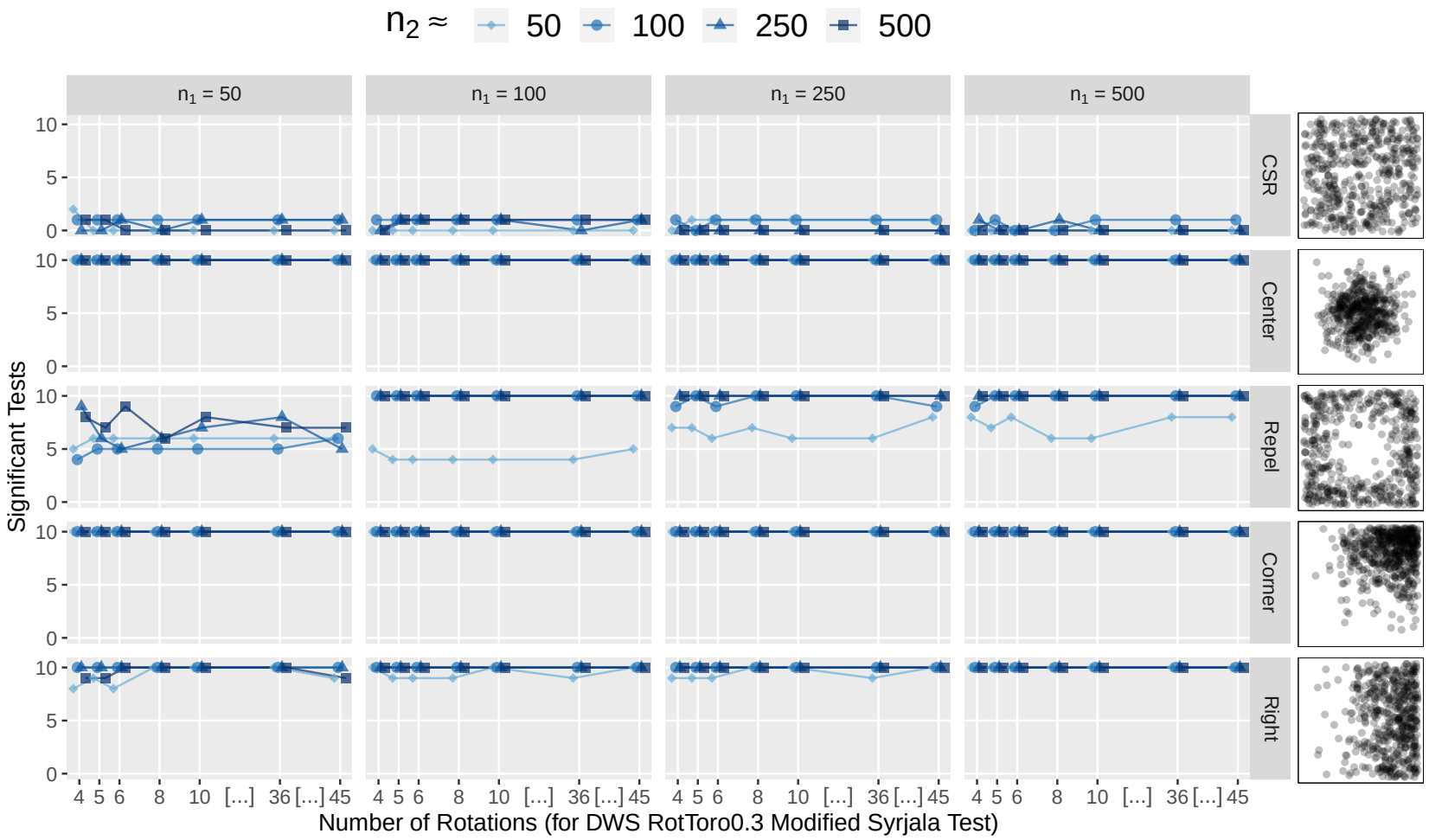


Fig. 76: A grid of line graphs showing the results of a simulation comparing multiple randomly selected proportions of points for the origins of the toroidal shifts of the modified Syrjala test across a number of rotations using double weightings of the squared differences in the ECDFs (DWS). The grid column name indicates the CSR sample size (n_1), and the grid row indicates the shape of the second sample. The colors of the lines and symbols indicate the second sample size (approximate n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) on CSR realizations of 50 points with Right realizations of approximately 50, 100, 250, and 500 points.

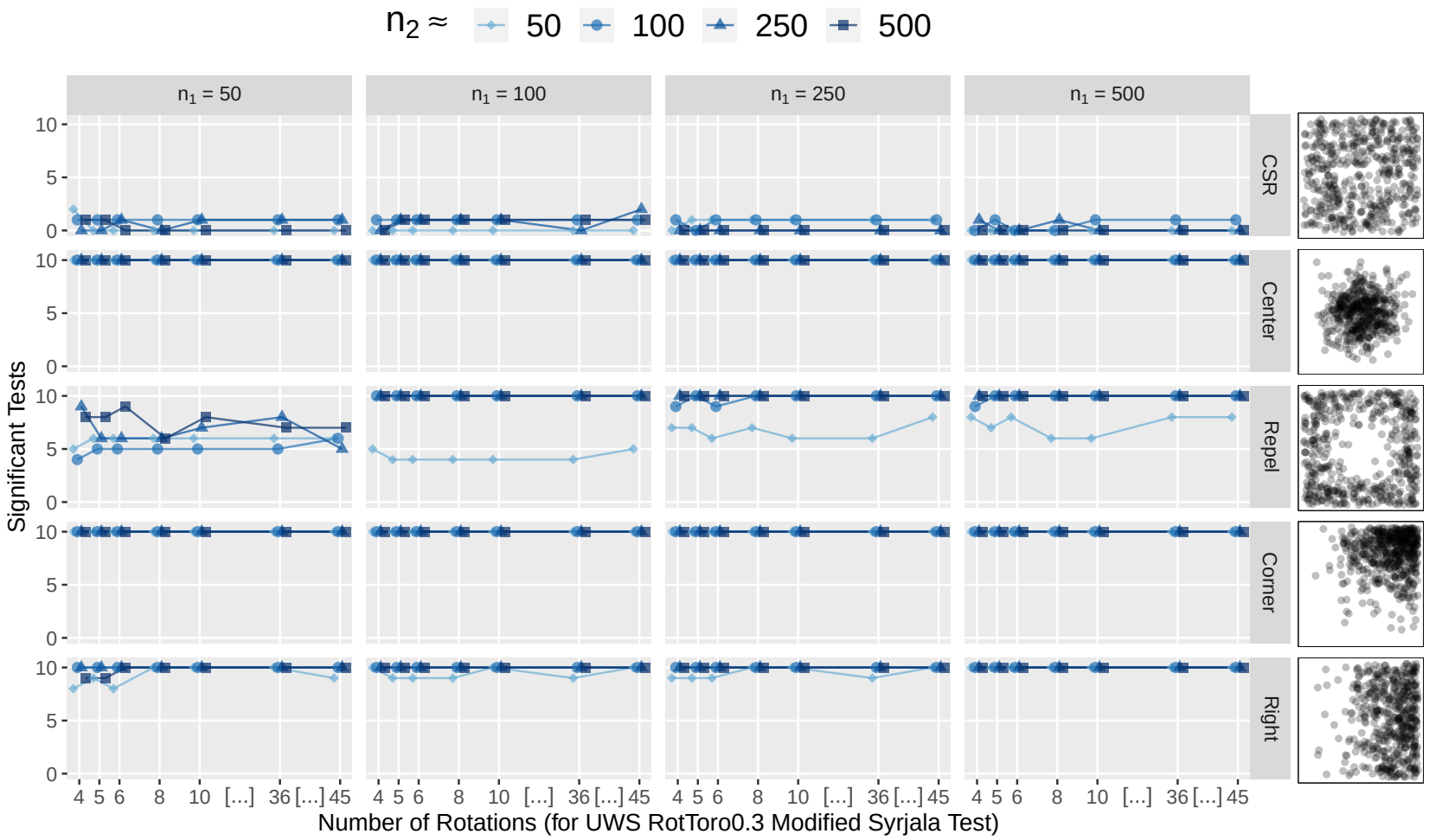


Fig. 77: A grid of line graphs showing the results of a simulation comparing multiple randomly selected proportions of points for the origins of the toroidal shifts of the modified Syrjala test across a number of rotations using uniform weightings of the squared differences in the ECDFs (UWS). The grid column name indicates the CSR sample size (n_1), and the grid row indicates the shape of the second sample. The colors of the lines and symbols indicate the second sample size (approximate n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) on CSR realizations of 50 points with Right realizations of approximately 50, 100, 250, and 500 points.

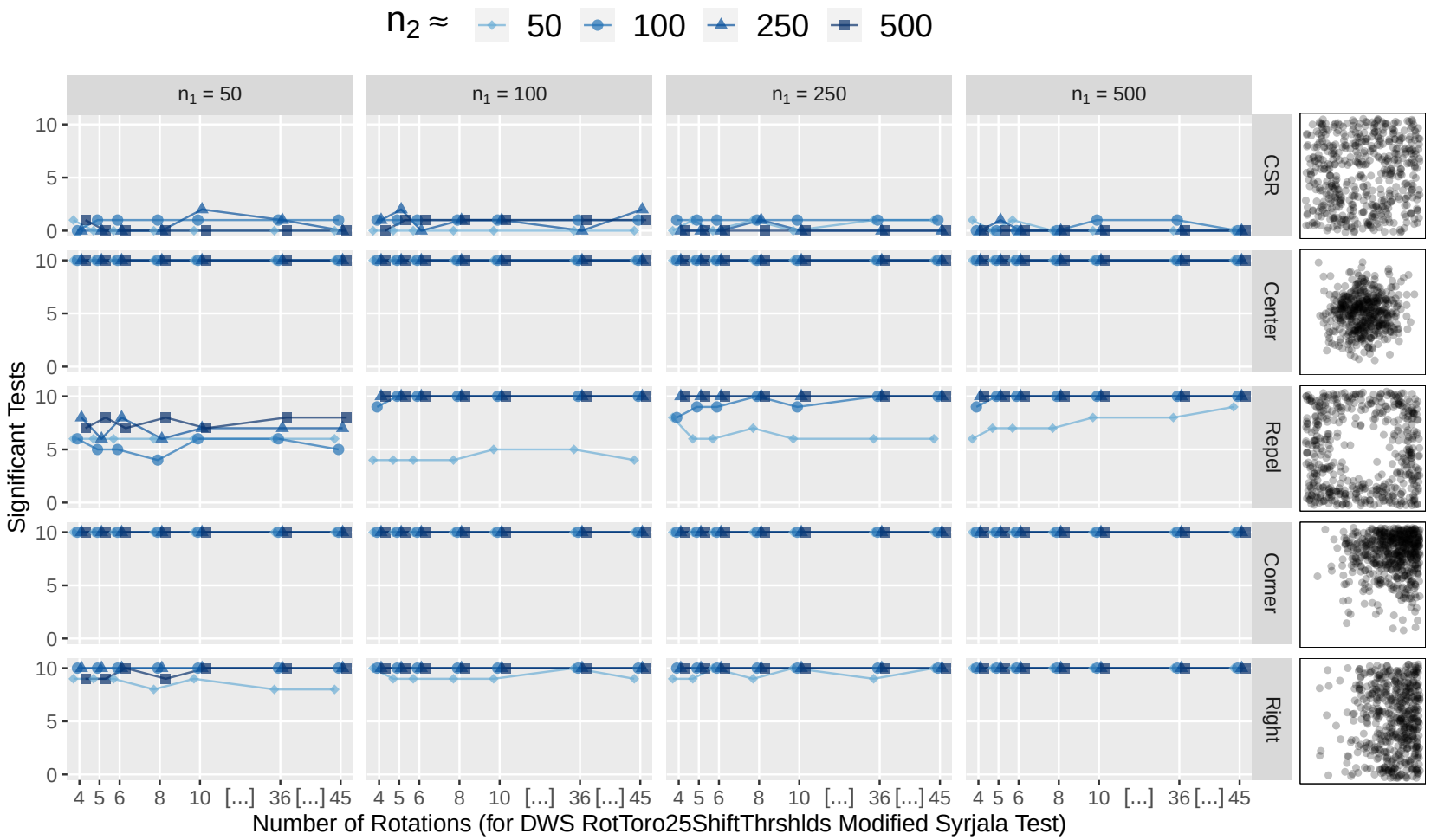


Fig. 78: A grid of line graphs showing the results of a simulation comparing toroidal shift thresholds of 25 points of the modified Syrjala test across a number of rotations using double weightings of the squared differences in the ECDFs (DWS). The grid column name indicates the CSR sample size (n_1), and the grid row indicates the shape of the second sample. The colors of the lines and symbols indicate the second sample size (approximate n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) on CSR realizations of 50 points with Right realizations of approximately 50, 100, 250, and 500 points.

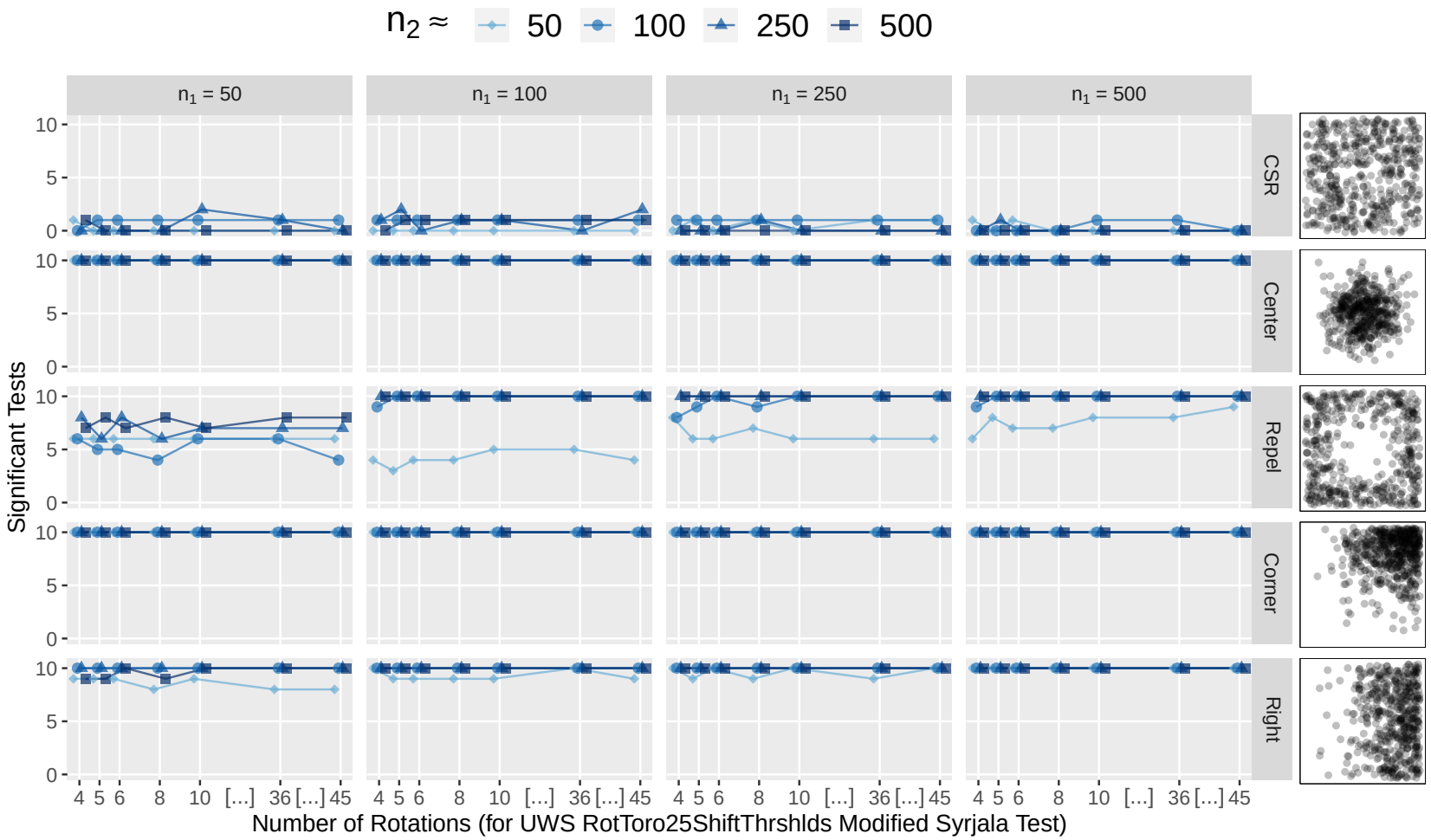


Fig. 79: A grid of line graphs showing the results of a simulation comparing toroidal shift thresholds of 25 points of the modified Syrjala test across a number of rotations using uniform weightings of the squared differences in the ECDFs (UWS). The grid column name indicates the CSR sample size (n_1), and the grid row indicates the shape of the second sample. The colors of the lines and symbols indicate the second sample size (approximate n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) on CSR realizations of 50 points with Right realizations of approximately 50, 100, 250, and 500 points.

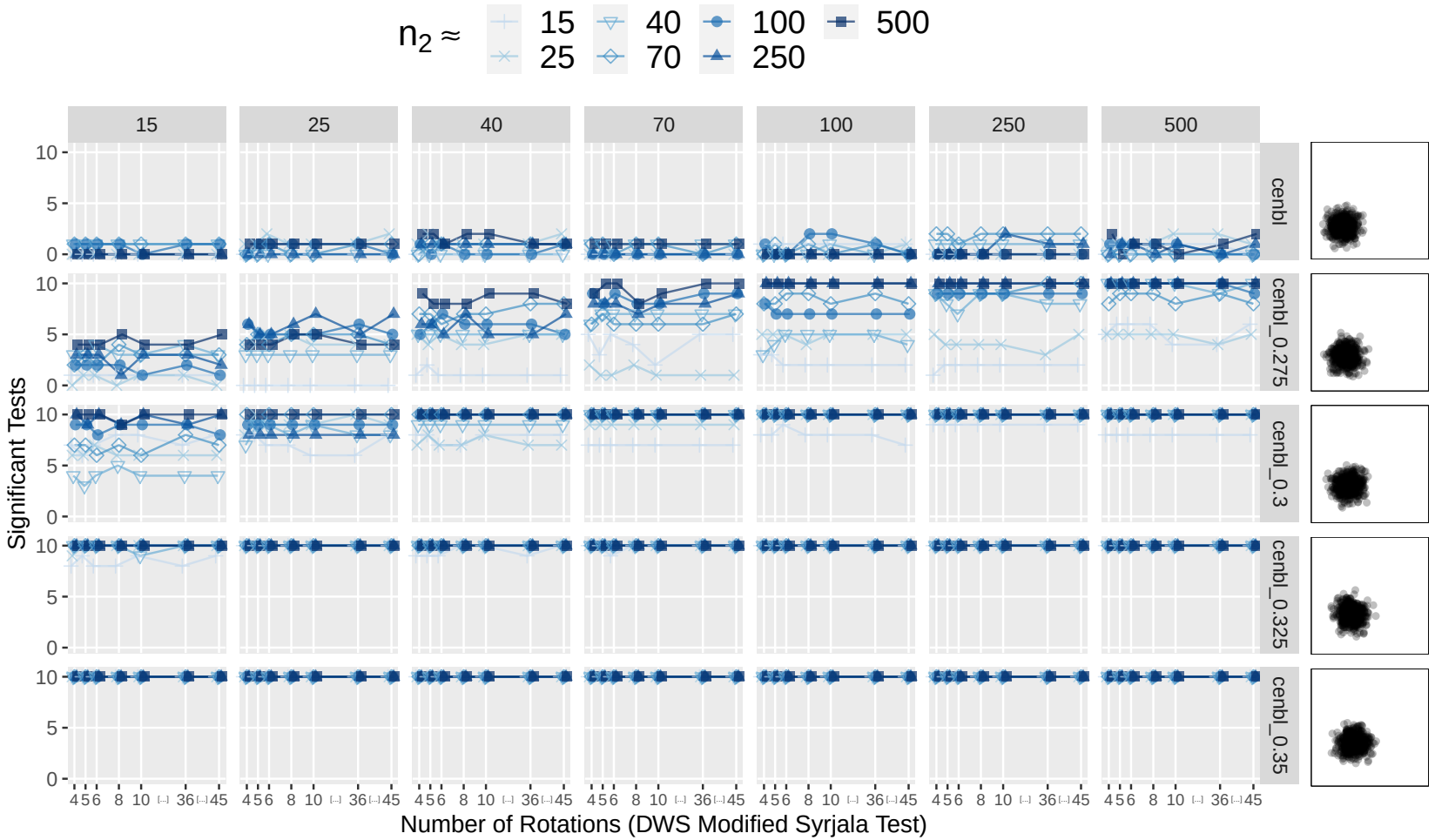


Fig. 80: A grid of line graphs showing the performance of the modified Syrjala test (using 0.1 proportion of points as origins of toroidal shifts, 8 rotations, and the DWS statistic) on simulated eye-tracking data where subjects concentrate on a single object at differing locations. The grid column name indicates the first sample size (n_1), and the grid row indicates the shape of the second sample. The shape of the first sample follows the same distribution exhibited in the first row. The horizontal axis indicates the second sample size (n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) between the first samples with 15 points and second samples with 15, 25, 40, 70, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

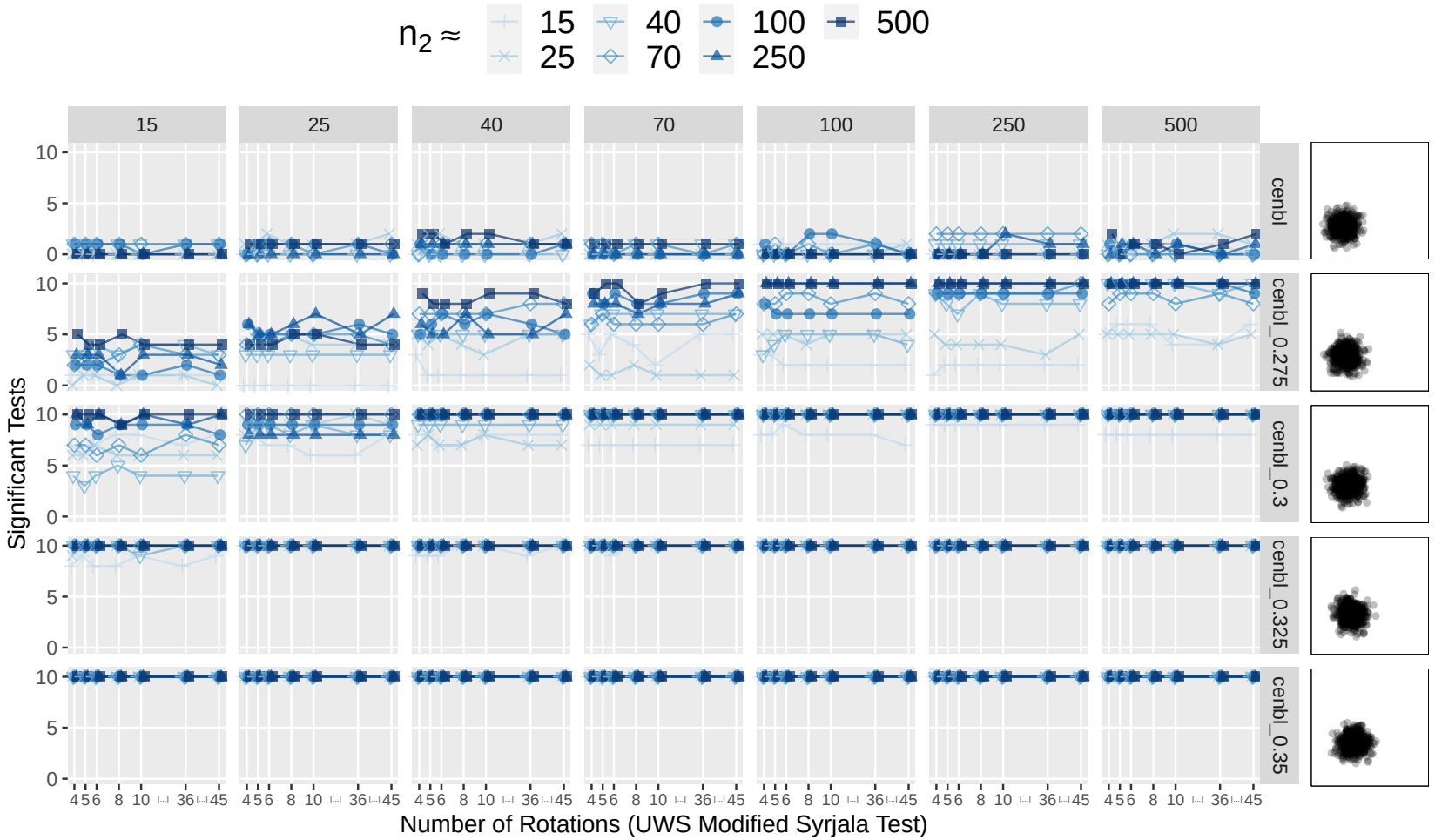


Fig. 81: A grid of line graphs showing the performance of the modified Syrjala test (using 0.1 proportion of points as origins of toroidal shifts, 8 rotations, and the UWS statistic) on simulated eye-tracking data where subjects concentrate on a single object at differing locations. The grid column name indicates the first sample size (n_1), and the grid row indicates the shape of the second sample. The shape of the first sample follows the same distribution exhibited in the first row. The horizontal axis indicates the second sample size (n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) between the first samples with 15 points and second samples with 15, 25, 40, 70, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

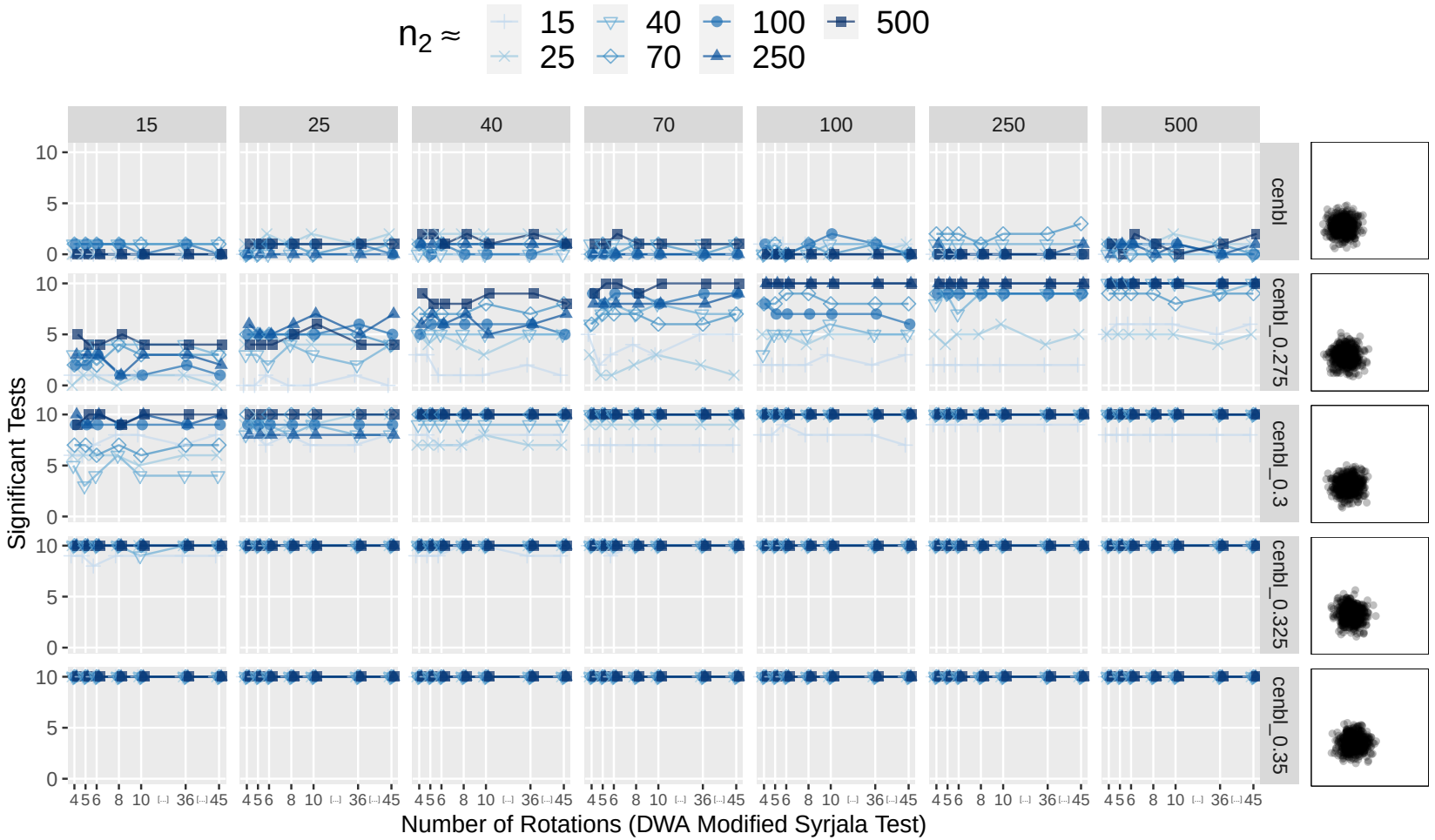


Fig. 82: A grid of line graphs showing the performance of the modified Syrjala test (using 0.1 proportion of points as origins of toroidal shifts, 8 rotations, and the DWA statistic) on simulated eye-tracking data where subjects concentrate on a single object at differing locations. The grid column name indicates the first sample size (n_1), and the grid row indicates the shape of the second sample. The shape of the first sample follows the same distribution exhibited in the first row. The horizontal axis indicates the second sample size (n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) between the first samples with 15 points and second samples with 15, 25, 40, 70, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

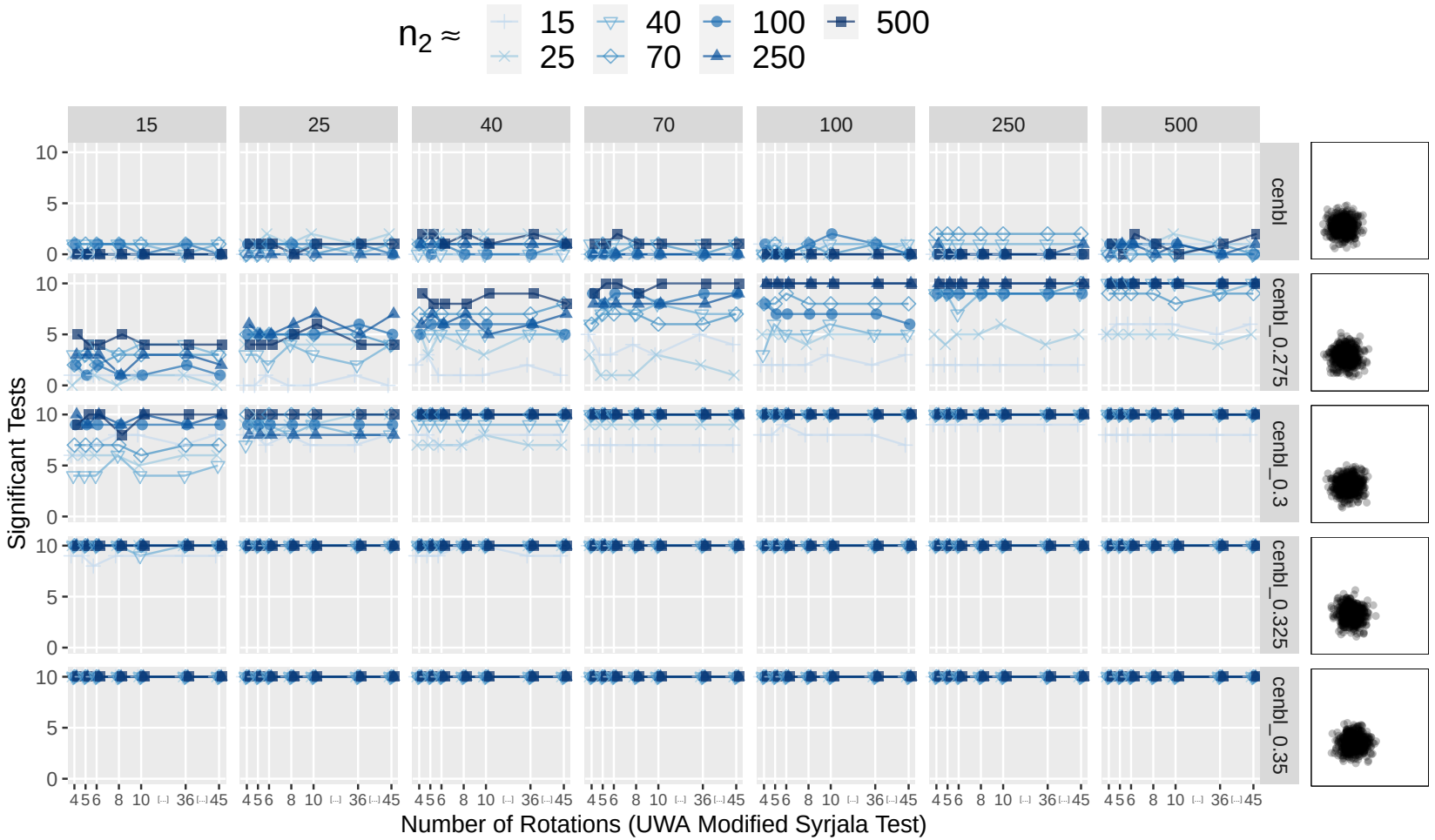


Fig. 83: A grid of line graphs showing the performance of the modified Syrjala test (using 0.1 proportion of points as origins of toroidal shifts, 8 rotations, and the UWA statistic) on simulated eye-tracking data where subjects concentrate on a single object at differing locations. The grid column name indicates the first sample size (n_1), and the grid row indicates the shape of the second sample. The shape of the first sample follows the same distribution exhibited in the first row. The horizontal axis indicates the second sample size (n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) between the first samples with 15 points and second samples with 15, 25, 40, 70, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

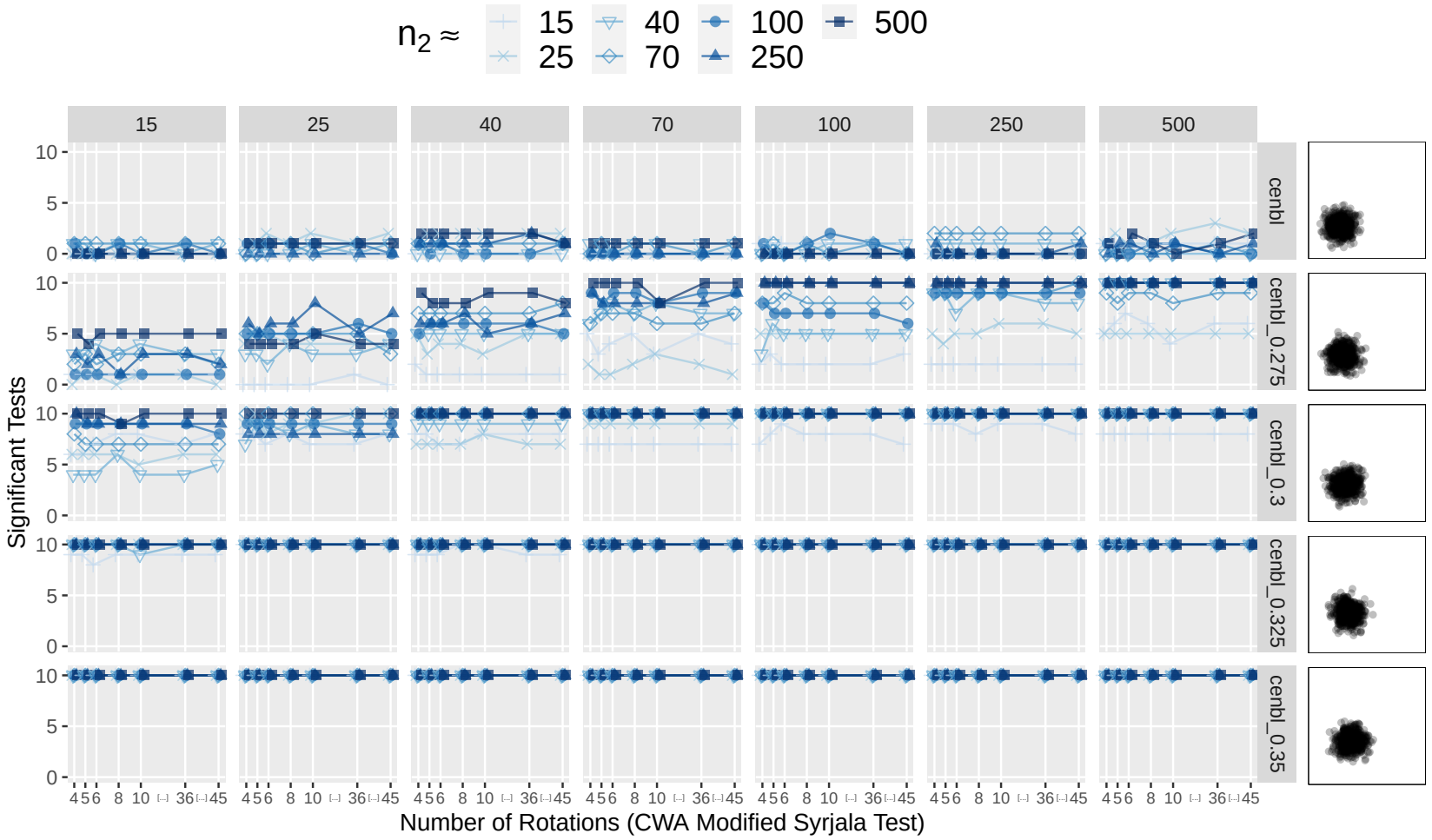


Fig. 84: A grid of line graphs showing the performance of the modified Syrjala test (using 0.1 proportion of points as origins of toroidal shifts, 8 rotations, and the CWA statistic) on simulated eye-tracking data where subjects concentrate on a single object at differing locations. The grid column name indicates the first sample size (n_1), and the grid row indicates the shape of the second sample. The shape of the first sample follows the same distribution exhibited in the first row. The horizontal axis indicates the second sample size (n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) between the first samples with 15 points and second samples with 15, 25, 40, 70, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

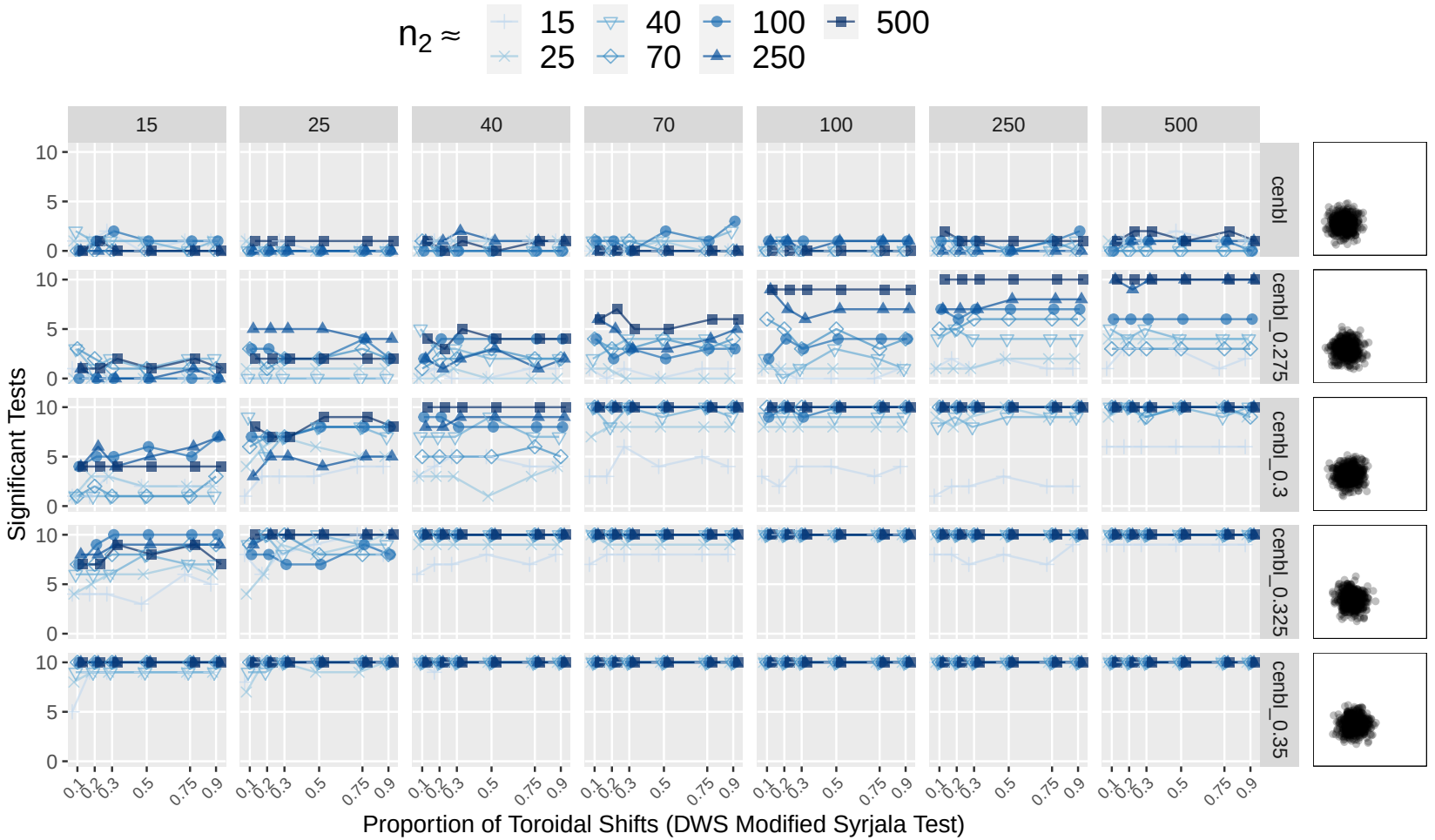


Fig. 85: A grid of line graphs showing the performance of the modified Syrjala test (using 0.1 proportion of points as origins of toroidal shifts, 8 rotations, and the DWS statistic) on simulated eye-tracking data where subjects concentrate on a single object at differing locations. The grid column name indicates the first sample size (n_1), and the grid row indicates the shape of the second sample. The shape of the first sample follows the same distribution exhibited in the first row. The horizontal axis indicates the second sample size (n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) between the first samples with 15 points and second samples with 15, 25, 40, 70, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

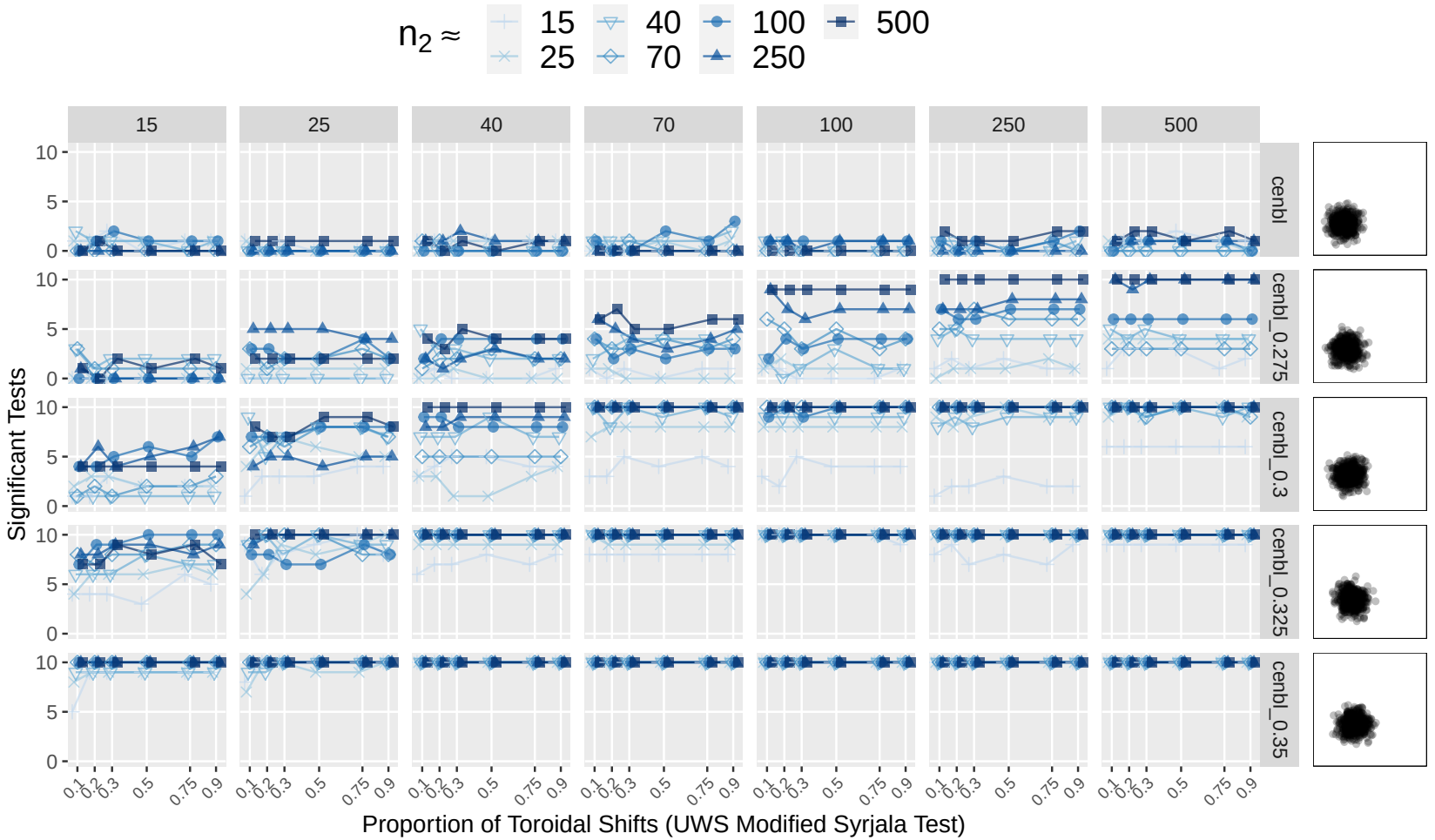


Fig. 86: A grid of line graphs showing the performance of the modified Syrjala test (using 0.1 proportion of points as origins of toroidal shifts, 8 rotations, and the UWS statistic) on simulated eye-tracking data where subjects concentrate on a single object at differing locations. The grid column name indicates the first sample size (n_1), and the grid row indicates the shape of the second sample. The shape of the first sample follows the same distribution exhibited in the first row. The horizontal axis indicates the second sample size (n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) between the first samples with 15 points and second samples with 15, 25, 40, 70, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

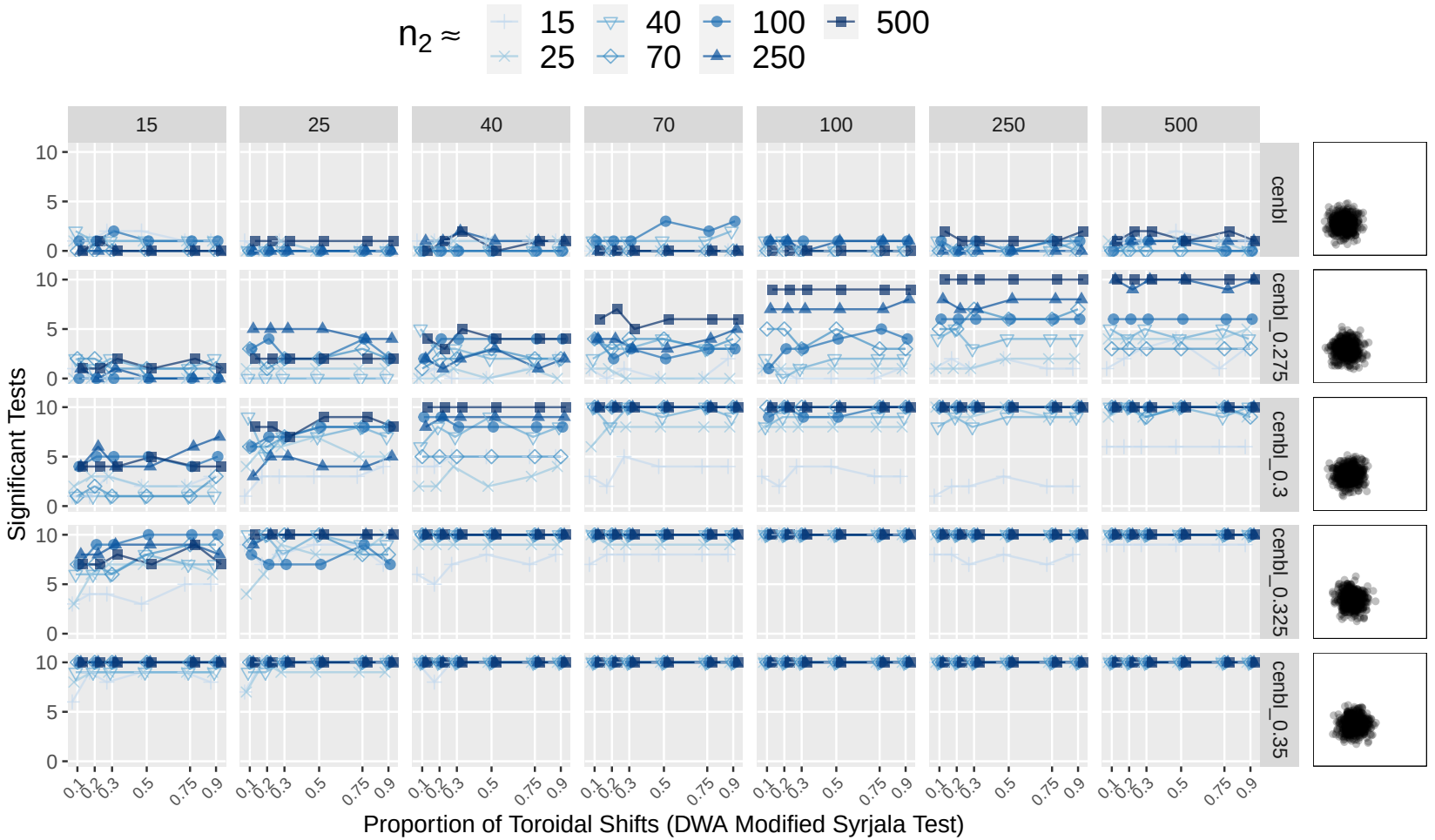


Fig. 87: A grid of line graphs showing the performance of the modified Syrjala test (using 0.1 proportion of points as origins of toroidal shifts, 8 rotations, and the DWA statistic) on simulated eye-tracking data where subjects concentrate on a single object at differing locations. The grid column name indicates the first sample size (n_1), and the grid row indicates the shape of the second sample. The shape of the first sample follows the same distribution exhibited in the first row. The horizontal axis indicates the second sample size (n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) between the first samples with 15 points and second samples with 15, 25, 40, 70, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

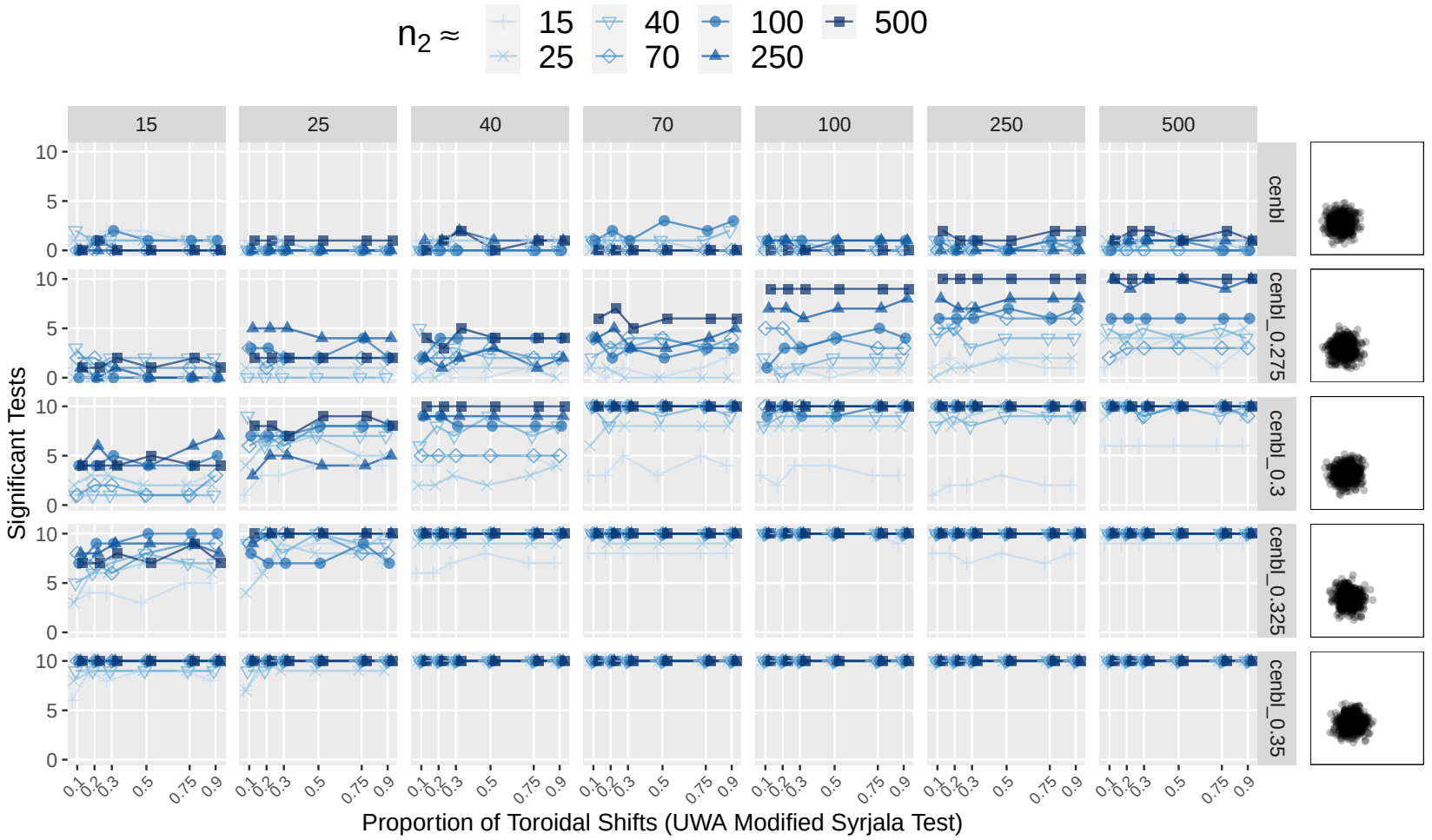


Fig. 88: A grid of line graphs showing the performance of the modified Syrjala test (using 0.1 proportion of points as origins of toroidal shifts, 8 rotations, and the UWA statistic) on simulated eye-tracking data where subjects concentrate on a single object at differing locations. The grid column name indicates the first sample size (n_1), and the grid row indicates the shape of the second sample. The shape of the first sample follows the same distribution exhibited in the first row. The horizontal axis indicates the second sample size (n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) between the first samples with 15 points and second samples with 15, 25, 40, 70, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

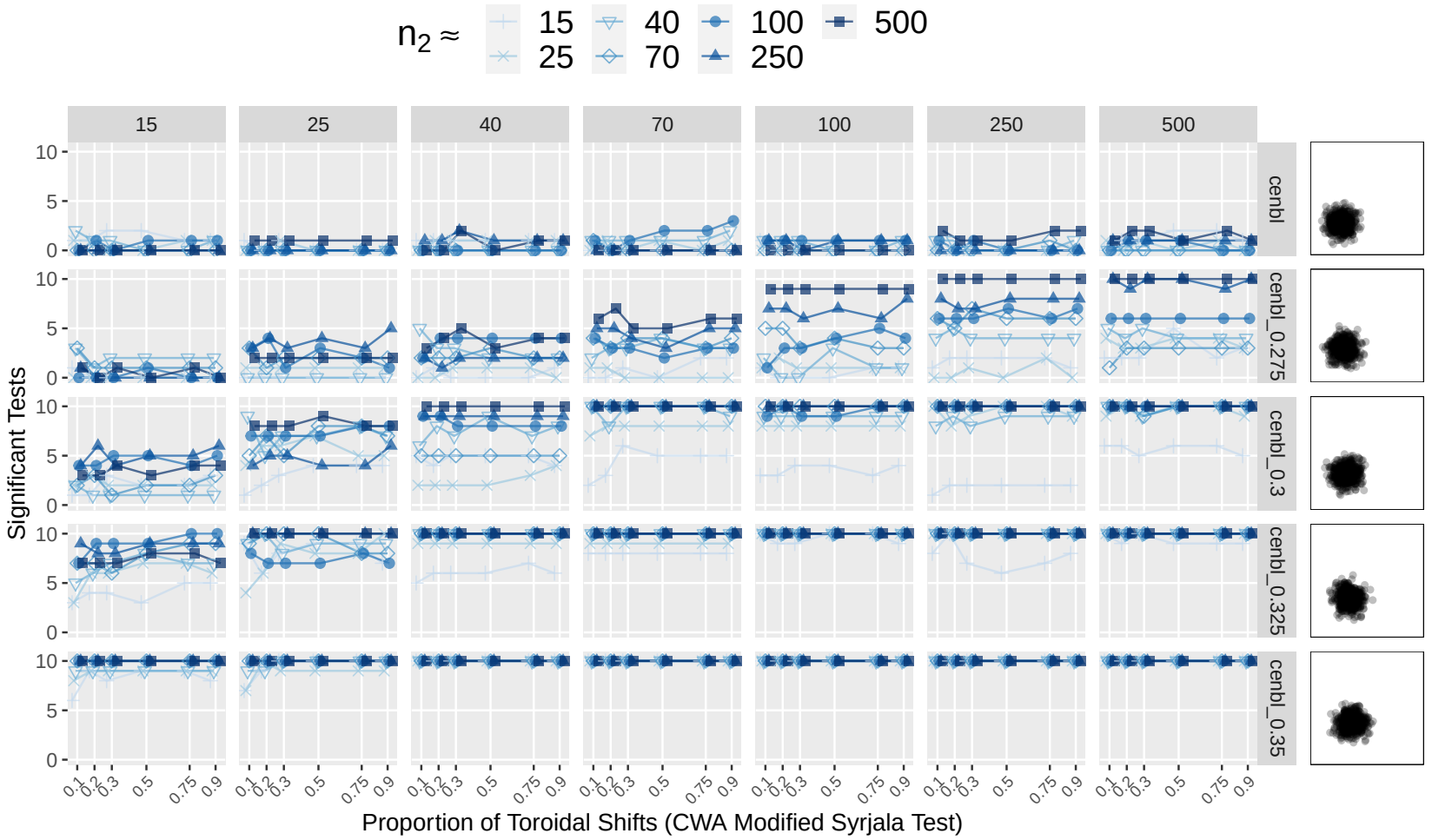


Fig. 89: A grid of line graphs showing the performance of the modified Syrjala test (using 0.1 proportion of points as origins of toroidal shifts, 8 rotations, and the CWA statistic) on simulated eye-tracking data where subjects concentrate on a single object at differing locations. The grid column name indicates the first sample size (n_1), and the grid row indicates the shape of the second sample. The shape of the first sample follows the same distribution exhibited in the first row. The horizontal axis indicates the second sample size (n_2). For example, the bottom left graph shows the number of significant test results (out of ten tests) between the first samples with 15 points and second samples with 15, 25, 40, 70, 100, 250, and 500 points. Note that the spaces between horizontal tick marks are only approximately represented.

APPENDIX C

Additional USU Posture Study Figures

This appendix provides further details of the setup and design of the USU Posture Study (Symanzik et al., 2017, 2018; Studenka et al., 2020; Coltrin et al., 2020; McKinney and Symanzik, 2019, 2021) described in more detail in Chapter 4.

C.1 Posture Identification Numbers

Figures 90–113 depict the 23 postures (24 if the calibration image is counted twice) shown to the subjects.

Fig. 90: Posture ID 0 in the USU Posture Study. This image served as the initial calibration image.

Fig. 91: Posture ID 1 in the USU Posture Study.

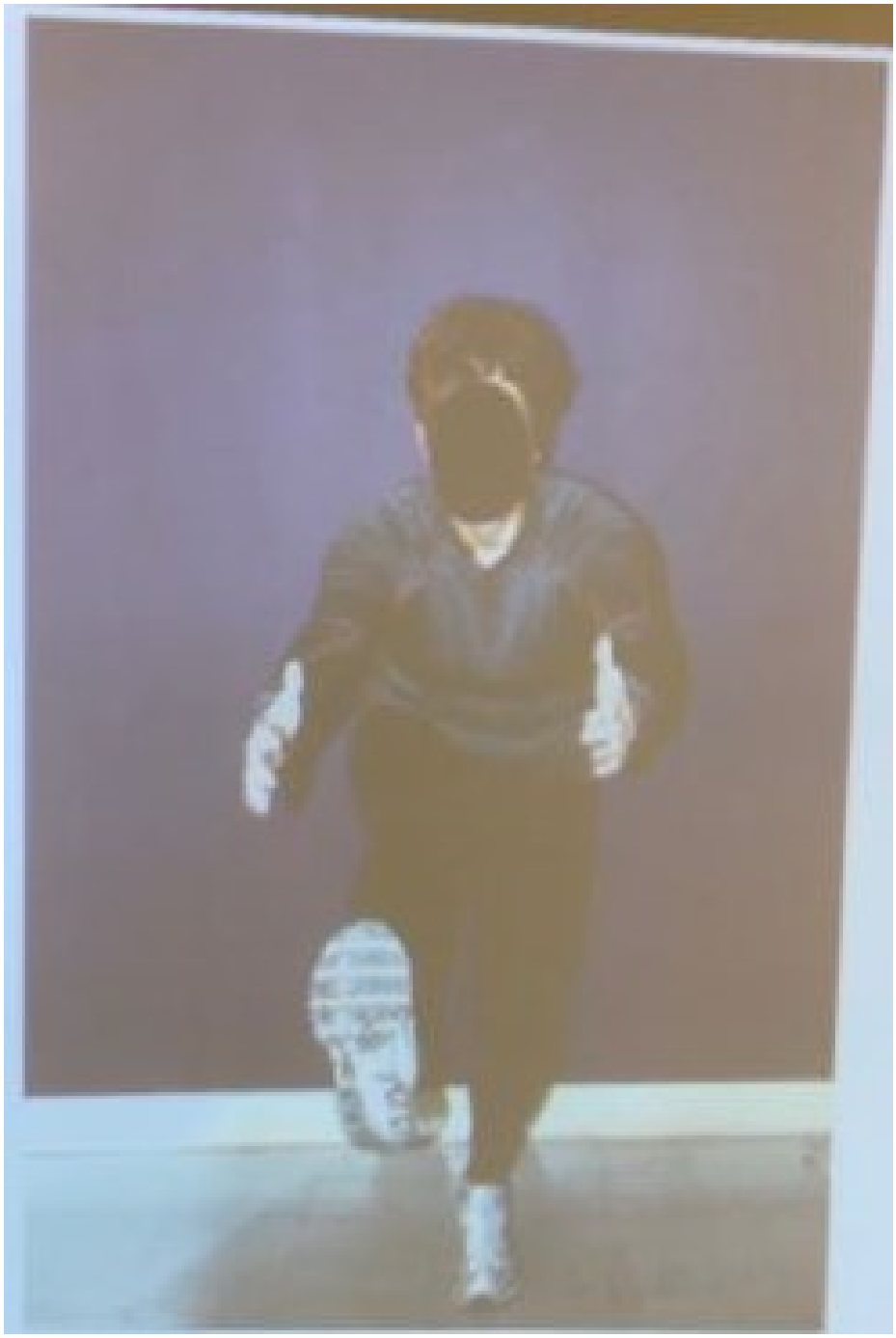


Fig. 92: Posture ID 2 in the USU Posture Study.

Fig. 93: Posture ID 3 in the USU Posture Study.

Fig. 94: Posture ID 4 in the USU Posture Study.

Fig. 95: Posture ID 5 in the USU Posture Study.

Fig. 96: Posture ID 6 in the USU Posture Study.

Fig. 97: Posture ID 7 in the USU Posture Study.

Fig. 98: Posture ID 8 in the USU Posture Study.

Fig. 99: Posture ID 9 in the USU Posture Study.

Fig. 100: Posture ID 10 in the USU Posture Study.

Fig. 101: Posture ID 11 in the USU Posture Study.

Fig. 102: Posture ID 12 in the USU Posture Study.

Fig. 103: Posture ID 13 in the USU Posture Study.

Fig. 104: Posture ID 14 in the USU Posture Study.

Fig. 105: Posture ID 15 in the USU Posture Study.

Fig. 106: Posture ID 16 in the USU Posture Study.

Fig. 107: Posture ID 17 in the USU Posture Study.

Fig. 108: Posture ID 18 in the USU Posture Study.

Fig. 109: Posture ID 19 in the USU Posture Study.

Fig. 110: Posture ID 20 in the USU Posture Study.

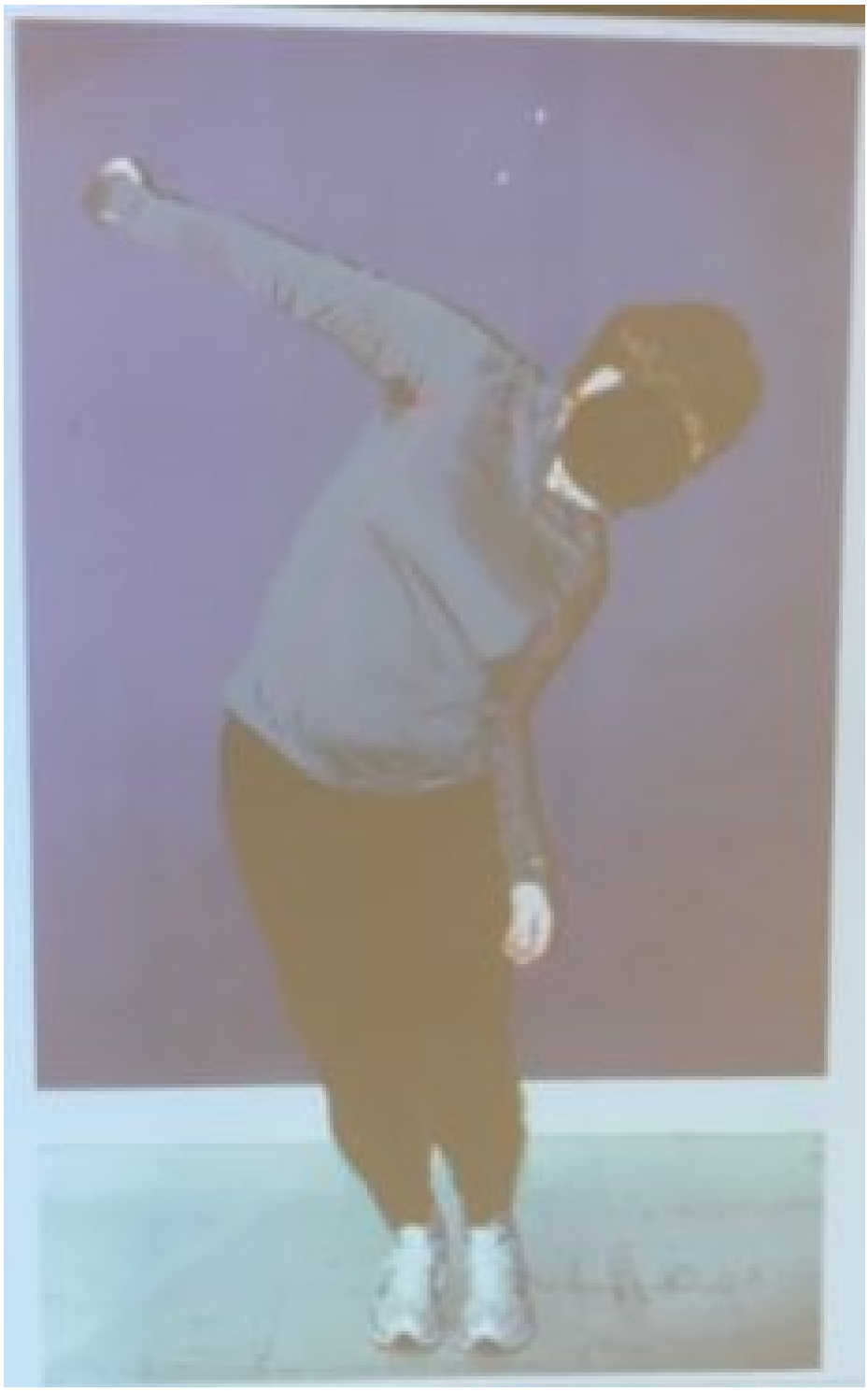


Fig. 111: Posture ID 21 in the USU Posture Study.

Fig. 112: Posture ID 22 in the USU Posture Study.

Fig. 113: Posture ID 23 in the USU Posture Study. This image served as the final calibration image and is identical to Figure 90.

REFERENCES

Anderson, N. H., Hall, P., Titterington, D. M., 1994. Two-Sample Test Statistics for Measuring Discrepancies Between Two Multivariate Probability Density Functions Using Kernel-Based Density Estimates. Journal of Multivariate Analysis 50 (1), 41–54.

Anderson, T. W., 1962. On the Distribution of the Two-Sample Cramér-von Mises Criterion. The Annals of Mathematical Statistics 33 (3), 1148–1159.

Anderson, T. W., Darling, D. A., 1952. Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes. The Annals of Mathematical Statistics 23 (2), 193–212.

Anderson, T. W., Darling, D. A., 1954. A Test of Goodness of Fit. Journal of the American Statistical Association 49 (268), 765–769.

Aronszajn, N., 1950. Theory of Reproducing Kernels. Transactions of the American Mathematical Society 68 (3), 337–404.

Baddeley, A., Turner, R., 2005. spatstat: An R Package for Analyzing Spatial Point Pattens. Journal of Statistical Software 12 (6), 1–42.
URL `http://www.jstatsoft.org/v12/i06/`

Baringhaus, L., Franz, C., 2004. On a New Multivariate Two-Sample Test. Journal of Multivariate Analysis 88 (1), 190–206.

Benayas, J. M. R., de la Montaã, E., Pérez-Camacho, L., de la Cruz, M., Moreno-Mateos, D., L., P. J., Seoane, S. S., Galván, I., 2010. Short-Term Dynamics and

Spatial Pattern of Nocturnal Birds Inhabiting a Mediterranean Agricultural Mosaic. Ardeola 57 (2), 303–320.

Benjamini, Y., Hochberg, Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological) 57 (1), 289–300.

Berman, M., 1986. Testing for Spatial Association Between a Point Process and Another Stochastic Process. Journal of the Royal Statistical Society: Series C (Applied Statistics) 35 (1), 54–62.

Berry, K. H., Coble, A. A., Yee, J. L., Mack, J. S., Perry, W. M., Anderson, K. M., Brown, M. B., 2015. Distance to Human Populations Influences Epidemiology of Respiratory Disease in Desert Tortoises. The Journal of Wildlife Management 79 (1), 122–136.

Berry, K. J., Johnston, J. E., Mielke Jr, P. W., 2011. Permutation Methods. Wiley Interdisciplinary Reviews: Computational Statistics 3 (6), 527–542.

Bickel, P. J., 1969. A Distribution Free Version of the Smirnov Two Sample Test in the p-Variate Case. The Annals of Mathematical Statistics 40 (1), 1–23.

Blascheck, T., Kurzhals, K., Raschke, M., Burch, M., Weiskopf, D., Ertl, T., 2017. Visualization of Eye Tracking Data: A Taxonomy and Survey. In: Computer Graphics Forum. Vol. 36 (8). Wiley Online Library, Hoboken, NJ, pp. 260–284.

Bonferroni, C. E., 1936. Teoria Statistica delle Classi e Calcolo delle Probabilità. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze 8, 3–62.

Botev, Z., Ridder, A., 2017. Variance Reduction. Wiley StatsRef: Statistics Reference Online, 1–6 https://doi.org/10.1002/9781118445112.stat07975.

Cantoni, V., Porta, M., De Maio, L., Distasi, R., Nappi, M., 2012. Towards a Novel Technique for Identification Based on Eye Tracking. In: Workshop on Biometric Measurements and Systems for Security and Medical Applications Proceedings. IEEE, Salerno, Italy, pp. 1–4.

Chetverikov, A., Kuvaldina, M., MacInnes, W. J., Jóhannesson, Ó. I., Kristjánsson, Á., 2018. Implicit Processing During Change Blindness Revealed with Mouse-Contingent and Gaze-Contingent Displays. Attention, Perception, & Psychophysics 80 (4), 844–859.

Chiu, S. N., Liu, K. I., 2009. Generalized Cramér-von Mises Goodness-of-Fit Tests for Multivariate Distributions. Computational Statistics & Data Analysis 53 (11), 3817–3834.

Choi, K., Marden, J., 1997. An Approach to Multivariate Rank Tests in Multivariate Analysis of Variance. Journal of the American Statistical Association 92 (440), 1581–1590.

Clark, P. J., Evans, F. C., 1954. Distance to Nearest Neighbor as a Measure of Spatial Relationships in Populations. Ecology 35 (4), 445–453.

Coltrin, J., McKinney, E., Studenka, B., Symanzik, J., 2020. Defining Areas of Interest for Eye-Tracking Data: Implementing a Systematic Approach. In: 2020 JSM Proceedings. American Statistical Association, Alexandria, VA, pp. 1144–1153.

Cramér, H., 1928. On the Composition of Elementary Errors: First Paper: Mathematical Deductions. Scandinavian Actuarial Journal 1928 (1), 13–74.

D'Agostino, R. B., 1986. Goodness-of-Fit Techniques. CRC Press, Boca Raton, FL.

Darling, D. A., 1957. The Kolmogorov-Smirnov, Cramer-von Mises Tests. The Annals of Mathematical Statistics 28 (4), 823–838.

Davison, A. C., Hinkley, D. V., 1997. Bootstrap Methods and their Application. Vol. 1. Cambridge University Press, New York, NY.

de la Cruz Rot, M., Maestre, F. T., Escudero, A., Bonet, A., 2008. Metodos para analizar datos puntuales. Asociacion Espanola de Ecologia Terrestre, Universidad Rey Juan Carlos and Caja de Ahorros del Mediterraneo, Madrid, Spain, Ch. 3, pp. 76–127.

Deniz, O., 2016. Group Eye Tracking. Ph.D. thesis, Department of Statistics, Middle East Technical University, http://etd.lib.metu.edu.tr/upload/12620362/index.pdf.

Díaz, E., Sebastian, R., Ayala, G., Díaz, M. E., Zoncu, R., Toomre, D., Gasman, S., 2008. Measuring Spatiotemporal Dependencies in Bivariate Temporal Random Sets with Applications to Cell Biology. IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (9), 1659–1671.

Diggle, P. J., Milne, R. K., 1983. Bivariate Cox Processes: Some Models for Bivariate Spatial Point Patterns. Journal of the Royal Statistical Society: Series B (Methodological) 45 (1), 11–21.

Dixon, P. M., 2014. Ripley's K Function. Wiley StatsRef: Statistics Reference Online, 1–12 https://doi.org/10.1002/9781118445112.stat07751.

Duchowski, A. T., 2007. Eye Tracking Methodology: Theory and Practice. Springer, New York, NY.

Edwards, W., 1977. How to Use Multiattribute Utility Measurement for Social Decisionmaking. IEEE Transactions on Systems, Man, and Cybernetics 7 (5), 326–340.

Epelboim, J., Suppes, P., 2001. A Model of Eye Movements and Visual Working Memory During Problem Solving in Geometry. Vision Research 41 (12), 1561–1574.

Fraser, D. A., 1951. Sequentially Determined Statistically Equivalent Blocks. The Annals of Mathematical Statistics 22 (3), 372–381.

Friedman, J. H., Rafsky, L. C., 1979. Multivariate Generalizations of the Wald-Wolfowitz and Smirnov Two-Sample Tests. The Annals of Statistics 7 (4), 697–717.

Fuller, T., Munguía, M., Mayfield, M., Sánchez-Cordero, V., Sarkar, S., 2006. Incorporating Connectivity into Conservation Planning: A Multi-Criteria Case Study from Central Mexico. Biological Conservation 133 (2), 131–142.

Glasserman, P., 2013. Monte Carlo Methods in Financial Engineering. Springer Science & Business Media, Berlin, Germany.

Glasserman, P., Yao, D. D., 1992. Some Guidelines and Guarantees for Common Random Numbers. Management Science 38 (6), 757–911.

Goldberg, J. H., Helfman, J. I., 2010. Comparing Information Graphics: A Critical Look at Eye Tracking. In: Proceedings of the 3rd BELIV'10 Workshop: BEyond time and errors: novel evaLuation methods for Information Visualization. Association for Computing Machinery, New York, NY, pp. 71–78.

Gordon, P. C., Moser, S., 2007. Insight into Analogies: Evidence from Eye Movements. Visual Cognition 15 (1), 20–35.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., Smola, A., 2012. A Kernel Two-Sample Test. Journal of Machine Learning Research 13 (25), 723–773.

Hall, P., 1984. Central Limit Theorem for Integrated Square Error of Multivariate Nonparametric Density Estimators. Journal of Multivariate Analysis 14 (1), 1–16.

Hall, P., Tajvidi, N., 2002. Permutation Tests for Equality of Distributions in High-Dimensional Settings. Biometrika 89 (2), 359–374.

Harris, C., 1993. On the Reversibility of Markov Scanning in Free Viewing. In: Gale, A. G., Carr, K., Brogan, D. (Eds.), Visual Search 2: Proceedings of the 2nd International Conference on Visual Search. Taylor & Francis, London, United Kingdom, pp. 123–135.

Henze, N., 1988. A Multivariate Two-Sample Test Based on the Number of Nearest Neighbor Type Coincidences. The Annals of Statistics 16 (2), 772–783.

Hessels, R. S., Kemner, C., van den Boomen, C., Hooge, I. T., 2016. The Area-Of-Interest Problem in Eyetracking Research: A Noise-Robust Solution for Face and Sparse Stimuli. Behavior Research Methods 48 (4), 1694–1712.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., Van de Weijer, J., 2011. Eye Tracking: A Comprehensive Guide to Methods and Measures. Oxford University Press, Oxford, United Kingdom.

Irwin, D. E., 1992. Visual Memory Within and Across Fixations. In: Rayner, K. (Ed.), Eye Movements and Visual Cognition. Springer, New York, NY, pp. 146–165.

Jonckheere, A. R., 1954. A Distribution-Free k-Sample Test Against Ordered Alternatives. Biometrika 41 (1/2), 133–145.

Karatzoglou, A., Smola, A., Hornik, K., 2019. kernlab: Kernel-Based Machine Learning Lab. R package version 0.9-29 (`https://cran.r-project.org/web/packages/kernlab/index.html`.

Kleijnen, J. P. C., 1975. Antithetic Variates, Common Random Numbers and Optimal Computer Time Allocation in Simulation. Management Science 21 (10), 1189–1214.

Kleijnen, J. P. C., 1976. Comparing Means and Variances of Two Simulations. Simulation 26 (3), 87–88.

Kleijnen, J. P. C., 1979. Analysis of Simulation with Common Random Numbers: A Note on Heikes et al.(1976). ACM SIGSIM Simulation Digest 11 (2), 7–13.

Kolmogorov, A., 1933. Sulla Determinazione Empirica di una Legge di Distributione. Giornale dell'Istituto Italiano degli Attuari 1933 (4), 83–91, (In Italian).

Krejtz, K., Duchowski, A., Szmidt, T., Krejtz, I., González Perilli, F., Pires, A., Vilaro, A., Villalobos, N., 2015. Gaze Transition Entropy. Association for Computing Machinery Transactions on Applied Perception 13 (1), 1–20.

Kruskal, W. H., Wallis, W. A., 1952. Use of Ranks in One-Criterion Variance Analysis. Journal of the American Statistical Association 47 (260), 583–621.

Kullback, S., Leibler, R. A., 1951. On Information and Sufficiency. The Annals of Mathematical Statistics 22 (1), 79–86.

Kumar, A., Burch, M., Brand, I., Castelijns, L., Ritchi, F., Rooks, F., Timmermans, N., Mueller, K., 2018. Eye Tracking for Exploring Visual Communication Differences. In: Transactions on Visualization & Computer Graphics. IEEE, Berlin, Germany, pp. 1–5.

Lai, L., Patel, J. H., Rinderknecht, T., Cheng, W., 2005. Hardware Efficient LBIST with Complementary Weights. In: International Conference on Computer Design. IEEE, pp. 479–481.

Li, C., 2017. Extracting and Visualizing Data from Mobile and Static Eye Trackers in R and Matlab. Ph.D. thesis, Department of Mathematics and Statistics, Utah State University, Logan, UT, https://digitalcommons.usu.edu/etd/6880/.

Lotwick, H., Silverman, B., 1982. Methods for Analysing Spatial Processes of Several Types of Points. Journal of the Royal Statistical Society: Series B (Methodological) 44 (3), 406–413.

Mann, H. B., Whitney, D. R., 1947. On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other. The Annals of Mathematical Statistics 18 (1), 50–60.

Matsuda, N., Takeuchi, H., 2012. Do Heavy and Light Users Differ in the Web-Page Viewing Patterns? Analysis of Their Eye-Tracking Records by Heat Maps and Networks of Transitions. International Journal of Computer Information Systems and Industrial Management Applications 4 (1), 109–120.

McAdam, B. J., Grabowski, T. B., Marteinsdóttir, G., 2012. Testing for Differences in Spatial Distributions from Individual Based Data. Fisheries Research 127 (1), 148–153.

McKinney, E., Symanzik, J., 2019. Modifications of the Syrjala Test for Testing Spatial Distribution Differences Between Two Populations. In: 2019 JSM Proceedings. American Statistical Association, Alexandria, VA, pp. 2518–2530.

McKinney, E., Symanzik, J., 2021. Extensions to the Syrjala Test with Eye-Tracking Analysis Applications. In: 2021 JSM Proceedings. American Statistical Association, Alexandria, VA, pp. 853–889.

Miller, R. G. J., 1981. Simultaneous Statistical Inference. Springer-Verlag, New York, NY.

Mondal, P. K., Biswas, M., Ghosh, A. K., 2015. On High Dimensional Two-Sample Tests Based on Nearest Neighbors. Journal of Multivariate Analysis 141 (9), 168–178.

Moreno-Fernández, D., Ledo, A., Cañellas, I., Montes, F., 2020. Strategies for Modeling Regeneration Density in Relation to Distance from Adult Trees. Forests 11 (1), 120.

Pettitt, A. N., 1976. A Two-Sample Anderson-Darling Rank Statistic. Biometrika 63 (1), 161–168.

Pieters, R., Rosbergen, E., Wedel, M., 1999. Visual Attention to Repeated Print Advertising: A Test of Scanpath Theory. Journal of Marketing Research 36 (4), 424–438.

Pratt, J. W., 1964. Robustness of Some Procedures for the Two-Sample Location Problem. Journal of the American Statistical Association 59 (307), 665–680.

R Core Team, 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, `http://www.R-project.org/`.

Rabiner, L., Juang, B., 1986. An Introduction to Hidden Markov Models. IEEE ASSP Magazine 3 (1), 4–16.

Rahmatallah, Y., Zybailov, B., Emmert-Streib, F., Glazko, G., 2017. GSAR: Bioconductor Package for Gene Set Analysis in R. BMC Bioinformatics 18 (61), 1–12.

Rey, W. J. J., 1986. Multivariate Data Analysis, Contributions and Shortcomings of Robustness in Practice. In: De Antoni, F., Lauro, N. (Eds.), Proceedings in Computational Statistics. Physica-Verlag, Heidelberg, Germany, pp. 197–204.

Rice, J. A., 2006. Mathematical Statistics and Data Analysis, 3rd Edition. Brooks/Cole, Cengage Learning, Belmont, CA.

Ripley, B. D., 1976. The Second-Order Analysis of Stationary Point Processes. Journal of Applied Probability 13 (2), 255–266.

Rizzo, M. L., Székely, G. J., 2016. Energy Distance. Wiley Interdisciplinary Reviews: Computational Statistics 8 (1), 27–38.

Rizzo, M. L., Szekely, G. J., 2019. energy: E-Statistics: Multivariate Inference via the Energy of Data. R package version 1.7-7 (`https://cran.r-project.org/web/packages/energy/index.html`).

Ruddock, K., Wooding, D., Mannan, S., 1995. Automatic Control of Saccadic Eye Movements Made in Visual Inspection of Briefly Presented 2-D Images. Spatial Vision 9 (3), 363–386.

Rudin, W., 1964. Principles of Mathematical Analysis, 3rd Edition. McGraw-Hill, New York City, NY.

Schilling, M. F., 1986. Multivariate Two-Sample Tests Based on Nearest Neighbors. Journal of the American Statistical Association 81 (395), 799–806.

Scholz, F. W., Stephens, M. A., 1987. K-sample Anderson–Darling Tests. Journal of the American Statistical Association 82 (399), 918–924.

Šidák, Z., 1967. Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. Journal of the American Statistical Association 62 (318), 626–633.

Song, L., June 2008. Learning via Hilbert Space Embedding of Distributions. Ph.D. thesis, University of Sydney, Sydney, Australia.

Steel, R. G., 1960. A Rank Sum Test for Comparing All Pairs of Treatments. Technometrics 2 (2), 197–207.

Steel, R. G., 1961. Some Rank Sum Multiple Comparisons Tests. Biometrics, 539–552.

Storey, J. D., 2003. The Positive False Discovery Rate: a Bayesian Interpretation and the q-value. The Annals of Statistics 31 (6), 2013–2035.

Studenka, B. E., Athens, M., Casos, K., Coltrin, J., McKinney, E., Symanzik, J., June 2020. The Influence of Postural Stability and Yoga Experience on Perceptions of Other's Postural Stability. In: Journal of Sport and Exercise Psychology. Vol. 42 of S1. North American Society for the Psychology of Sport and Physical Activity (NASPSPA), Virtual Conference, p. 58, https://doi.org/10.1123/jsep.2020-0172.

Symanzik, J., Li, C., Zhang, B., Studenka, B. E., McKinney, E., 2017. Eye-Tracking in Practice: A First Analysis of a Study on Human Postures. In: 2017 JSM Proceedings. American Statistical Association, Alexandria, VA, pp. 2212–2226.

Symanzik, J., McKinney, E., Studenka, B. E., Bean, B., Athens, M., Hansen, M., 2018. Eye-Tracking in Practice: Results from a Study on Human Postures. In: 2018 JSM Proceedings. American Statistical Association, Alexandria, VA, pp. 2697–2706.

Syrjala, S. E., 1996. A Statistical Test for a Difference Between the Spatial Distributions of Two Populations. Ecology 77 (1), 75–80.

Székely, G. J., Rizzo, M. L., 2004. Testing for Equal Distributions in High Dimension. InterStat 10 (11), 1249–1272.

Terpstra, T. J., 1952. The Asymptotic Normality and Consistency of Kendall's Test Against Trend, when Ties are Present in One Ranking. Indagationes Mathematicae 14 (3), 327–333.

Upton, G., Fingleton, B., Stoyan, D., 1985. Spatial Data Analysis by Example. Volume 1: Point Pattern and Quantitative Data. John Wiley & Sons Ltd., Hoboken, NJ.

von Mises, R., 1928. Wahrscheinlichkeit Statistik und Wahrheit. Springer-Verlag, Wien.

Voronoi, G., 1908. Nouvelles Applications des Paramètres Continus à la Théorie des Formes Quadratiques. Deuxième Mémoire. Recherches sur les Parallélоèdres Primitifs. Journal für die reine und angewandte Mathematik (Crelles Journal) 134, 198–287.

Wald, A., Wolfowitz, J., 1944. Statistical Tests Based on Permutations of the Observations. The Annals of Mathematical Statistics 15 (4), 358–372.

Weiss, R. S., Remington, R., Ellis, S. R., 1989. Sampling Distributions of the Entropy in Visual Scanning. Behavior Research Methods, Instruments, & Computers 21 (3), 348–352.

Wilks, S. S., 1941. Determination of Sample Sizes for Setting Tolerance Limits. The Annals of Mathematical Statistics 12 (1), 91–96.

Wilson, R. J., 1996. Introduction to Graph Theory, 4th Edition. Longman Group Ltd, Harlow, United Kingdom.

Yarbus, A. L., 2013. Eye Movements and Vision. Springer, New York, NY.

Zardari, N. H., Ahmed, K., Shirazi, S. M., Yusop, Z. B., 2015. Weighting Methods and Their Effects on Multi-Criteria Decision Making Model Outcomes in Water Resources Management. Springer, New York, NY.

Zech, G., Aslan, B., 2003. A Multivariate Two-Sample Test Based on the Concept of Minimum Energy. In: Lyons, L., Mount, R., Reitmeyer, R. (Eds.), Statistical Problems in Particle Physics, Astrophysics, and Cosmology — PHYSTAT 2003. Stanford Linear Accelerator Center, Stanford, CA, pp. 97–100.

CURRICULUM VITAE

The author's curriculum vitae (CV) is provided on the following two pages. The author can be reached via the contact information provided within the CV. Any questions or requests regarding the content of this dissertation are welcome.

Eric McKinney

ericmckinney77@gmail.com https://ericmckinney.net

Education

Ph.D. Statistics **May 2022**
Utah State University, Logan UT *GPA 4.0*

M.S. Statistics **May 2017**
Utah State University, Logan UT *GPA 3.97*

B.S. Mathematics **Apr 2014**
Weber State University, Ogden UT *GPA 3.76*
Cum Laude honors. Minor in Business Administration.

Experience

Signal Processing / Artificial Intelligence Engineer **July 2021–present**
Space Dynamics Laboratory
- Produced efficient geolocation algorithms for NASA's Atmospheric Waves Experiment sensor pixel line-of-sight data
- Analyzed meteor data from Geostationary Lightning Mapper satellites for the Planetary Defense Coordination Office

Reinforcement Learning Engineering Assistant **Dec 2020–June 2021**
Space Dynamics Laboratory
- Developed cutting-edge machine learning cybersecurity solutions for SDL's Advanced Processing and Analytics branch
- Leveraged 8 NVIDIA GPUs for parallelized model training of 70 deep neural networks used for malicious traffic detection

Mathematics Course Developer & Instructor **May 2019–Aug 2020**
School of Veterinary Medicine at USU
- Awarded a $6,167 contract to redevelop a graduate level mathematics course which streamlined average veterinary student completion time by 71% (currently maintained by Dr. Michael S. Bishop, Director of Student and Academic Affairs)

Statistical Analyst Intern **May 2018–Aug 2018**
Intermountain Healthcare Inc.
- Engineered predictive models for SelectHealth's David L. Larsen, Director of Quality Control, involving web scraping, hierarchical cluster analysis, and time series predictions of Center for Medicaid and Medicare performance data
- Developed business intelligence SQL queries for comparing internal performance metrics for top executives in Salt Lake City

Graduate Instructor **Aug 2015–Dec 2020**
Utah State University
- Instructed 165 business students as the large-lecture instructor for a challenging undergraduate business statistics course
- Awarded "Excellence in Teaching" 4 consecutive years for coaching over 487 students in 20 college level math/stat courses
- Increased student engagement by developing 5 interactive Geogebra applets hosted on `https://ericmckinney.net/`

Projects

Description	**Programming and Computer Skills**
◦ Cybersecurity Network Anomaly Detector using Machine Learning	Python, TensorFlow, Git, Docker
◦ Convolutional Neural Network of Augmented Eye-Tracking Data	Python, Scikit-learn, PyTorch, MatPlotLib
◦ 2018 Data Expo R Shiny National Weather App (see `https://ericmckinney.shinyapps.io/vis_project_2/`)	R, ggplot2, R Shiny, Leaflet, Git
◦ Monte-Carlo Simulation of Bivariate Two-Sample Test on U of U's CHPC	R, C++, Slurm, Linux, Bash, GitHub
◦ Published the `distdiffR` package (https://github.com/EricMcKinney77/distdiffR)	R, Rcpp, Git, GitHub
◦ Detecting Counterfeit Euros with Multivariate Classification Methods *Awarded Outstanding Graduate Poster Presentation at USU's SRS 2018	R, ggplot2, PowerPoint
◦ Built and maintain `https://ericmckinney.net` hosted via `Netlify.com`	Rblogdown, Rmarkdown, HUGO, Git

Leadership and Service

President, USU Data Science Group Dec 2015–May 2016
- Founded the Data Science Group as USU's ASA student chapter, which was awarded USU competitive funding

- Organized two machine learning competitions for 46 members and participated in the Hack USU 2019 hackathon

Activities Director, Ogden Institute's Activities Council Apr 2009–Dec 2009

- Led a team which planned, organized, and directed activities for over 3000 students

President, WSU Math Factor Club Apr 2013–Apr 2014

Publications

McKinney, Eric and Daniel Mortensen. Deep Anomaly Detection for Network Traffic. In *2021 55th Asilomar Conference on Signals, Systems, and Computers*, pages 1279–1283, New York, NY, 2021. IEEE.

McKinney, Eric and Jürgen Symanzik. Modifications of the syrjala test for testing spatial distribution differences between two populations. In *2019 JSM Proceedings.*, pages 2518–2530, Alexandria, VA, 2019. American Statistical Association.

McKinney, Eric. Prediction of stress increase in unbonded tendons using sparse principal component analysis. Master's thesis, Utah State University, 2017.

McKinney, Eric, Minwoo Chang, Marc Maguire, and Yan Sun. Prediction of stress increase at ultimate in unbonded tendons using sparse principal component analysis. *International Journal of Concrete Structures and Materials*, 13(1):20, Mar 2019. ISSN 2234-1315. doi: 10.1186/s40069-019-0339-y. URL `https://doi.org/10.1186/s40069-019-0339-y`.

Joanna Coltrin, **McKinney, Eric**, Breanna Studenka, and Jürgen Symanzik. Defining areas of interest for eye-tracking data: Implementing a systematic approach. In *2020 JSM Proceedings.*, pages 1144–1153, Alexandria, VA, 2020. American Statistical Association.

Breanna E. Studenka, Melanie Athens, Kristina Casos, Joanna Coltrin, **McKinney, Eric**, and Jürgen Symanzik. The Influence of Postural Stability and Yoga Experience on Perceptions of Other's Postural Stability. In *Journal of Sport and Exercise Psychology*, volume 42 of *S1*, page 58, Virtual Conference, June 2020. North American Society for the Psychology of Sport and Physical Activity (NASPSPA). `https://doi.org/10.1123/jsep.2020-0172`.

Jürgen Symanzik, **McKinney, Eric**, Joanna Coltrin, and Breanna Erin Studenka. Eye-tracking in practice: An application to human postures. Sankt-Peterburg, Russia, 2020. International Scientific and Practice Conference, ISPC.

Jürgen Symanzik, Chunyang Li, Breanna Erin Studenka, Boyu Zhang, **McKinney, Eric**, Brennan Bean, Melanie Athens, and Madison Hansen. Eye-tracking in practice: The eyetrackr R package and its use in a study on human postures. Taipei, Taiwan, 2018a. Institute of Statistical Science, Academia Sinica, 2018 Statistics Week.

Jürgen Symanzik, **McKinney, Eric**, Breanna Erin Studenka, Brennan Bean, Melanie Athens, and Madison Hansen. Eye-tracking in practice: Results from a study on human postures. In *2018 JSM Proceedings.*, pages 2697–2706, Alexandria, VA, 2018b. American Statistical Association.

Jürgen Symanzik, Chunyang Li, Boyu Zhang, Breanna Erin Studenka, and **McKinney, Eric**. Eye-tracking in practice: A first analysis of a study on human postures. In *2017 JSM Proceedings.*, pages 2212–2226, Alexandria, VA, 2017. American Statistical Association.