Extensions to the Syrjala Test with Eye-Tracking Analysis Applications

Eric McKinney * Jürgen Symanzik †

Abstract
A series of generalized versions of the Syrjala (1996) two-sample test of distributional equality are proposed. The new tests makes use of both rotations and toroidal shifts of the data. Additionally, the tests exhibit stability across a variety of explored test statistics. While computationally more complex, simulations establish the test which employs both rotations and toroidal shifts as a more powerful and conservative choice over its predecessor. This test is then applied to eye-tracking data that originates from the Utah State University Posture Study.

Key Words: Continuous Bivariate Data; Two-Sample; Toroidal Shift; Permutation Test; Eye-Tracking Data

1. Introduction

Often the question arises as to whether or not two samples were drawn from the same distribution. Well known two-sample tests of distributional equality for univariate samples include the two-sample Kolmogorov-Smirnov test (Kolmogorov, 1933), the Cramér-von Mises test (Cramér, 1928; von Mises, 1928) generalized to the two-sample case by Anderson (1962), and the two-sample version to the Anderson-Darling statistic (Darling, 1957; Anderson and Darling, 1952, 1954). The two-sample Kolmogorov-Smirnov, Cramér-von Mises, and Anderson-Darling tests are classified as permutation tests (Berry et al., 2011). However, there also exists another classification of two-sample tests, called rank tests, including the Mann-Whitney U test (Mann and Whitney, 1947), and the Wald and Wolfowitz (1944) runs test for measuring whether two binary values which occur in a sequence are drawn independently from the same distribution.

Many of these univariate two-sample tests have also been generalized to the bivariate setting and beyond. For example, Friedman and Rafsky (1979) generalized the two-sample Kolmogorov-Smirnov, and Wald-Wolfowitz runs tests to the multivariate case by use of a minimal spanning tree on the pooled sample.

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. 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).

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. The proposed tests

*Department of Mathematics and Statistics, Utah State University, Logan, UT 84322–3900, USA. E-mail: ericmckinney77@gmail.com

†Department of Mathematics and Statistics, Utah State University, Logan, UT 84322–3900, USA. E-mail: symanzik@math.usu.edu

presented in this paper (discussed further in Section 2) involves several modifications to the Syrjala (1996) test, which is a bivariate extension to the Cramer-von Mises test. The new tests are referred to herein as the modified Syrjala tests, several of which are shown to be more powerful and more appropriately conservative by use of Monte-Carlo simulations than the original Syrjala test.

The rest of this paper will proceed as follows: Section 2 details the proposed modifications to the Syrjala test. Section 3 discusses the structure and results of a comparative simulation study. The test which employs both rotational and toroidal shifts is then applied to eye-tracking data from the Utah State University (USU) Posture Study (Symanzik et al., 2017, 2018; Coltrin et al., 2020) in Section 4. Sections 5 and 6 contain conclusions and future work, respectively. Additional simulation results are also provided in the appendix.

2. The Modified Syrjala Test

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 2.1), and (4) implementing the use of toroidal shifts of the data within the test (discussed in Section 2.2). A combination of both the rotational and toroidal shift modifications is detailed in Section 2.3.

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 (where n_1 and n_2 are the respective sample sizes) 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 some coordinate pair (x, y)

In contrast to the Syrjala test, $\Gamma_1^*(x, y)$ and $\Gamma_2^*(x, y)$ in this test 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 test statistics are proposed as follows:

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

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

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

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

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

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

These statistics explore the use of squared (ξ^{DWS}, ξ^{UWS}, and ξ^{RWS}) vs. absolute (ξ^{DWA}, ξ^{UWA}, and ξ^{RWA}) differences between the ECDFs along with three different types of weightings, namely, double (ξ^{DWS} and ξ^{DWA}), unweighted (ξ^{UWS} and ξ^{UWA}), and reversed (ξ^{RWS} and ξ^{RWA}) weightings. The abbreviations DW, UW, and RW refer to the different types of weightings in the ξ statistics, i.e., double, unweighted, and reverse weighted, respectfully. The S and A are abbreviations for squared or absolute differences, respectively, computed between the ECDFs from the two samples. 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). 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. The unweighted weightings omit these scaling ratios, and reversed weightings apply the scaling ratios to the opposite sums.

The six statistics are chosen to further explore the interactive and individual effects of squared vs. absolute ECDF differences along with the weightings of the respective differences. Similarities can be seen between Equations 1–6 (especially Equation 2) and the univariate two-sample Cramer-von Mises test statistic (Cramér, 1928) 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.

2.1 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. Consequently, the original Syrjala test was rotated four times in an attempt to remove a dependency of the test on data which lies closer to this origin (Syrjala, 1996). 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

$$\Psi^R = \frac{1}{R} \sum_{r=1}^{R} \xi_r^*, \tag{7}$$

where ξ_r^* is one of the six statistics defined by Equations 1–6 which redefines Γ_1^* as $\Gamma_{1,r}^*$ and Γ_2^* as $\Gamma_{2,r}^*$ for each of the r^{th} rotations, and R is a discrete number of rotations within 360°. Hence, our modified test statistic (Ψ^R) is the average of the respective statistic computed across the rotations of the data.

2.2 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 by McKinney and Symanzik (2019). To overcome this, an additional modification is employed which uses toroidal shifts.

The toroidal shift is a well established technique in the spatial statistics literature (Lotwick and Silverman, 1982; Diggle and Milne, 1983; Upton et al., 1985; Berman, 1986; Díaz et al., 2008; Dixon, 2014; Moreno-Fernández et al., 2020). This modification also addresses the ECDF origin issue (see Section 2.1 for more details). 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.

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 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 (Upton et al., 1985).

In order to ensure that every ECDF of the shifted data is equally likely, we apply the toroidal shift as follows: A random data value is first 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 1 shows an example of two subject's data which undergo a random toroidal shift.

These new toroidal shifted data provide 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 1–6 which redefines Γ_1^* as $\Gamma_{1,t}^*$ and Γ_2^* as $\Gamma_{2,t}^*$ for each of the t toroidal shifts. 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 7, except that the individual computations are weighted according to the number of toroidal shifts R_T, as seen in the following

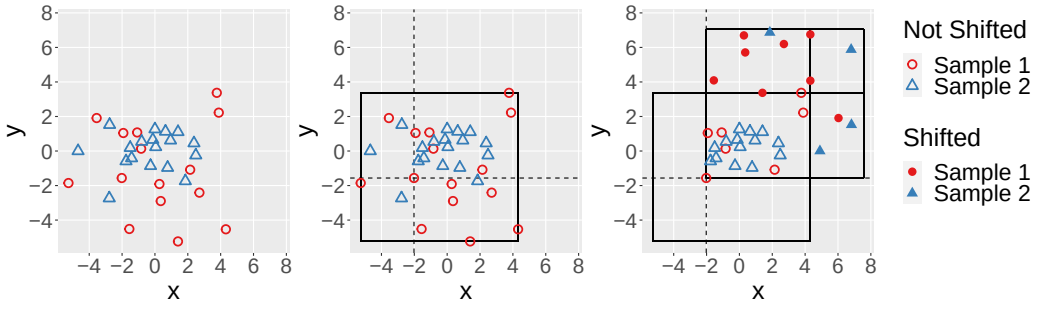


Figure 1: A visualization of two-sample data before (left) and after (right) a toroidal shift transformation. The center plot shows a bounding rectangle around combined samples along with the randomly selected data value which serves as the origin of the toroidal shift. The data values unaffected by the toroidal shift are indicated by hollow shapes, whereas those affected are indicated by filled-in shapes.

equation:

$$\Psi^T = \frac{1}{R_T} \sum_{t=1}^{R_T} \xi_t^*. \tag{8}$$

2.3 Combining Modifications

In an effort to remove both the dependency of the Syrjala test on the ECDF's origin (see Section 2.1 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 2.1) and toroidal shift (see Section 2.2) 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 7 and 8 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 1–6 which redefines Γ_1^* as $\Gamma_{1,r,t}^*$ and Γ_2^* as $\Gamma_{2,r,t}^*$ for each of the r rotations and t toroidal shifts.

2.4 Permutation Test Computations

Let Ψ^* be one of the previously discussed test statistics Ψ^R, Ψ^T, or Ψ^{RT} (see Sections 2.1–2.3). 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 $N \ll N_{\max}$ 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.

Figure 2 outlines the process in which both rotational and toroidal shift modifications are integrated into the Modified Syrjala test. 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 2.3 and 2.4, respectively.

3. Simulations

The rotational modified Syrjala test has already been shown to be both more powerful and more appropriately conservative as compared to the original Syrjala test (McKinney and Symanzik, 2019). However, additional simulations are discussed here which investigate both the power and false positive rate of not only the rotational modification, but also the toroidal shift, and combined modifications (see Sections 2.1–2.3).

3.1 Simulation Setup

To assess the tests when the null hypothesis is true, two realizations of independent, uniformly distributed, or completely spatially random (CSR), data were simulated on $[0, 1] \times [0, 1]$ square regions. To assess the tests when the null hypothesis is false, four separate comparisons were made, 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.

$$\begin{aligned}
f_{Center}(x, y) &= a_1 \cdot \exp\Big\{ -20 \cdot \big[(x - 0.5)^2 + (y - 0.5)^2\big] \Big\} \\
f_{Repel}(x, y) &= a_2 \cdot \Big(1 - \exp\Big\{ -80 \cdot \big[(x - 0.5)^4 \\
&\qquad + (y - 0.5)^4\big] \Big\}\Big) \\
f_{Corner}(x, y) &= a_3 \cdot \exp\Big\{ -5 \cdot \big[(x - 1)^2 + (y - 1)^2\big] \Big\} \\
f_{Right}(x, y) &= a_4 \cdot \exp\Big\{ -5 \cdot (x - 1)^2 \Big\}
\end{aligned}$$

Let μ be the average number of points within the unit square for the heterogeneous Poisson process. For reproducibility, Table 1 shows the values for the height parameters that achieve a specified intensity μ for each departure from CSR. The coefficients were chosen to ensure a sufficient departure from CSR. For each of the five comparisons (CSR compared with CSR, Center, Repel, Corner, Right), CSR realizations of 50, 100, 250, and 500 points was compared to the same four sample sizes.

Additionally, in order to ensure reproducibility of the simulation results, the simulation data is generated up front with predefined random number seeds.

Furthermore, to reduce the overall variability of the statistics on the simulated data, the method of common random numbers (CRNs) is employed (Glasserman, 2013; Botev and Ridder, 2017). CRNs (also called correlated sampling, matched

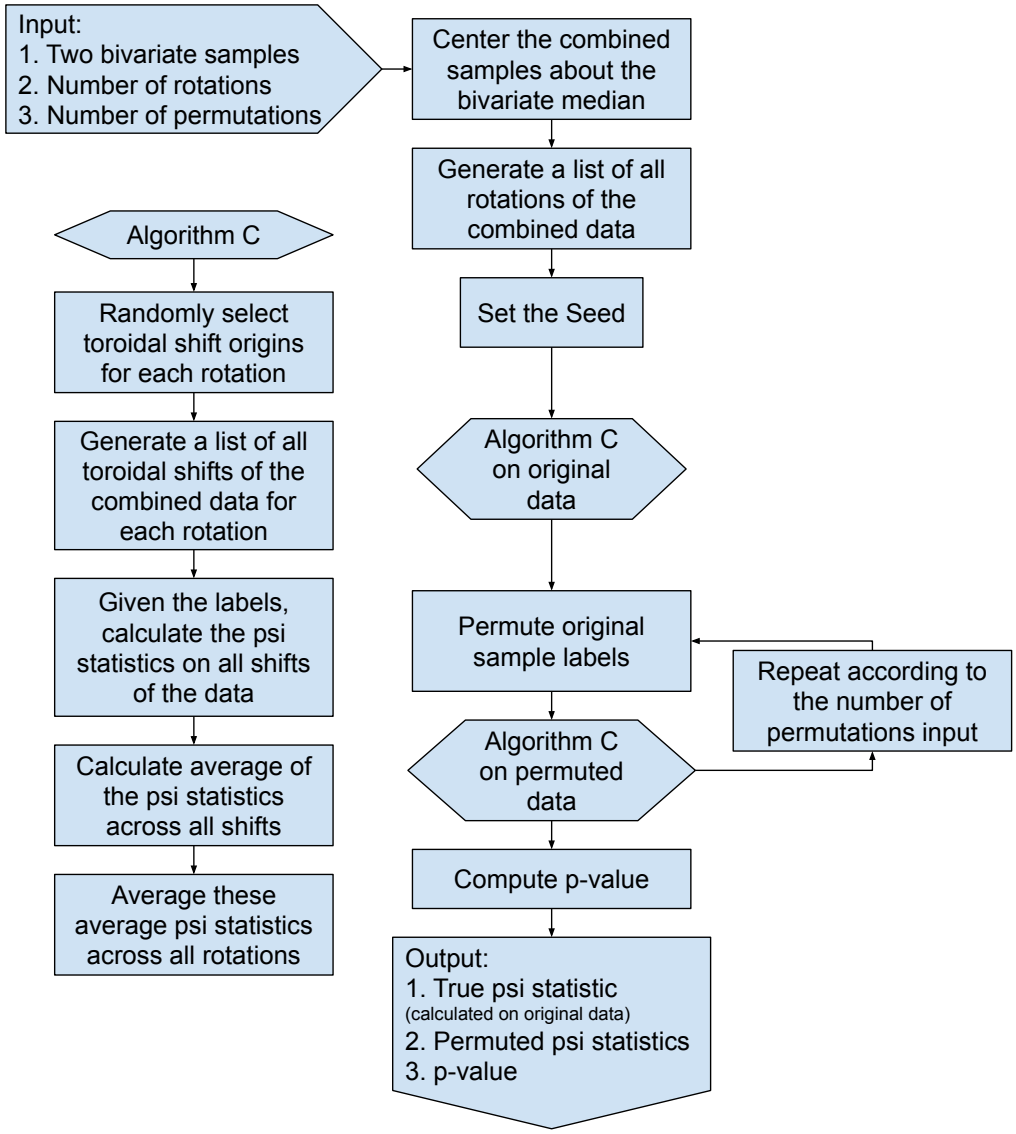


Figure 2: A flowchart which displays the process in which a combination of both the rotational and toroidal shift modifications are integrated into the Modified Syrjala test. The psi statistics computed on the original data and permuted data referred to in the figure are the Ψ^{RT} and Ψ_l^{RT} discussed in Sections 2.3 and 2.4, respectively.

Table 1: 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

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 simulation, which reduces the overall variation in the simulation statistics.

This work builds upon that of McKinney and Symanzik (2019) in the five ways: (1) the varying sample sizes when comparing two samples now includes all possible combinations of 50, 100, 250, and 500 for the first sample and 50, 100, 250, and 500 for the second sample (in McKinney and Symanzik (2019) only sample sizes of 500 in the first sample were compared to sample sizes of 50, 100, 250, and 500 for the second sample), (2) the performance of the rotational extension of the Syrjala test is observed at 4, 5, 6, 8, 10, 36, and 45 rotations within 360° (i.e., 4 90° rotations, 5 72° rotations, etc.), whereas McKinney and Symanzik (2019) only explored 4, 6, 8, 10, and 36 rotations within 360°, (3) five additional statistics are explored whereas McKinney and Symanzik (2019) only employed ξ^{DWS}, (4) the toroidal shift and combined rotational and toroidal shift extensions are investigated, and (5) a computationally feasible instance of the modified Syrjala tests (employing both rotational and toroidal shifts) is applied to the USU Posture Study data in Section 4.

3.2 Simulation Results

Due to the necessity of identical sampling locations for the Syrjala test, 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).

3.2.1 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.

3.2.2 Syrjala Test Simulation Results

Figure 3 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 3.2.1) on the Syrjala test. The horizontal axis displays which type of binning, either regular (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 3.2.1. 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 3), 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 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

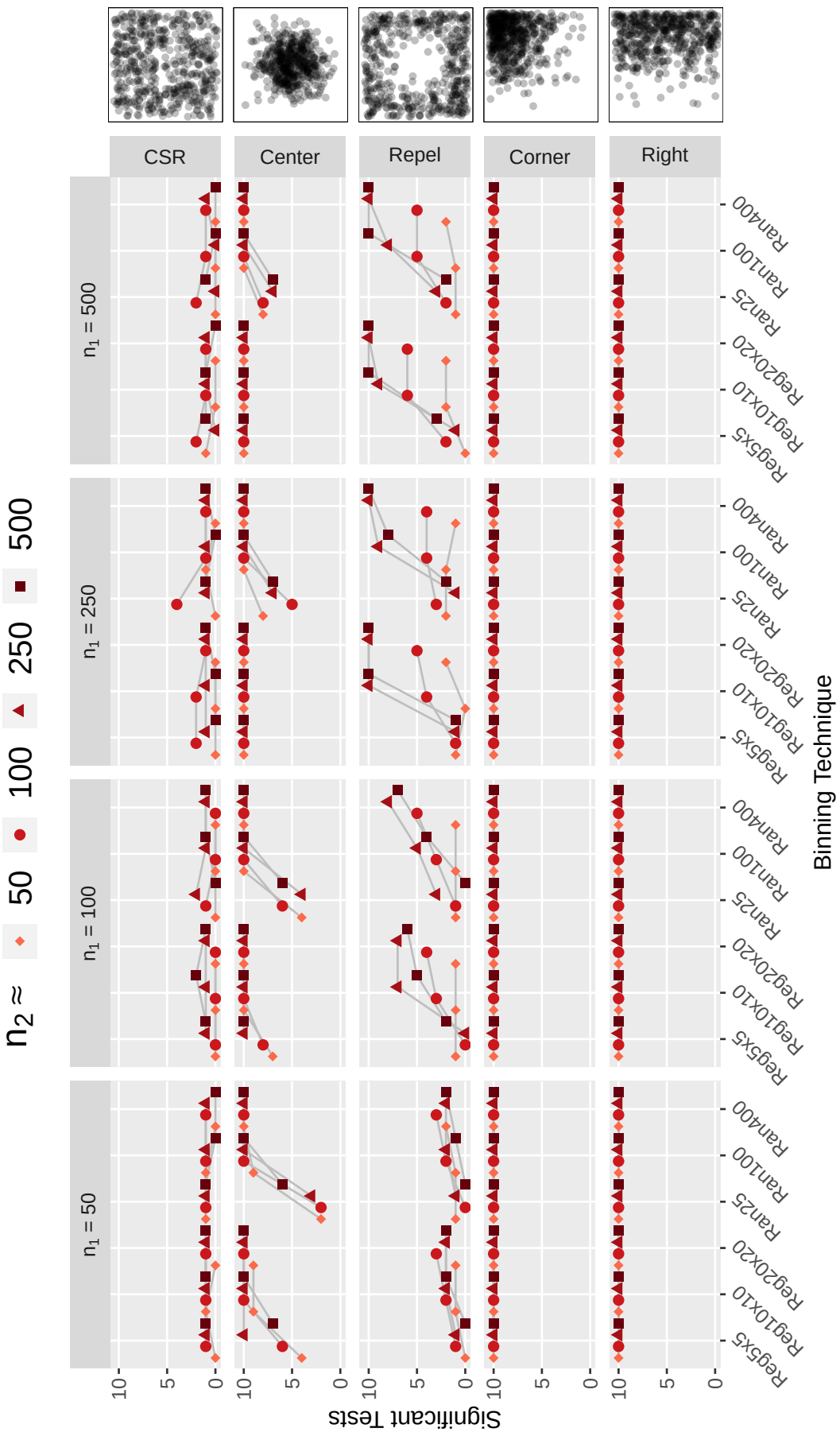


Figure 3: A grid of line graphs showing the results of a simulation comparing the effect of two types of data binning (lower horizontal axis), abbreviated as Reg or Ran, on the Syrjala test. The grid column indicates the CSR sample size (n_1), and the grid row indicates the distribution of the second 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 3.2.1

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 one and two more significant test results as the binning granularity increases for random and regular binning, respectively. However, when $n_1 = 500$ and $n_2 = 500$ the Syrjala test is able to detect eight and seven more significant test results as the binning granularity increases for random and regular 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 3 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).

3.2.3 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 Section 2), 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 3.2.6.

3.2.4 Rotational Modification Simulation Results

Figure 4 shows the results of the simulation study for the rotational modification when using the RWS statistic. The remaining five proposed versions of the statistic ξ (see Section 2 for more details) are also explored. However, since Figures 4 and 12–16 show almost the same behavior aside from some chance variation, the latter

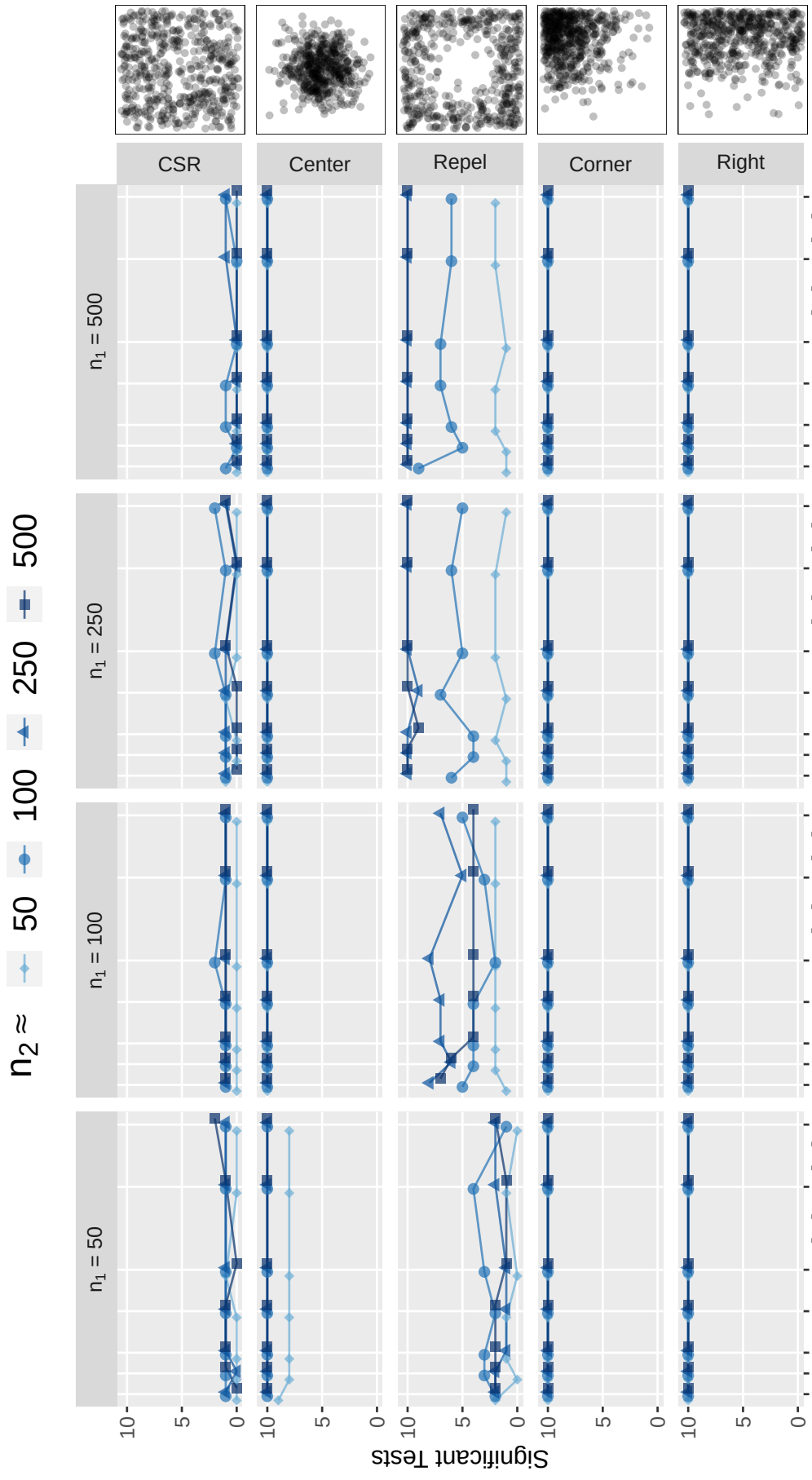


Figure 4: A grid of line graphs showing the results of a simulation comparing multiple rotations of the modified Syrjala test using reverse weightings of the squared differences in the ECDFs (RWS). The grid column name indicates the CSR sample size (n_1), and the grid row indicates the shape of the second 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.

figures (Figures 12–16) for the DWS, UWS, DWA, UWA, and RWA simulations (respectively) are provided in the Appendix. Recall from Section 2 that the abbreviations DW, UW, and RW refer to the different types of weightings in the ξ statistics, i.e., double weighted, unweighted, and reversed weighted, respectfully. The S and A are abbreviations for squared or absolute differences, respectively, computed between the ECDFs from the two samples. Figures 4 and 12–16 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 3.3), the choice in proposed versions of the statistic ξ makes little difference in the overall performance of the rotational modification test. Additionally, Figure 4 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 3.2.6 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 the figures). 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 3.3.

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 4 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 with Figure 3 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 3.2.2, 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 test also places less emphasis on differences located near the center of the bounding region similar to the Syrjala test (McAdam et al., 2012). However, the rotational modification Syrjala test overcomes some of these issues (see Section 3.2.2). 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 3.2.5 and 3.2.6 are discussed in Section 3.3. In comparison, the far right column ($n_1 = 500$) of Figure 12 match the results found by McKinney and Symanzik (2019).

3.2.5 Toroidal Shift Modification Simulation Results

Similar to the previous section, Figure 5 shows the results of the simulation study for the rotational modification when using the RWS statistic. The remaining five proposed versions of the statistic ξ (see Section 2 for more details) are also explored. However, since Figures 5 and 17–21 show almost the same behavior aside from some chance variation, the latter figures (Figures 17–21) for the DWS, UWS, DWA, UWA, and RWA simulations (respectively) are provided in the Appendix. The layout of these figures is identical to Figure 4 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 3.2.4, while the squared statistics perform marginally better than the absolute value statistics in Figure 5 (this is shown more explicitly in Section 3.3), the choice in proposed versions of the statistic ξ makes little difference in the overall performance of the toroidal shift modification test. Additionally, Figure 5 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 3.2.6 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 3.3.

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 5 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. In comparison, the rotational modified Syrjala test correctly computed significance for all of the Right distribution cases (as seen in the bottom row of graphs of Figure 4).

Similar to the rotational test results (in Section 3.2.4), 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 4, Figure 5 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

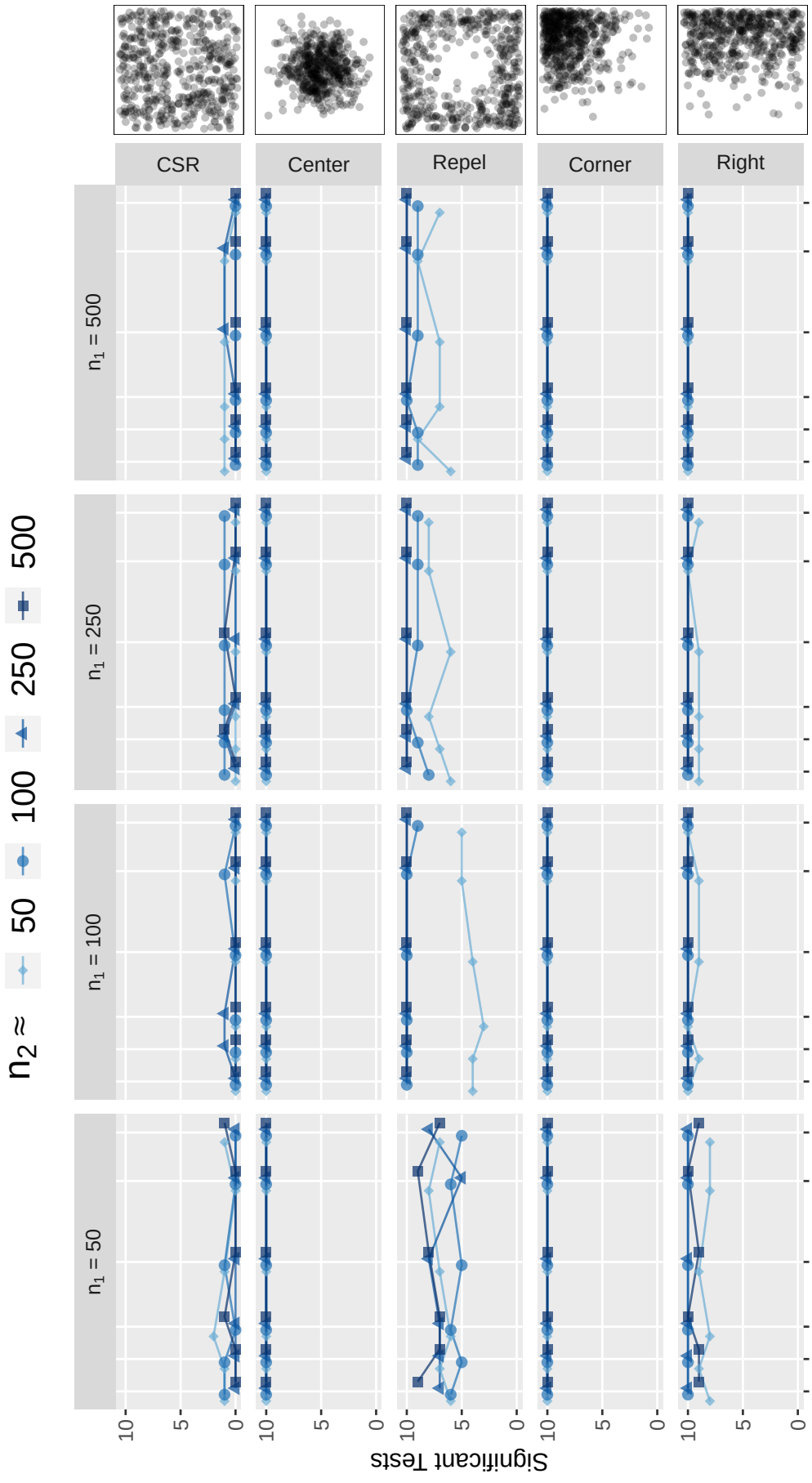


Figure 5: A grid of line graphs showing the results of a simulation comparing multiple randomly selected proportions of points for the origins of the toroidal shifts of the modified Syrjala test using reverse weightings of the squared differences in the ECDFs (RWS). The grid column name indicates the 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.

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 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 3.3, 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 3.2.6.

3.2.6 Simulation Results for the Modified Syrjala Test which Combines both Rotational and Toroidal Shift Modifications

While little difference is observed in the versions of the ξ statistics across both the rotational (Figures 4 and 12–16) and toroidal shift (Figures 5 and 17–21) modifications, squared differences in the statistics perform marginally better than the absolute differences. Hence, while an argument could be made for the absolute differences, the squared difference is only considered in the simulations of the test involving both rotational and toroidal shift modifications. Furthermore, there is little indication of superiority in the choice of weights within the ξ statistic. Consequently, for the sake of counteracting a difference in sample size between the two samples, ξ^{RWS} is used for combined modification simulations. See Section 2 for more details about the different ξ statistics.

Hence, Figures 6–8 show the results of the simulation study for the combination of both rotational and toroidal shift modifications using only the ξ^{RWS} statistic. Similar to Figures 4, 5, and 12–21, 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. Figures 6–8 use the proportions of 0.1, 0.2, and 0.3, respectively. The 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 6–8 similar to those seen in Section 3.2.5 suggesting that running these additional simulations may be unnecessary.

Overall, Figures 6–8 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 3.2.5. 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). 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. This is shown more explicitly and discussed further in Section 3.3.

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 6–8 exhibit any difficulty in correctly identifying all of the significant differences for the

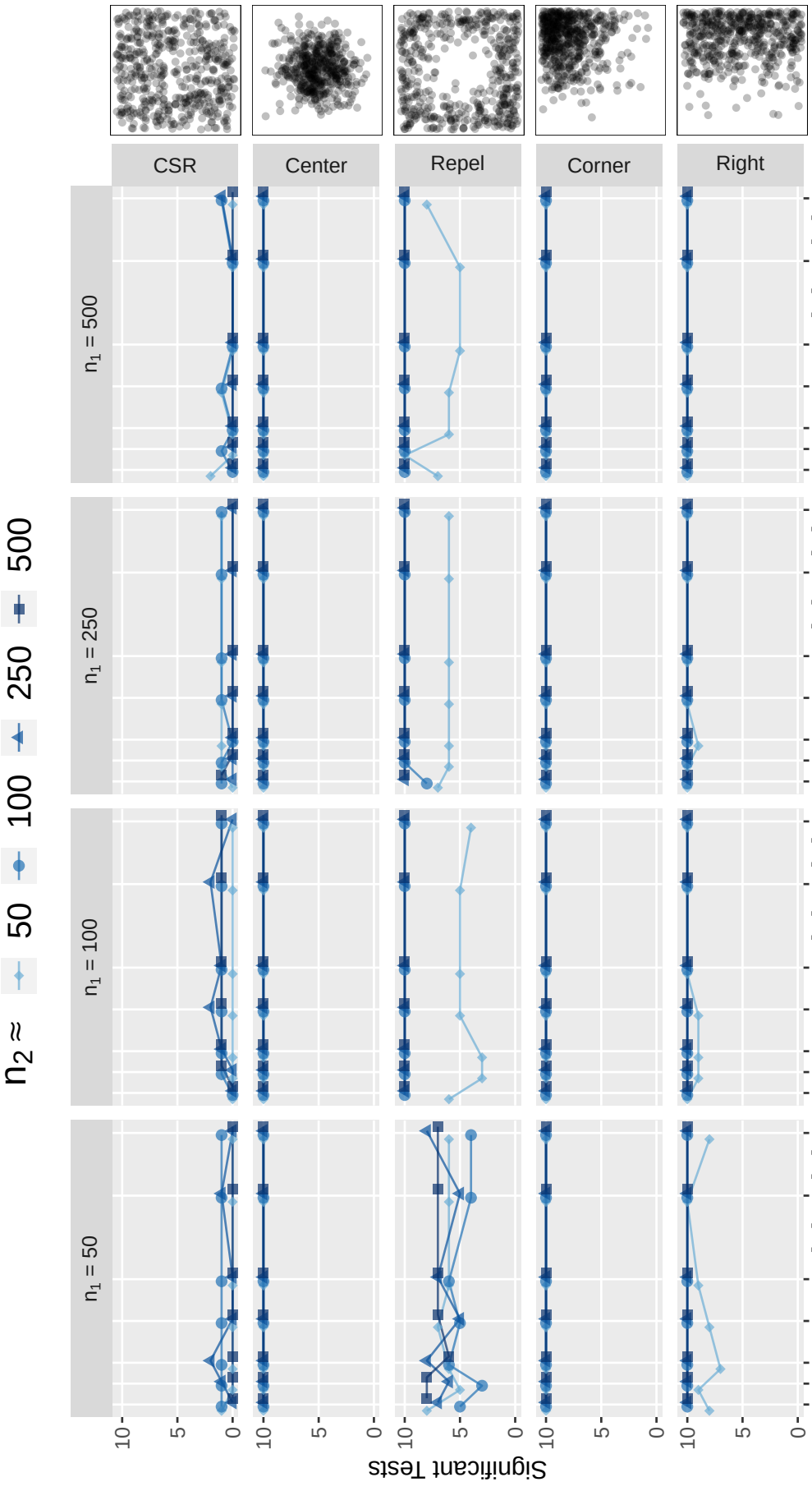


Figure 6: A grid of line graphs showing the results of a simulation comparing multiple rotations of the modified Syrjala test using 0.1 proportion of points as origins of toroidal shifts, and reverse weightings of the squared differences in the ECDFs (RWS). The grid column name indicates the CSR sample 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.

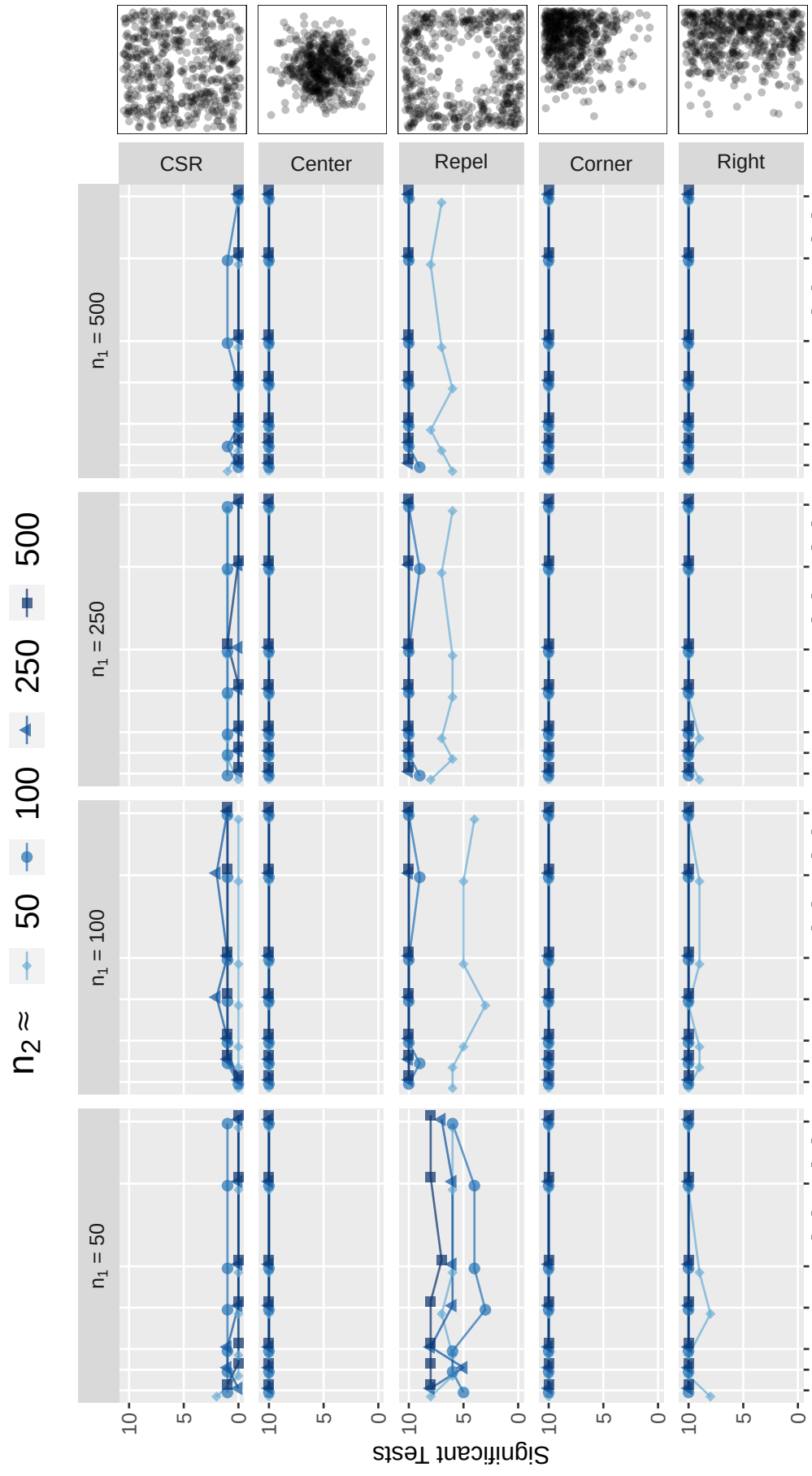


Figure 7: A grid of line graphs showing the results of a simulation comparing multiple rotations of the modified Syrjala test using 0.2 proportion of points as origins of toroidal shifts, and reverse weightings of the squared differences in the ECDFs (RWS). The grid column name indicates the CSR sample 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.

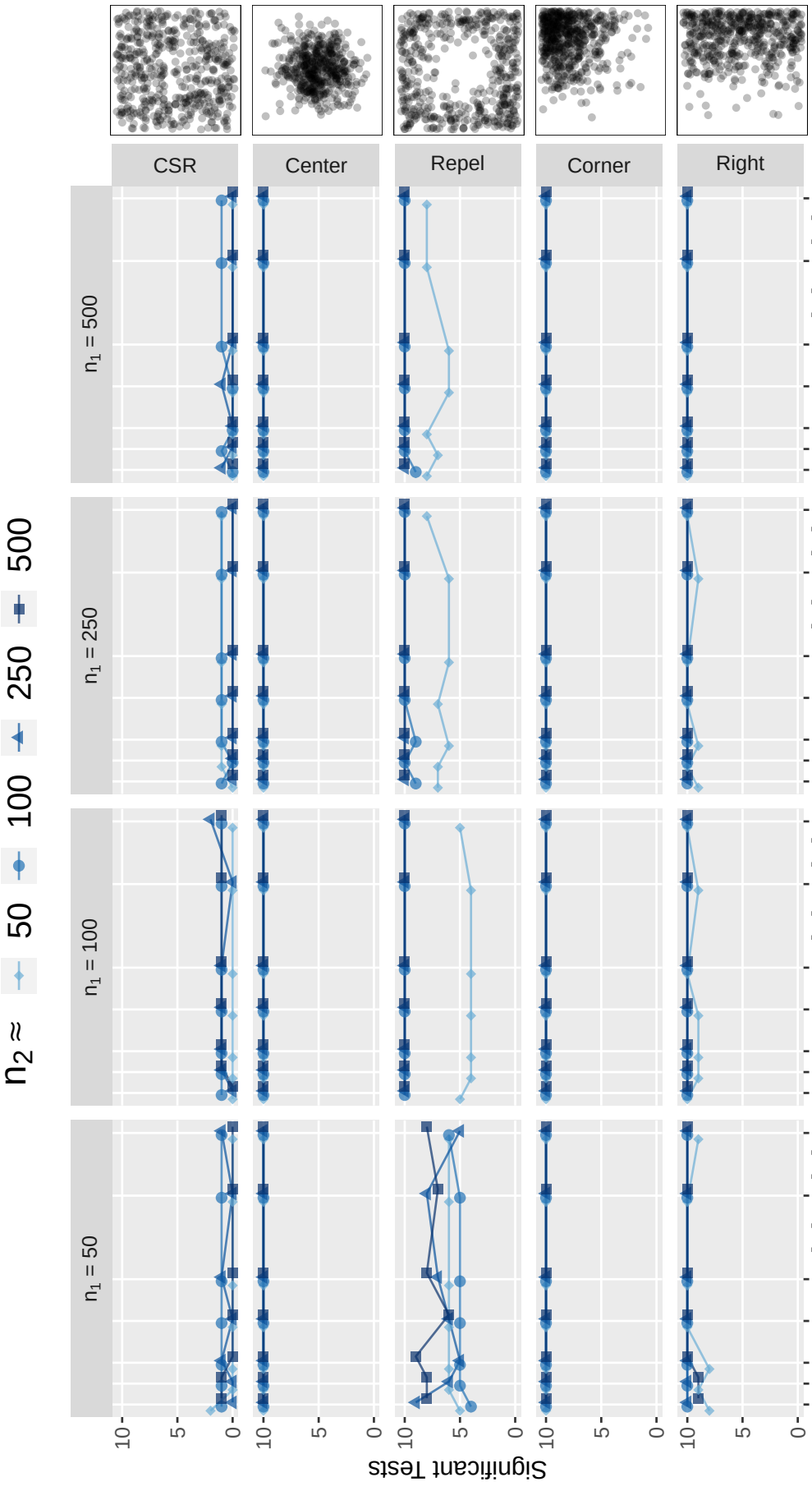


Figure 8: A grid of line graphs showing the results of a simulation comparing multiple rotations of the modified Syrjala test using 0.3 proportion of points as origins of toroidal shifts, and reverse weightings of the squared differences in the ECDFs (RWS). The grid column name indicates the CSR sample 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.

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 3.2.5). 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 3.2.5), 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 3.2.5), 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).

3.3 Summary of Results

In this section, the power (see Figure 9) and false positive rate (see Figure 10) are computed and graphed for the original Syrjala test across the different binning techniques as well as the rotational, toroidal shift, and combined modified Syrjala tests across the six different ξ statistics. The false positive rate is computed as the ratio of significant tests out of all tests when the null hypothesis is true. The power of a test is computed as the ratio of significant tests out of all tests when the null hypothesis is false. In these figures, the use of the weightings DW, UW, and RW in connection with the squared (Sqrd) versus absolute (AbsV) differences correspond to the six statistics ξ^{DWS}, ξ^{UWS}, ξ^{RWS}, ξ^{DWA}, ξ^{UWA}, and ξ^{RWA}.

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 from the same CSR distribution (first row of graphs in Figures 3–8 and 12–21). For example, the Syrjala test (see Figure 3) 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 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 10, 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 to reject the null when it is actually true). In Figure 9, 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.

Overall, while Figures 9 and 10 show a considerable difference between the toroidal shift and rotational modifications, little difference is seen across the different weightings (DW, UW, and RW) or squared (Sqrd) versus absolute (AbsV)

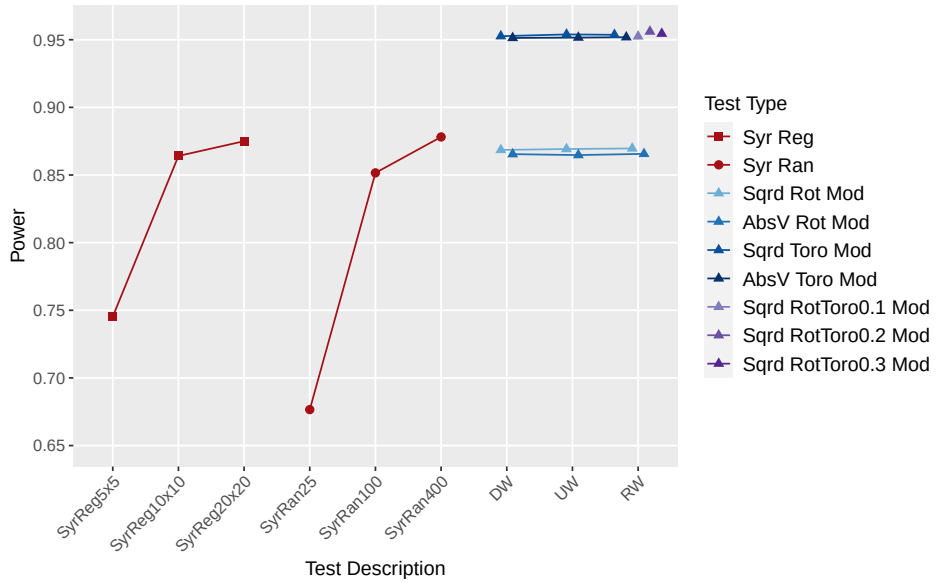


Figure 9: A comparison of the power achieved by the tests discussed in Sections 3.

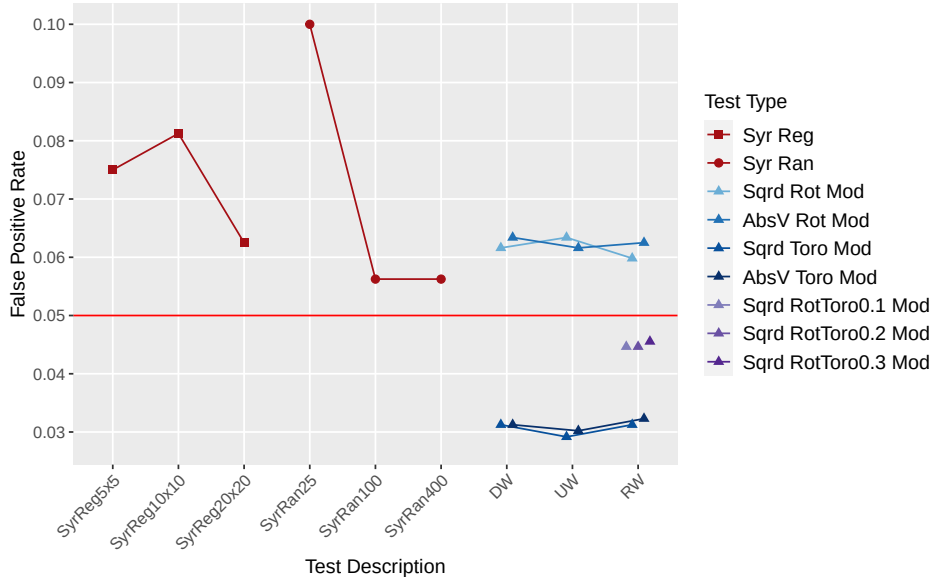


Figure 10: A comparison of the false positive rate achieved by the tests discussed in Sections 3. The horizontal red line at 0.05 indicates the significance level of the tests.

differences in the ECDF values. Specifically, the power of the tests which employ toroidal shift modifications is about 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 about 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.

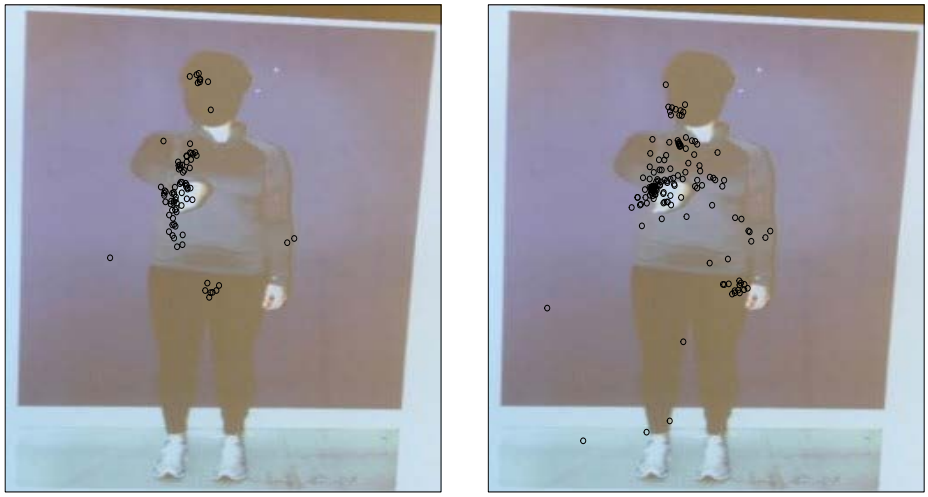
While the horizontal axes in Figures 9 and 10 describes either the binning technique for the Syrjala tests or the modified Syrjala test statistic weighting, the legend also describes which ECDF difference was used in addition to the modification employed for the modified Syrjala tests. For example, the combined rotational and toroidal shift modification tests which used the test statistic ξ^{RWS} while exploring all previously used rotations, and 0.1, 0.2, and 0.3 as the proportion of toroidal shifts are labeled in the figure legends as Sqrd RotToro0.1 Mod, Sqrd RotToro0.2 Mod, and Sqrd RotToro0.3 Mod, respectively.

Figure 9 shows the superiority of the toroidal shift and combined modifications over the original Syrjala test and rotational modifications. While the toroidal shift modification outperforms the rotational modification, the combination of both modifications did not achieve a considerably higher power than the toroidal shift modification alone. However, Figure 10 shows that the combined test provides a false positive rate closer to the significance level of 0.05. 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. Additionally, since the false positive rate is below the significance level this implies that the test is more appropriately conservative as compared to the toroidal shift modification alone. Due to the trade-off between power and false positive rates in tests, this suggests that while the toroidal shift modification is more conservative as compared to the combined modification it may not be as powerful in the face of all possible departures from the null as the combined modifications test. Thus, the combined rotational and toroidal shift modified Syrjala test appears to be the superior choice.

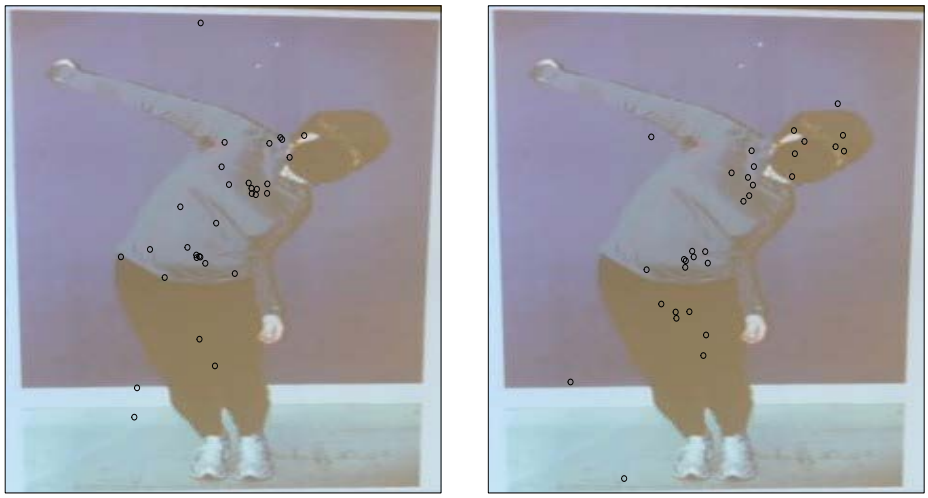
4. Application

In this section, we apply the combined rotational and toroidal shift modified Syrjala test to eye-tracking data taken from two pairs of participants from the USU Posture Study (Symanzik et al., 2017, 2018; Coltrin et al., 2020). The test employs eight rotations, 0.1 proportion of points for toroidal shifts, and the ξ^{RWS} statistic. Two pairs of gaze scatterplots are displayed in Figure 11. While applying tests which employed 99 permutations of the data and a significance level of 0.05, we obtain significant results for the postures shown in (a) and (b) (p-value = 0.01), and non-significant results for the postures shown in (c) and (d) (p-value = 0.22) in Figure 11. The sample sizes for (a)–(d) are 84, 144, 32, and 32, respectively.

The significant test result for the comparison of gaze points in (a) and (b) (in Figure 11) is confirmed by the visual difference in gaze point distributions for the postures in (a) and (b). In (a), the subject assessed the postural stability of the actor by focusing more on the actor's right torso, and between the upper thighs, whereas in (b) the subject focused on the upper right torso, and left hip / wrist. While both subjects spent time on the right wrist, the differences between the overall gaze point distributions was different enough to reject the null hypothesis and conclude



(a) (b)



(c) (d)

Figure 11: Comparison of the gaze scatterplots for two pairs of subjects. Images (a) and (b) exhibit different viewing patterns for one of the postures, while images (c) and (d) exhibit similar viewing patterns for a different posture.

significant differences between the subject's respective gaze point distributions for this posture. Furthermore, while there are some minor differences between the postures shown in (c) and (d), the non-significant test result leads to the conclusion that there is not enough evidence to suggest that these two subjects have different gaze point distributions for this posture.

5. Conclusion

The proposed generalized version of the Syrjala (1996) test overcomes many of the Syrjala test's limitations, including much of its dependency on the origin of the bivariate ECDF, and the necessity of identical sampling locations. Additionally, the combined rotational and toroidal shift modified Syrjala test, while more computationally complex, has been shown to be a more powerful and more appropriately conservative choice. It has also been shown to be useful in applications of two-sample spatial data such as in an eye-tracking data analysis.

6. Future Work

Additional analyses are being conducted on the performance of the modified Syrjala tests in comparison to several other alternative multivariate two-sample tests of distributional equality as summarized in Section 1. A series of other simulations are also being conducted which demonstrate the performance of the modified Syrjala tests on generated data which is patterned more closely to that found in eye-tracking analyses. Furthermore, an `R` package called `distdiffR` is currently being developed to make the current research more easily reproducible, aid in the extension of the methodology, and facilitate in the distribution of the underlying `R` code for further applications.

References

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. and 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. and Darling, D. A. (1954). A Test of Goodness of Fit. *Journal of the American Statistical Association*, 49(268):765–769.

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. J., Johnston, J. E., and Mielke Jr, P. W. (2011). Permutation Methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(6):527–542.

Botev, Z. and Ridder, A. (2017). Variance Reduction. *Wiley StatsRef: Statistics Reference Online*, pages 1–6. `https://doi.org/10.1002/9781118445112.stat07975`.

Chetverikov, A., Kuvaldina, M., MacInnes, W. J., Jóhannesson, Ó. I., and Kristjánsson, Á. (2018). Implicit Processing During Change Blindness Revealed with Mouse-Contingent and Gaze-Contingent Displays. *Attention, Perception, & Psychophysics*, 80(4):844–859.

Coltrin, J., McKinney, E., Studenka, B., and Symanzik, J. (2020). Defining areas of interest for eye-tracking data: Implementing a systematic approach. In *2020 JSM Proceedings.*, pages 1144–1153, Alexandria, VA. American Statistical Association.

Cramér, H. (1928). On the Composition of Elementary Errors: First Paper: Mathematical Deductions. *Scandinavian Actuarial Journal*, 1928(1):13–74.

Darling, D. A. (1957). The Kolmogorov-Smirnov, Cramer-von Mises Tests. *The Annals of Mathematical Statistics*, 28(4):823–838.

Díaz, E., Sebastian, R., Ayala, G., Díaz, M. E., Zoncu, R., Toomre, D., and 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. and 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*, pages 1–12. https://doi.org/10.1002/9781118445112.stat07751.

Friedman, J. H. and 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., and 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.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13(25):723–773.

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.

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 Distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 1933(4):83–91. (In Italian).

Lotwick, H. and 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. and 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.

McAdam, B. J., Grabowski, T. B., and Marteinsdóttir, G. (2012). Testing for Differences in Spatial Distributions from Individual Based Data. *Fisheries Research*, 127(1):148–153.

McKinney, E. and Symanzik, J. (2019). Modifications of the Syrjala Test for Testing Spatial Distribution Differences Between Two Populations. In *2019 JSM Proceedings.*, pages 2518–2530, Alexandria, VA. American Statistical Association.

Mondal, P. K., Biswas, M., and 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., and Montes, F. (2020). Strategies for Modeling Regeneration Density in Relation to Distance from Adult Trees. *Forests*, 11(1):120.

Schilling, M. F. (1986). Multivariate Two-Sample Tests Based on Nearest Neighbors. *Journal of the American Statistical Association*, 81(395):799–806.

Symanzik, J., Li, C., Zhang, B., Studenka, B. E., and McKinney, E. (2017). Eye-Tracking in Practice: A First Analysis of a Study on Human Postures. In *2017 JSM Proceedings.*, pages 2212–2226, Alexandria, VA. American Statistical Association.

Symanzik, J., McKinney, E., Studenka, B. E., Bean, B., Athens, M., and Hansen, M. (2018). Eye-Tracking in Practice: Results from a Study on Human Postures. In *2018 JSM Proceedings.*, pages 2697–2706, Alexandria, VA. American Statistical Association.

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. and Rizzo, M. L. (2004). Testing for Equal Distributions in High Dimension. *InterStat*, 10(11):1249–1272.

Upton, G., Fingleton, B., and 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.

Wald, A. and Wolfowitz, J. (1944). Statistical Tests Based on Permutations of the

Observations. *The Annals of Mathematical Statistics*, 15(4):358–372.

Zech, G. and Aslan, B. (2003). A Multivariate Two-Sample Test Based on the Concept of Minimum Energy. In Lyons, L., Mount, R., and Reitmeyer, R., editors, *Statistical Problems in Particle Physics, Astrophysics, and Cosmology — PHYSTAT 2003*, pages 97–100, Stanford, CA. Stanford Linear Accelerator Center.

Appendix

Since Figures 4 and 12–16 show almost the same rotational modified test behavior aside from some chance variation, the latter figures (Figures 12–16) for the DWS, UWS, DWA, UWA, and RWA simulations (respectively) are provided in this appendix. Similarly, since Figures 17–21 show almost the same toroidal shift modified test behavior aside from some chance variation as Figure 5, they are also provided in this appendix. Definitions for the test statistic abbreviations in these plots can be found in Section 2. Additionally, explanations of the figure graph features for Figures 12–16 and 17–21 can be found in Sections 3.2.3 and 3.2.5, respectively.

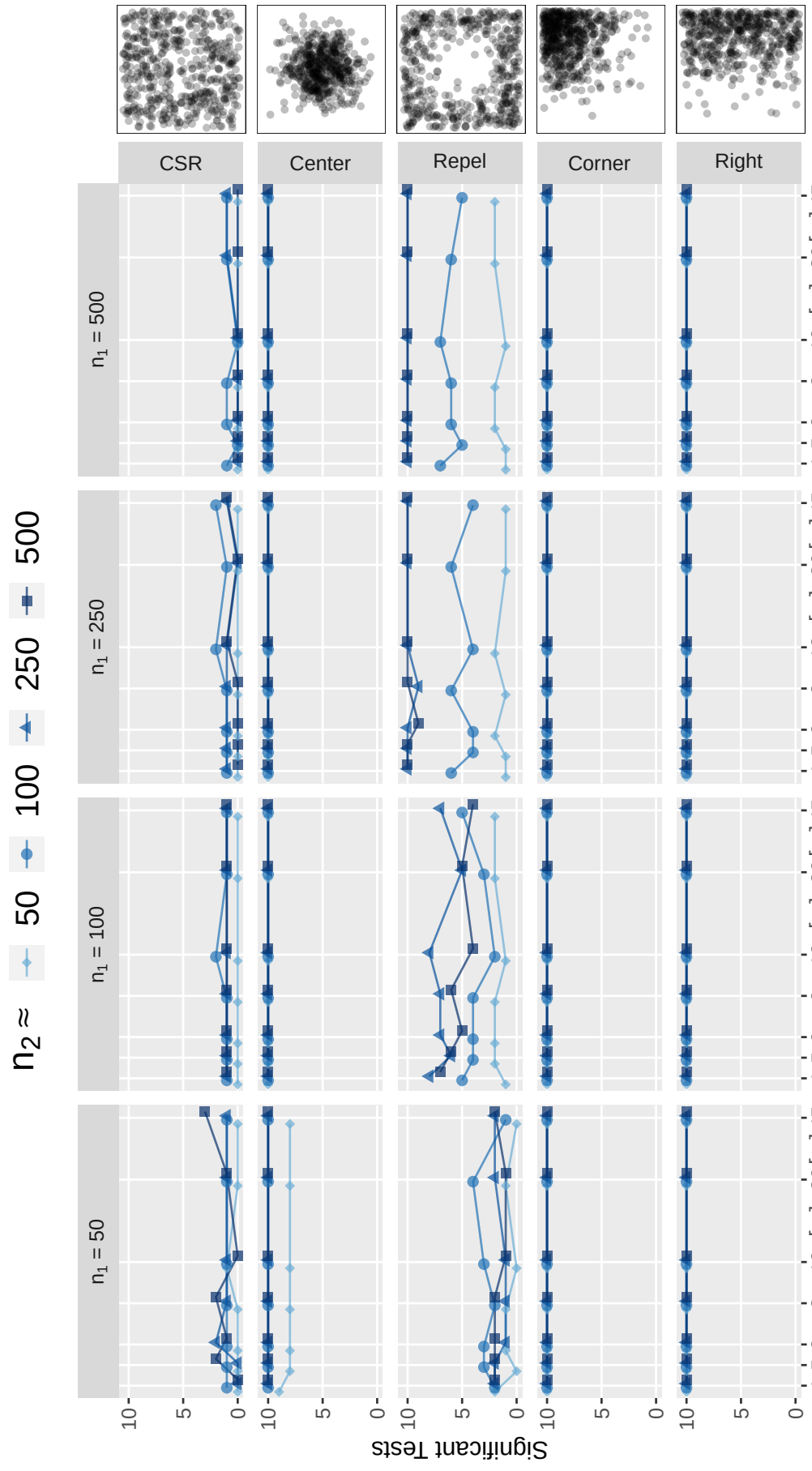


Figure 12: A grid of line graphs showing the results of a simulation comparing multiple rotations of the modified Syrjala test using double weightings of the squared differences in the ECDFs (DWS). The grid column name indicates the CSR sample size (n_1), and the grid row indicates the shape of the second 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.

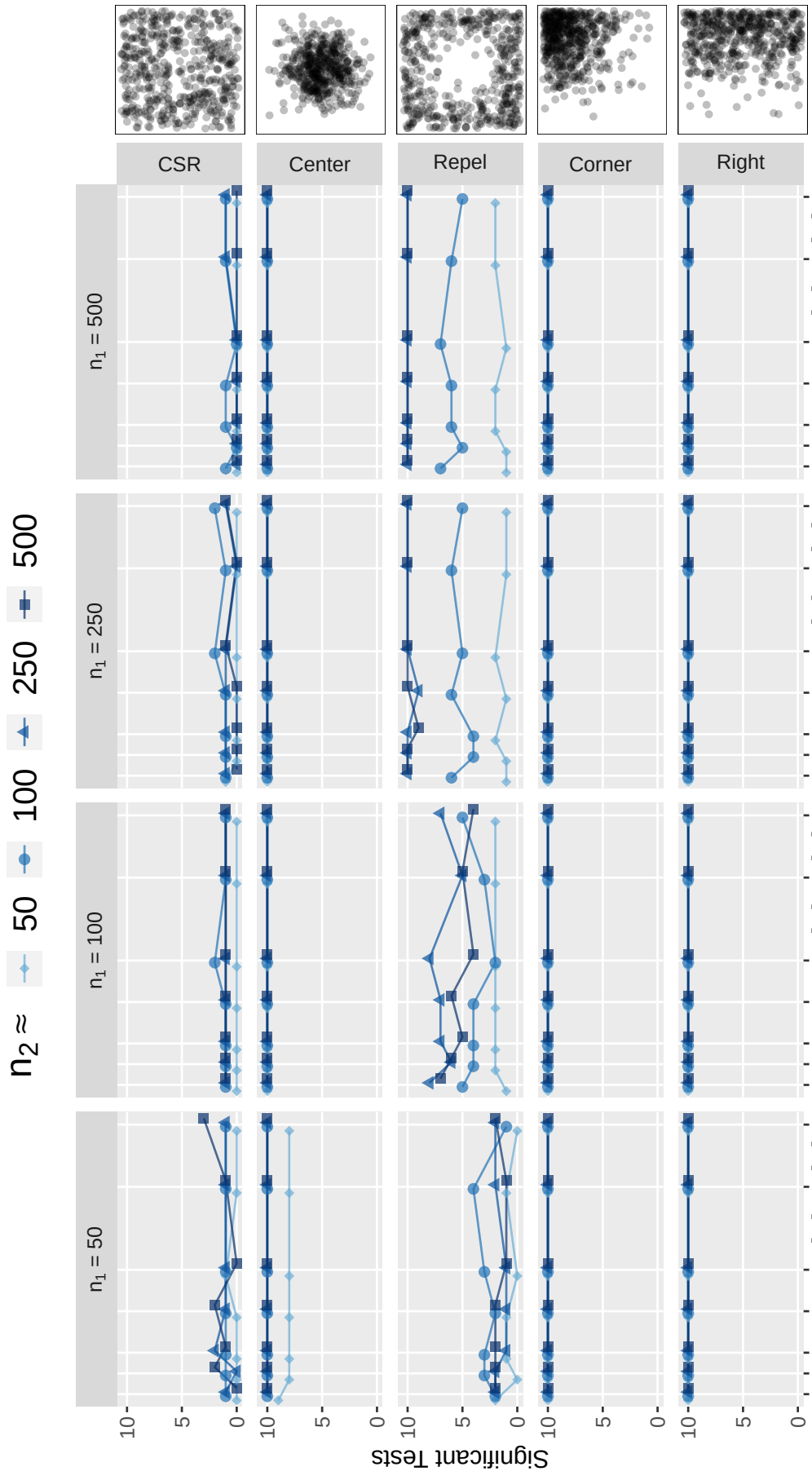


Figure 13: A grid of line graphs showing the results of a simulation comparing multiple rotations of the modified Syrjala test using unweighted squared differences in the ECDFs (UWS). The grid column name indicates the CSR sample size (n_1), and the grid row indicates the shape of the second sample. 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.

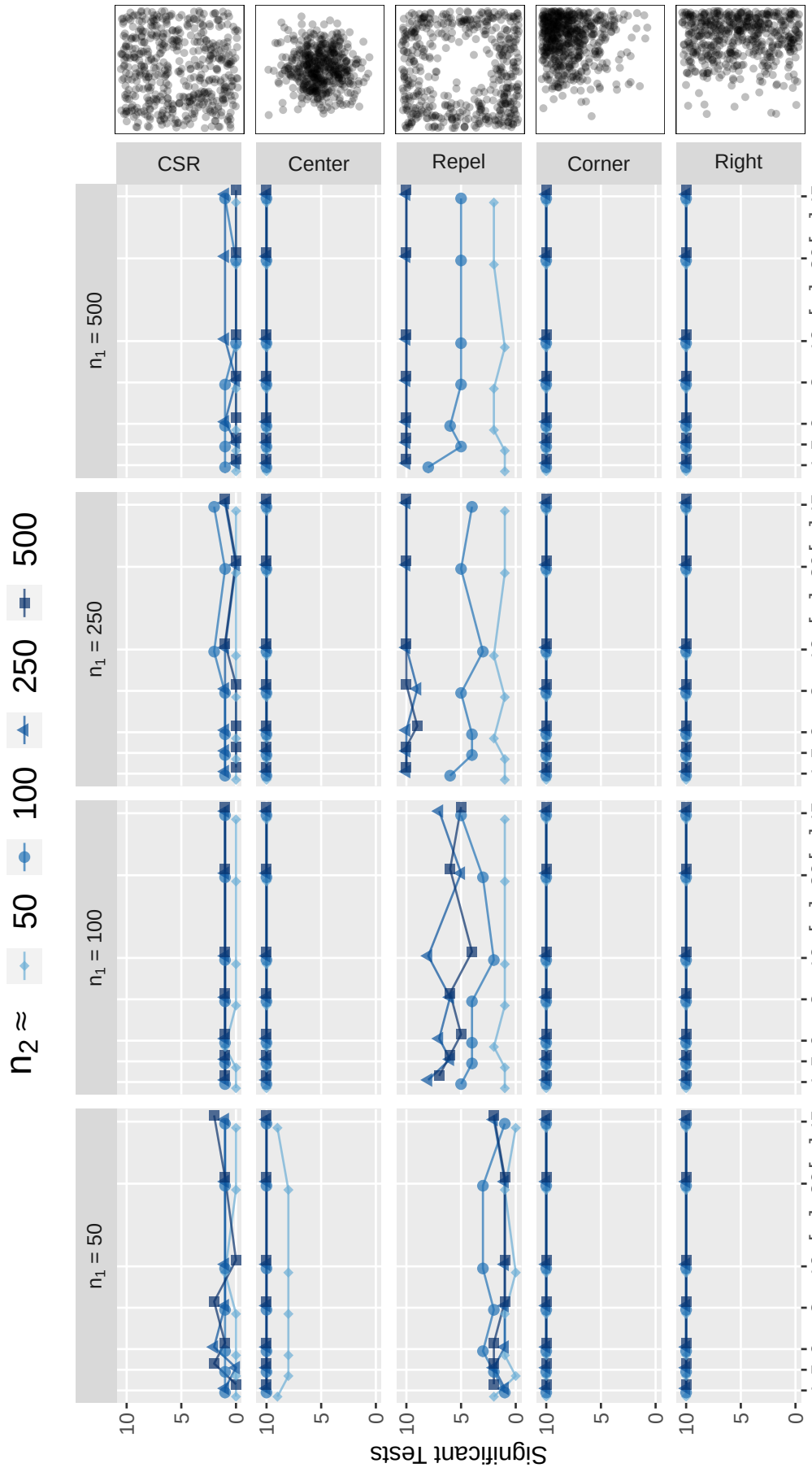


Figure 14: A grid of line graphs showing the results of a simulation comparing multiple rotations of the modified Syrjala test using double weightings of the absolute differences in the ECDFs (DWA). The grid column name indicates the CSR sample size (n_1), and the grid row indicates the shape of the 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.

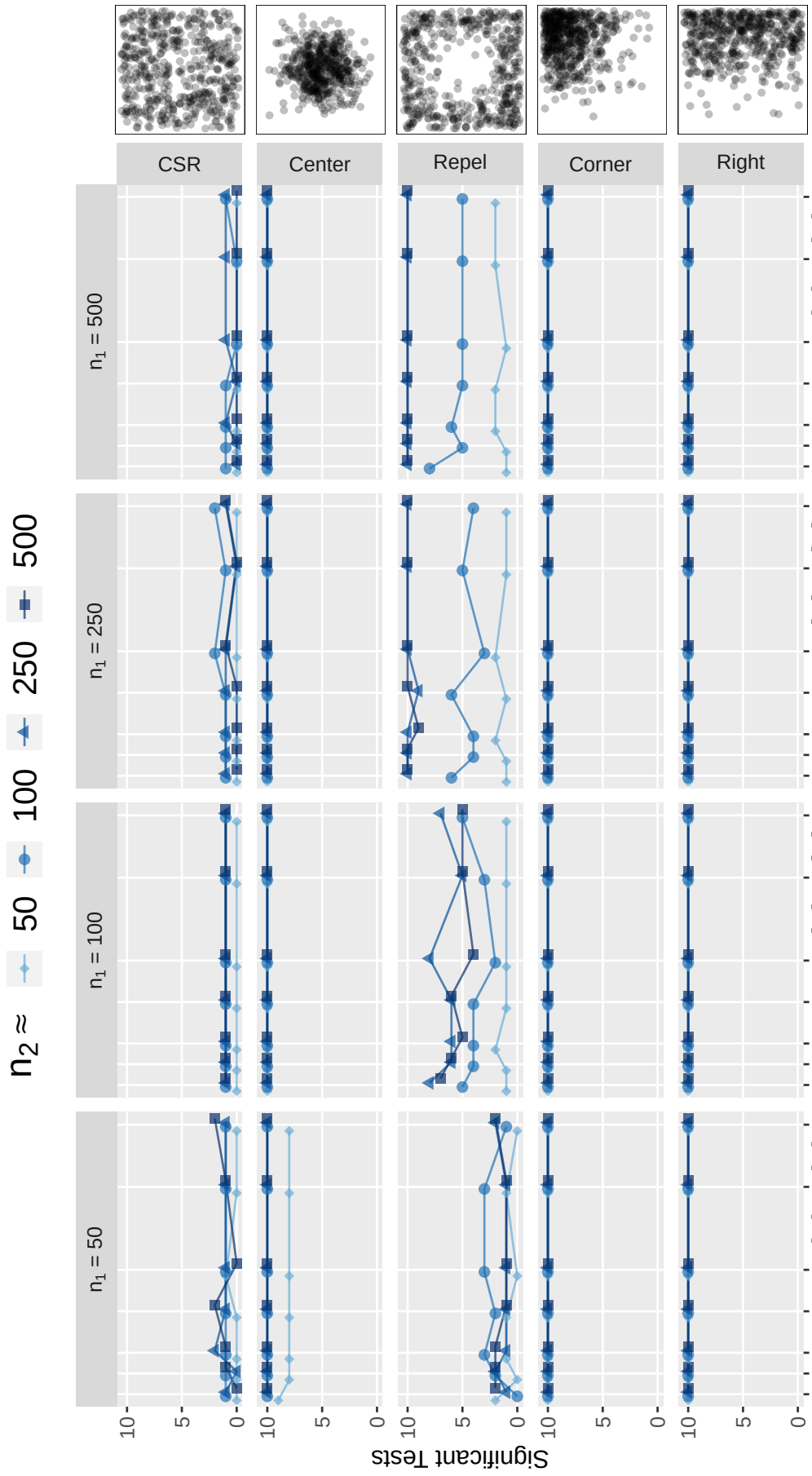


Figure 15: A grid of line graphs showing the results of a simulation comparing multiple rotations of the modified Syrjala test using unweighted absolute differences in the ECDFs (RWS). The grid column name indicates the CSR sample size (n_1), and the grid row indicates the shape of the second sample. 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.

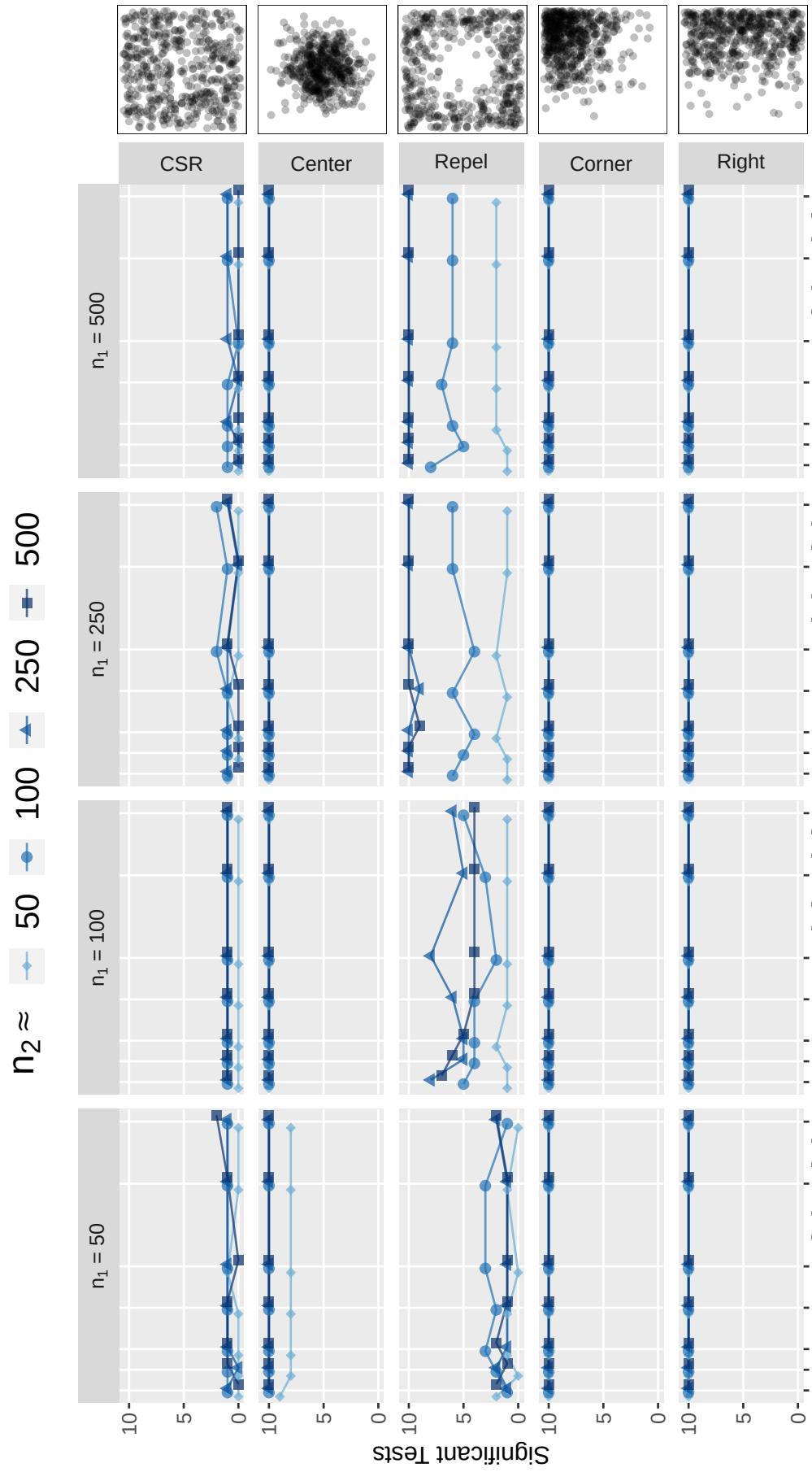


Figure 16: A grid of line graphs showing the results of a simulation comparing multiple rotations of the modified Syrjala test using reverse weightings of the absolute differences in the ECDFs (RWA). The grid column name indicates the CSR sample size (n_1), and the grid row indicates the shape of the 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.

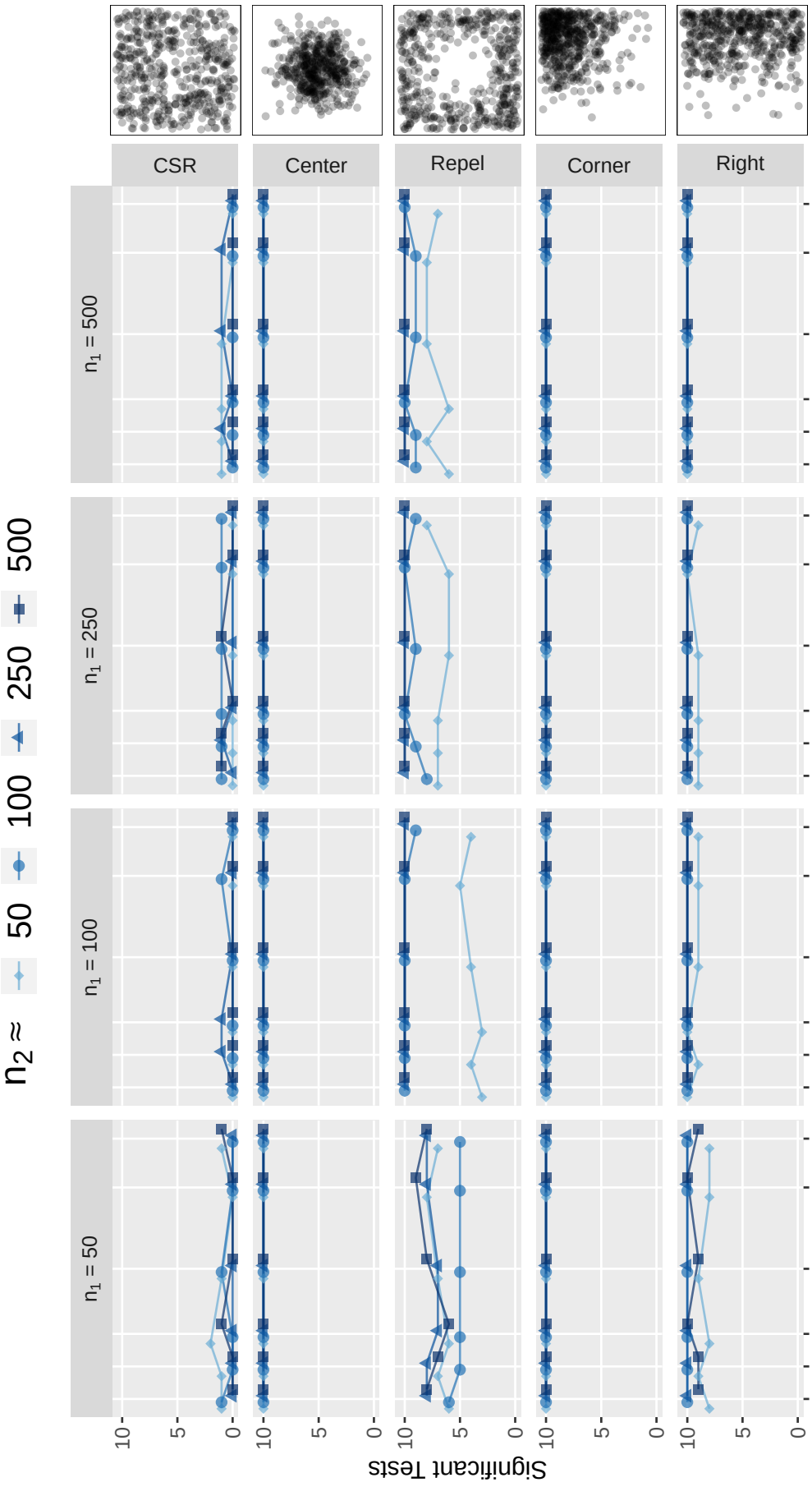


Figure 17: A grid of line graphs showing the results of a simulation comparing multiple randomly selected proportions of points for the origins of the toroidal shifts of the modified Syrjala test using double weightings of the squared differences in the ECDFs (DWS). The grid column name indicates the 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.

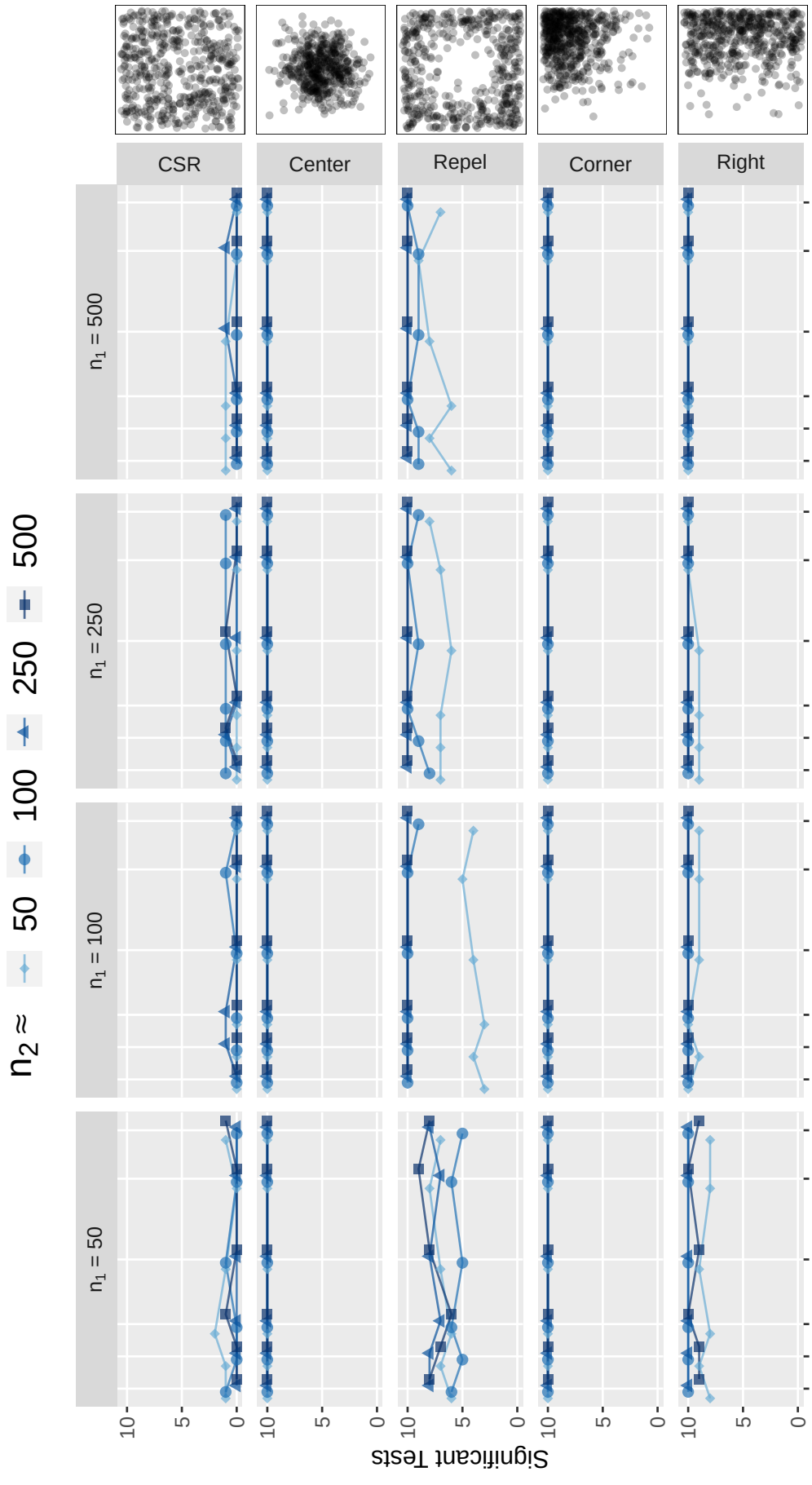


Figure 18: A grid of line graphs showing the results of a simulation comparing multiple randomly selected proportions of points for the origins of the toroidal shifts of the modified Syrjala test using unweighted squared differences in the ECDFs (UWS). The grid column name indicates the CSR sample size (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.

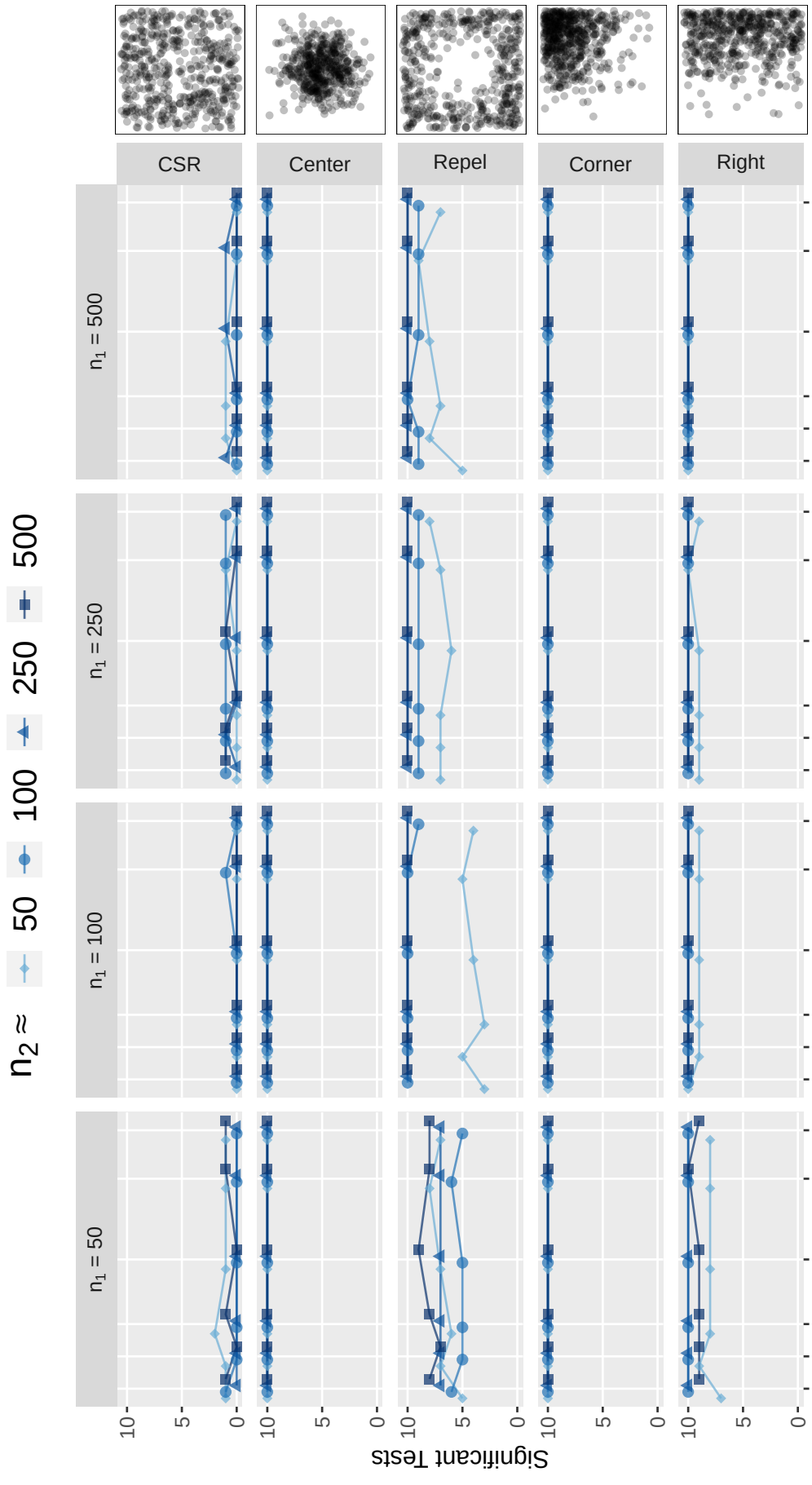


Figure 19: A grid of line graphs showing the results of a simulation comparing multiple randomly selected proportions of points for the origins of the toroidal shifts of the modified Syrjala test using double weightings of the absolute differences in the ECDFs (DWA). The grid column name indicates the 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.

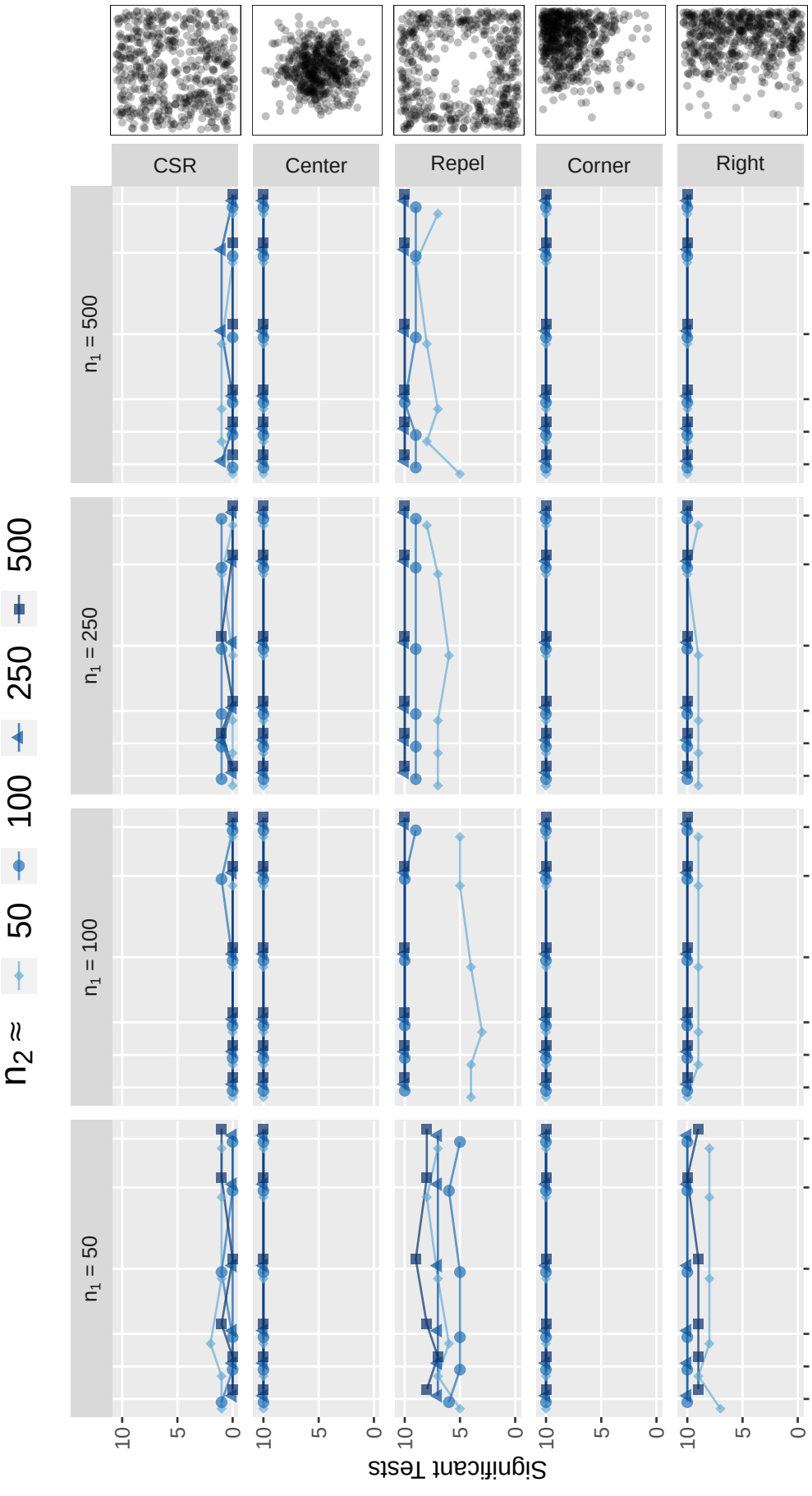


Figure 20: A grid of line graphs showing the results of a simulation comparing multiple randomly selected proportions of points for the origins of the toroidal shifts of the modified Syrjala test using unweighted absolute differences in the ECDFs (UWA). The grid column name indicates the CSR sample size (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.

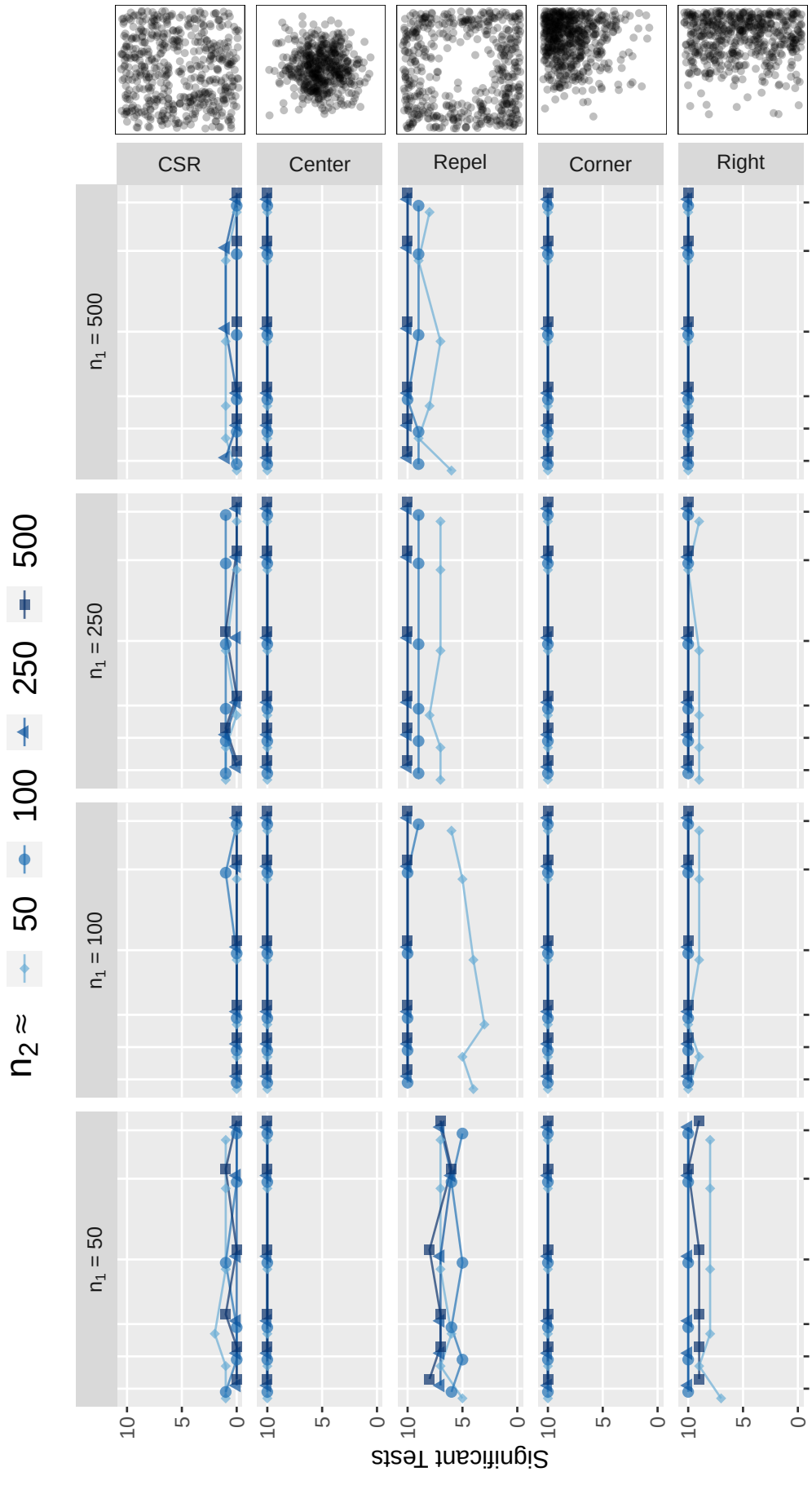


Figure 21: A grid of line graphs showing the results of a simulation comparing multiple randomly selected proportions of points for the origins of the toroidal shifts of the modified Syrjala test using reverse weightings of the absolute differences in the ECDFs (RWA). The grid column name indicates the 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.