The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only ANOVA Procedures

Jacob O. Wobbrock,[1] Leah Findlater,[1] Darren Gergle,[2] James J. Higgins[3]

[1]The Information School
DUB Group
University of Washington
Seattle, WA 98195 USA
{wobbrock, leahkf}@uw.edu

[2]School of Communication
Northwestern University
Evanston, IL 60208 USA
dgergle@northwestern.edu

[3]Department of Statistics
Kansas State University
Manhattan, KS 66506 USA
jhiggins@ksu.edu

ABSTRACT

Nonparametric data from multi-factor experiments arise often in human-computer interaction (HCI). Examples may include error counts, Likert responses, and preference tallies. But because multiple factors are involved, common nonparametric tests (*e.g.*, Friedman) are inadequate, as they are unable to examine interaction effects. While some statistical techniques exist to handle such data, these techniques are not widely available and are complex. To address these concerns, we present the *Aligned Rank Transform (ART)* for nonparametric factorial data analysis in HCI. The ART relies on a preprocessing step that "aligns" data before applying averaged ranks, after which point common ANOVA procedures can be used, making the ART accessible to anyone familiar with the F-test. Unlike most articles on the ART, which only address two factors, we generalize the ART to *N* factors. We also provide *ARTool* and *ARTweb*, desktop and Web-based programs for aligning and ranking data. Our re-examination of some published HCI results exhibits advantages of the ART.

Author Keywords: Statistics, analysis of variance, ANOVA, factorial analysis, nonparametric data, F-test.

ACM Classification Keywords: H.5.2 [Information interfaces and presentation]: User interfaces – *evaluation/methodology*, *theory and methods*.

General Terms: Experimentation, Measurement, Theory.

INTRODUCTION

Studies in human-computer interaction (HCI) often generate nonparametric data from multiple independent variables. Examples of such data may include error counts, Likert responses, or preference tallies. Often complicating this picture are correlated data arising from repeated measures. Two analysis approaches for data like these appear regularly in the HCI literature. The first simply uses a parametric F-test, which risks violating ANOVA assumptions and inflating Type I error rates. The second uses common *one-way* nonparametric tests (*e.g.*, Friedman), foregoing the examination of interaction effects. Although some methods for handling nonparametric factorial data exist (see Table 1) [12], most are not well known to HCI researchers, require advanced statistical knowledge, and are not included in common statistics packages.[1]

Table 1. Some possible analyses for nonparametric data.

Method	Limitations
General Linear Models (GLM)	Can perform factorial parametric analyses, but cannot perform nonparametric analyses.
Mann-Whitney *U*, Kruskal-Wallis	Can perform nonparametric analyses, but cannot handle repeated measures or analyze multiple factors or interactions.
Wilcoxon, Friedman	Can perform nonparametric analyses and handle repeated measures, but cannot analyze multiple factors or interactions.
χ^2, Logistic Regression, Generalized Linear Models (GZLM)	Can perform factorial nonparametric analyses, but cannot handle repeated measures.
Generalized Linear Mixed Models (GLMM), Generalized Estimating Equations (GEE)	Can perform factorial nonparametric analyses and handle repeated measures, but are not widely available and are complex.
Kaptein *et al.*'s [7] nonparametric method	Can perform factorial nonparametric analyses and handle repeated measures, but requires different mathematics and software modules for each type of experiment design. ([7] focused only on 2×2 mixed designs.)
Aligned Rank Transform (ART)	Can perform factorial nonparametric analyses and handle repeated measures. Requires only an ANOVA after data alignment and ranking, provided for by *ARTool* or *ARTweb*.

A remedy to the paucity of nonparametric factorial analyses would be a procedure that retains the familiarity and interpretability of the familiar parametric F-test. We present just such an analysis called the *Aligned Rank Transform* (*ART)*. The ART relies on an alignment and ranking step before using F-tests. We offer two equivalent tools to do the alignment and ranking, one for the desktop (*ARTool*) and one on the Web (*ARTweb*). Unlike the advanced methods

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.
Copyright 2011 ACM 978-1-4503-0267-8/11/05....$10.00.

[1] A useful page from statistics consulting at UCLA for choosing the right statistical test has a conspicuous omission marked with "???", which appears in the slot for analyzing factorial ordinal data. See http://www.ats.ucla.edu/stat/mult_pkg/whatstat/default.htm. Similarly, another useful page giving the rationale for myriad statistical tests fails to describe any nonparametric factorial analysis for repeated measures data. See http://www.ats.ucla.edu/stat/spss/whatstat/whatstat.htm. Both pages were accessed January 12, 2011.

shown in Table 1, researchers only familiar with ANOVA can use, interpret, and report results from the ART.

We describe the ART and contribute: (1) its generalization from two factors to N factors, (2) the *ARTool* and *ARTweb* programs for easy alignment and ranking, and (3) a re-examination of some published HCI data.

THE ALIGNED RANK TRANSFORM FOR *N* FACTORS

This section describes the Aligned Rank Transform, generalizing it to an arbitrary number of factors. The ART is for use in circumstances similar to the parametric ANOVA, except that the response variable may be continuous or ordinal, and is not required to be normally distributed.

Background

Rank transformations have appeared in statistics for years [12]. Conover and Iman's [1] Rank Transform (RT) applies ranks, averaged in the case of ties, over a data set, and then uses the parametric F-test on the ranks, resulting in a nonparametric factorial procedure. However, it was subsequently discovered that this process produces inaccurate results for interaction effects [5,11], making the RT method unsuitable for factorial designs.

The Aligned Rank Transform (ART) [2,10] corrects this problem, providing accurate nonparametric treatment for both main and interaction effects [4,5,11]. It relies on a preprocessing step that first "aligns" the data for each effect (main or interaction) before assigning ranks, averaged in the case of ties. Data alignment is an established process in statistics [6] by which effects are estimated as marginal means and then "stripped" from the response variable so that all effects but one are removed.

Let's consider an example: In a two-factor experiment with effects A, B, and $A*B$, and response Y, testing for the significance of effect A means first stripping from Y estimates of effects B and $A*B$, leaving only a possible effect of A behind. This alignment results in Y_A', whose values are then ranked, producing Y_A''. Lastly, a full-factorial ANOVA is run with Y_A'' as the response and model terms A, B, and $A*B$, but importantly, only the effect of A is examined in the ANOVA table; B and $A*B$ are ignored. This process is then repeated for the effects of B and $A*B$, *i.e.*, on aligned ranks Y_B'' and Y_{A*B}'', respectively. Thus, to use the ART, responses Y from a study must be aligned and ranked for each effect of interest. This is a tedious process to do by hand, but *ARTool* or *ARTweb* make it easy.

ART Procedure for *N* Factors

We present the ART procedure in five steps:

Step 1. Compute residuals. For each raw response Y, compute its residual as

$$residual = Y - cell\ mean\ .$$

The cell mean is the mean for Y's "cell," *i.e.*, the average of all responses whose factors' levels match that of the Y response for which we're computing the residual. As Table 2 shows, cell means are computed using Y values from rows with matching levels of independent variables (X_1 and X_2). Thus, the cell mean in row 1 is the mean Y of s1 and s5.

Table 2. Example calculation of cell means.

Subject	X_1	X_2	Y	*cell mean*
s1	a	a	12	(12+19)/2
s2	a	b	7	(7+16)/2
s3	b	a	14	(14+14)/2
s4	b	b	8	(8+10)/2
s5	a	a	19	(12+19)/2
s6	a	b	16	(7+16)/2
s7	b	a	14	(14+14)/2
s8	b	b	10	(8+10)/2

Step 2. Compute estimated effects for all main and interaction effects. Let A, B, C, and D be factors with levels A_i, $i = 1..a$; B_j, $j = 1..b$; C_k, $k = 1..c$; D_ℓ, $\ell = 1..d$.

Let $\overline{A_i}$ be the mean response Y_i for rows where factor A is at level i. Let $\overline{A_iB_j}$ be the mean response Y_{ij} for rows where factor A is at level i and factor B is at level j. And so on. Let μ be the grand mean of Y over all rows.

One-way effects. The estimated main effect for a factor A with response Y_i is

$$= \overline{A_i} - \mu\,.$$

Two-way effects. The estimated effect for an $A*B$ interaction with response Y_{ij} is

$$= \overline{A_iB_j} - \overline{A_i} - \overline{B_j} + \mu\,.$$

Three-way effects. The estimated effect for an $A*B*C$ interaction with response Y_{ijk} is

$$= \overline{A_iB_jC_k} - \overline{A_iB_j} - \overline{A_iC_k} - \overline{B_jC_k} + \overline{A_i} + \overline{B_j} + \overline{C_k} - \mu\,.$$

Four-way effects. The estimated effect for an $A*B*C*D$ interaction with response $Y_{ijk\ell}$ is

$$\begin{aligned}&= \overline{A_iB_jC_kD_\ell} - \overline{A_iB_jC_k} - \overline{A_iB_jD_\ell} - \overline{A_iC_kD_\ell} - \overline{B_jC_kD_\ell}\\&+ \overline{A_iB_j} + \overline{A_iC_k} + \overline{A_iD_\ell} + \overline{B_jC_k} + \overline{B_jD_\ell} + \overline{C_kD_\ell}\\&- \overline{A_i} - \overline{B_j} - \overline{C_k} - \overline{D_\ell} + \mu\,.\end{aligned}$$

N-way effects. The generalized estimated effect for an N-way interaction is

$$\begin{aligned}&= \overline{N\text{ way}}\\&- \sum\left(\overline{N\text{-1 way}}\right) + \sum\left(\overline{N\text{-2 way}}\right) - \sum\left(\overline{N\text{-3 way}}\right) + \sum\left(\overline{N\text{-4 way}}\right)\\&\ldots\\&- \sum\left(\overline{N\text{-}h\text{ way}}\right)\ \text{// if } h \text{ is odd, or}\\&+ \sum\left(\overline{N\text{-}h\text{ way}}\right)\ \text{// if } h \text{ is even}\\&\ldots\\&- \mu\ \text{// if } N \text{ is odd, or}\\&+ \mu\ \text{// if } N \text{ is even}\,.\end{aligned}$$

Step 3. Compute aligned response Y'. The calculation is

$$\begin{aligned}Y' &= \text{residual} + \text{estimated effect, } \textit{i.e.}\text{,}\\&= \text{result from Step 1} + \text{result from Step 2.}\end{aligned}$$

Step 4. Assign averaged ranks *Y″*. Assign averaged ranks to a column of aligned observations *Y′* to create *Y″*. The smallest *Y′* receives rank 1, the next smallest *Y′* receives rank 2, and so on until the largest of *r* values receives rank *r*. In the case of a tie among *k* values, the average rank is the sum of ranks divided by *k*.

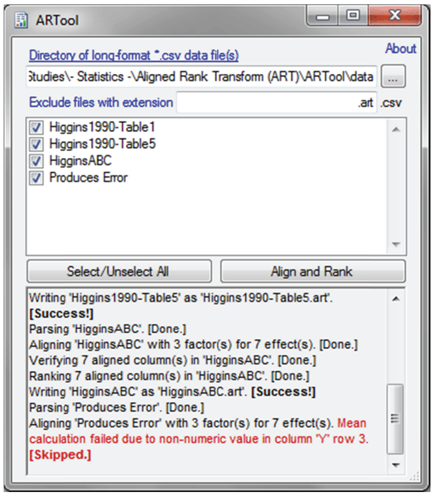
Step 5. Perform a full-factorial ANOVA on *Y″*. All main and interaction effects should be included in the model, but only the result corresponding to the effect for which *Y* was aligned as *Y′* should be considered. A fixed-effects ANOVA or a mixed-effects model analysis of variance can be used, the latter being useful for repeated measures [8].[2] Also, *post hoc* comparisons can be used; however, comparisons should be made only within effects for which the data was aligned.

Ensuring Correctness

ART users have two opportunities for ensuring correctness. First, every column of aligned responses *Y′* should sum to zero; *ARTool* and *ARTweb* verify this for the user. Second, a full-factorial ANOVA performed on the aligned (not ranked) responses *Y′* should show all effects stripped out (F=0.00, *p*=1.00) except for the effect for which the data were aligned. (An ANOVA on ranked responses *Y″* will often show nearly F=0.00, *p*=1.00, but the stripping is rarely as complete as for *Y′*.) The ability to check one's results is why *ARTool* and *ARTweb* produce not only ranked columns *Y″*, but also aligned columns *Y′*.

ARTOOL AND *ARTWEB*

To make alignment and ranking easy, we built *ARTool* (Fig. 1), which parses long-format data tables and produces aligned and ranked responses for all main and interaction effects. For example, with two factors and their interaction, *ARTool* produces three aligned columns and three ranked columns, all in *.csv format.



Directory of long-format *.csv data file(s)
Studies\- Statistics -\Aligned Rank Transform (ART)\ARTool\data

Exclude files with extension .art .csv

☑ Higgins1990-Table1
☑ Higgins1990-Table5
☑ HigginsABC
☑ Produces Error

```
Writing 'Higgins1990-Table5' as 'Higgins1990-Table5.art'.
[Success!]
Parsing 'HigginsABC'. [Done.]
Aligning 'HigginsABC' with 3 factor(s) for 7 effect(s). [Done.]
Verifying 7 aligned column(s) in 'HigginsABC'. [Done.]
Ranking 7 aligned column(s) in 'HigginsABC'. [Done.]
Writing 'HigginsABC' as 'HigginsABC.art'. [Success!]
Parsing 'Produces Error'. [Done.]
Aligning 'Produces Error' with 3 factor(s) for 7 effect(s). Mean
calculation failed due to non-numeric value in column 'Y' row 3.
[Skipped.]
```

Figure 1. *ARTool* processes long-format data tables giving detailed feedback.

ARTool makes no assumptions about column names, level types or values, or row order. *ARTool* does assume that data is in long-format (one response per row), that the first column is the subject identifier (*S*), and that the last column is a numeric response (*Y*). All intervening columns are assumed to be factors (X_i's). In the case of an error, *ARTool* produces descriptive messages to aid researchers in locating and remedying problems. Because *ARTool* reproduces an *N*-factor table's original columns, *ARTool*'s output contains $(2+N) + 2(2^N\text{-}1)$ columns. *ARTool* is implemented in C# using the .NET 2.0 framework. Its Web-based equivalent is *ARTweb*, a Java program that runs in a Web browser. Both tools are available at http://faculty.washington.edu/wobbrock/art/ .

RE-EXAMINATION OF DATA

In this section, we briefly re-examine some published HCI data. Because the ART procedure has been vetted in the statistics literature, our purpose is to show its utility and relevance, not prove its correctness. The first case shows how the approach can be used to provide interaction effects not available with Friedman tests. The second case shows how the ART can be used to free analysts from the distributional assumptions of ANOVA. The third case shows the nonparametric testing of repeated measures data.

Findlater *et al.* (CHI 2009)

In a study concerning adaptive menus, Findlater *et al.* [3] collected satisfaction ratings (1-7) from 24 participants on two independent factors: *Accuracy* (2 levels, between-subjects) and *Interface* (3 levels, within-subjects). In their paper, the authors used two separate Friedman tests to find satisfaction differences across the three levels of *Interface* within each level of *Accuracy*. The authors found no such differences, but commented on a possible interaction unexaminable by the Friedman test: "[I]n the High accuracy condition the mean rating for Short-Onset was lowest, whereas in the Low accuracy condition it was highest" [3] (p. 1660). An (inappropriate) parametric repeated measures ANOVA yields nonsignificant main effects for *Accuracy* ($F_{1,22}$=0.01, *p*=.936) and *Interface* ($F_{2,44}$=0.41, *p*=.668), but a significant interaction ($F_{2,44}$=4.47, *p*=.017). The nonparametric ART method also yields nonsignificant main effects for *Accuracy* ($F_{1,22}$=0.01, *p*=.920) and *Interface* ($F_{2,44}$=0.65, *p*=.529) and a significant interaction ($F_{2,44}$=4.12, *p*=.023). Incidentally, because this was a 2×2 mixed design, Kaptein *et al.*'s [7] online tool can also be used, and yields a similar interaction *p*-value (*p*=.020).

MacKenzie & Zhang (CHI 1999)

Nonparametric analyses are useful when ANOVA assumptions may be violated. In a 20-session longitudinal study comparing the OPTI and QWERTY stylus keyboards, MacKenzie & Zhang [9] used a parametric repeated measures ANOVA to analyze error rates for within-subjects factors *Session* ($F_{19,76}$=4.43, *p*<.001) and *Method* ($F_{1,4}$=12.14, *p*=.025), finding that OPTI had fewer errors but that both methods' errors increased over time. However, no *Session*Method* interaction was reported. Our re-examination of their data shows that under a parametric ANOVA, the interaction was marginal ($F_{19,76}$=1.57, *p*=.087). It also shows, however, that these data were non-normal for both OPTI (Shapiro-Wilk *W*=.95, *p*=.001) and QWERTY (*W*=.96, *p*=.003). In fact, error rates were lognormal for both OPTI (Kolmogorov's *D*=.08, *p*=.139) and QWERTY (*D*=.04, *p*=.150). If the data had been log-transformed, the *Session*Method* interaction would have, in fact, been

[2] A discussion of fixed-effects ANOVAS versus models with random effects is beyond the current scope.

significant ($F_{19,76}$=3.46, *p*<.001). An alternative to using a log-transform to correct for non-normality would have been to use a nonparametric procedure on the original data. With the ART, the main effects of *Session* ($F_{19,76}$=5.19, *p*<.001) and *Method* ($F_{1,4}$=10.64, *p*=.031) are still significant, and now their interaction is also significant ($F_{19,76}$=1.85, *p*=.032). Although OPTI had fewer overall errors, its errors increased faster than that of QWERTY, being quite similar from the 15th-20th sessions.

Wobbrock *et al.* (UIST 2007)

Wobbrock *et al.* [14] compared a new stroke recognizer to two published recognizers in a within-subjects study of gestures made at three different speeds by 10 participants. The researchers found that recognition errors were rare and highly skewed towards zero. They therefore regarded errors as "rare events" and used Poisson regression [13], which can be appropriate for such data. However, Poisson regression is a Generalized Linear Model (GZLM) and cannot handle repeated measures (see Table 1). Wobbrock *et al.* found significant effects of *Recognizer* ($\chi^2_{(2,N=780)}$=867.33, *p*<.001), *Speed* ($\chi^2_{(2,N=780)}$=24.56, *p*<.001), and *No. Train* ($\chi^2_{(1,N=780)}$=125.24, *p*<.001) on error counts. However, they found no significant interactions. By contrast, an (inappropriate) parametric mixed-effects model analysis of variance [8] gives the same significant main effects, but also a significant *Recognizer*No. Train* interaction ($F_{2,757}$=98.25, *p*<.001) and a marginal *Recognizer*Speed* interaction ($F_{4,757}$=2.27, *p*=.060). More appropriately, using the ART reduces the skew in the data, gives the same significant main effects, and now gives both significant *Recognizer*No. Train* ($F_{2,757}$=42.49, *p*<.001) and *Recognizer*Speed* ($F_{4,757}$=4.37, *p*=.002) interactions.

Some Limitations of the Aligned Rank Transform

In the above examples, the ART enabled the nonparametric testing of interactions, the avoidance of distributional assumptions, and the nonparametric testing of repeated measures. But the ART has limitations. For data exhibiting very high proportions of ties, the ART simply replaces those ties with tied ranks. If data exhibits extreme skew (*e.g.*, power-law distributions), the ART, as with any rank-based transform, will reduce that skew, which may be undesirable if distributions are meaningful. Lastly, alignment works best for completely randomized designs; it also works for other designs, but effects may not be entirely stripped out.

CONCLUSION

We have presented the Aligned Rank Transform for nonparametric analysis of factorial experiments using the familiar F-test. The ART offers advantages over more complex methods in its simplicity and usability. We offered the first generalized mathematics for an *N*-way ART and programs called *ARTool* and *ARTweb* to make alignment and ranking easy. By providing three examples of published data re-examined using the ART, we exhibited its benefits. It is our hope that researchers in HCI will find this convenient nonparametric method and our tool as useful as we have.

ACKNOWLEDGEMENTS

Thanks to Nelson Nogales, I. Scott MacKenzie, Krzysztof Z. Gajos, Parmit Chilana, and Jessica J. Tran. This work was supported in part by National Science Foundation grants IIS-0811063 and IIS-0811884.

REFERENCES

1. Conover, W.J. and Iman, R.L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician 35* (3), 124-129.
2. Fawcett, R.F. and Salter, K.C. (1984). A Monte Carlo study of the F test and three tests based on ranks of treatment effects in randomized block designs. *Communications in Statistics: Simulation and Computation 13* (2), 213-225.
3. Findlater, L., Moffatt, K., McGrenere, J. and Dawson, J. (2009). Ephemeral adaptation: The use of gradual onset to improve menu selection performance. *Proc. CHI '09*. New York: ACM Press, 1655-1664.
4. Higgins, J.J., Blair, R.C. and Tashtoush, S. (1990). The aligned rank transform procedure. *Proc. Conf. on Applied Statistics in Agriculture*. Kansas State, 185-195.
5. Higgins, J.J. and Tashtoush, S. (1994). An aligned rank transform test for interaction. *Nonlinear World 1* (2), 201-211.
6. Hodges, J.L. and Lehmann, E.L. (1962). Rank methods for combination of independent experiments in the analysis of variance. *Annals of Mathematical Statistics 33* (2), 482-497.
7. Kaptein, M., Nass, C. and Markopoulos, P. (2010). Powerful and consistent analysis of Likert-type rating scales. *Proc. CHI '10*. New York: ACM Press, 2391-2394.
8. Littell, R.C., Henry, P.R. and Ammerman, C.B. (1998). Statistical analysis of repeated measures data using SAS procedures. *J. Animal Science 76* (4), 1216-1231.
9. MacKenzie, I.S. and Zhang, S.X. (1999). The design and evaluation of a high-performance soft keyboard. *Proc. CHI '99*. New York: ACM Press, 25-31.
10. Salter, K.C. and Fawcett, R.F. (1985). A robust and powerful rank test of treatment effects in balanced incomplete block designs. *Communications in Statistics: Simulation and Computation 14* (4), 807-828.
11. Salter, K.C. and Fawcett, R.F. (1993). The art test of interaction: A robust and powerful rank test of interaction in factorial models. *Communications in Statistics: Simulation and Computation 22* (1), 137-153.
12. Sawilowsky, S.S. (1990). Nonparametric tests of interaction in experimental design. *Review of Educational Research 60* (1), 91-126.
13. Vermunt, J.K. (1997). *Log-linear Models for Event Histories*. Thousand Oaks, CA: Sage Publications.
14. Wobbrock, J.O., Wilson, A.D. and Li, Y. (2007). Gestures without libraries, toolkits or training: A $1 recognizer for user interface prototypes. *Proc. UIST '07*. New York: ACM Press, 159-168.