JDSSV

Journal of Data Science, Statistics, and Visualisation

MMMMMM YYYY, Volume VV, Issue II. *doi: XX.XXXXX/jdssv.v000.i00*

The effect of user interaction for understanding variable contributions to structure in linear projections

Nicholas Spyrison
Monash University

Dianne Cook
Monash University

Kimbal Marriott
Monash University

Abstract

Principal component analysis and other eigenvalue decomposition enjoy wide when it comes to visualizing multivariate spaces. The grand tour adds the continuity of near-by linear projections while moving towards a randomly selected basis. The radial tours keep this continuity of projection space while offering a means of controlling this animation. We perform a between-participant user study evaluating these biplots of the above methods. We measure accuracy and speed for a task of supervised clusters where participants indicate which variables contain the separation between two target clusters. Prolific.co is used to crowdsourced N=108 participants across 648 task evaluations. We vary across 3 dimensions of block parameterizations: location, shape, and data dimensionality. In summary, we find that use of radial tour tends to increase accuracy compared with the alternatives. This increase is relatively large compared with the effect from block dimension, though relatively small compared to the random effect of the participant. The use of the grand tour decreases the time takenm, which is also of modest size compared with the effect of the participant. Participants subjectively prefer to use the radial tour for this task.

Keywords: multivariate data, exploratory data analysis, grand tour, manual tour, dimension reduction, linear projections, linear embeddings, R.

1. Introduction

Imagine, it is that time of year, again. Your favorite data event is right around the corner. They did a great job tantalizing releasing new information piece by piece. You know the topic, you have your tool belt is at the ready. Your plan of attack is sound and thought-through. But what about the communication, you've seen this before, so-many measures to look at and digest let alone to convey to others. The audience reading your submission also needs to be rapidly brought up to speed on such a complex problem. The visualizations you use can make or break your submission, or more dangerously may even obscure an insight to the author.

Multivariate data (and measure space) is ubiquitous. Yet exploratory data analysis (EDA) (Tukey 1977) of such spaces becomes difficult, increasingly so as dimension increases. Numeric statistic summarization of data often doesn't explain the full complexity of the data or worse can lead to missing obvious visual patterns (Anscombe 1973; Coleman 1986; Goodman 2008; Matejka and Fitzmaurice 2017). Ideally, data would be visually explored, in its original variable-space, before and after model application, and statistical summarization. This is crucial to validating assumptions, identify outliers, finding visual peculiarities, and galvanizing insight in the best direction to take a study.

Previous user studies in dimension reduction compare across embedding- and display-dimensionality (Gracia et al. 2016; Wagner Filho et al. 2018). There are empirical metrics and comparisons used to describe non-linear reduction and how well and faithfully they embed the data (Bertini, Tatu, and Keim 2011; Liu et al. 2017; Sedlmair, Munzner, and Tory 2013; Maaten and Hinton 2008). There is a notable absence comparing linear techniques, finding which excel and under which contexts they should be employed, and when to best convey information to the audience.

Here, we compare and contrast the efficacy of 3 linear projection techniques through the application of a within-participant user study. Specific visualization is the primary factor of the study with a null hypothesis that visual has no impact on the accuracy nor the speed of evaluating a task. We design a task and measure evaluating its accuracy that involves participants indicating which variable(s) contain a difference in groups of different clusters. Data is simulated over several block parameterizations.

Section 2 discusses the visualization methods. Section 3 goes into the user study. Subsection 3.2 digs into the task and its evaluation. The results of the study are in section 4. Conclusion and future directions are covered in section 5. An accompanying tool is discussed in section 6.

2. Background, visual methods

2.1. Linear projection notation

Consider a numeric data matrix with n observations of p variables,

$$\mathbf{X}_{[n,p]} = (\mathbf{x}_1,\ \ldots\ \mathbf{x}_p)$$
$$\mathbf{x}_i = (x_{1i},\ \ldots\ x_{ni}) \mid i \in [1, p]$$

Let Y_{n*d} be the d-dimensional projection or embedding of $\mathbf{X}_{n*p}$ via matrix multiplication of a particular orthonormal basis matrix $\mathbf{B}_{p*d}$.

$$\mathbf{Y}_{n*d} = \mathbf{X}_{n*p}\mathbf{B}_{p*d} \mid \mathbf{B} \text{ is orthonormal}$$
$$\mathbf{y}_j = (y_{1j},\ \ldots\ y_{nj}) \mid j \in [i, d]$$

A matrix is said to be orthonormal if they are 1) orthogonal, that is all column pairs are independent, having a cross product of 0, and 2) normal, each column has a norm length of 1.

2.2. Principal component analysis

Principal component analysis is a good baseline of comparison for linear projections because of its frequent and broad use across discipline. Principal component analysis (PCA) (Pearson 1901) creates new components that are linear combinations of the original variables. The creation of these variables is ordered by decreasing variation which is orthogonally constrained to all previous components. While the full dimensionality is intact, the benefit comes from the ordered nature of the components. The first two or three components are typically used to approximate the variation multivariate data set, while the rest are discarded. We will allow participants to choose the components on the x and y axes from the first 4 principal components.

2.3. Data visualization tours

A data visualization *tour* is an animation of many linear projections across several identified target bases. One of the key features of the tour is the object permanence of the data points; that is to say by watching nearby frames one can track the relative changes of observations as the basis moves toward the next target basis. Various types of tours are enumerated by the selection or generation of their basis paths[lee_review_2021; cook_grand_2008]. To contrast with PCA, we compare with the *grand* and *radial* tours. Both of these methods use geodesically-interpolated orthonormal frames with a fixed step distance between frames.

Grand tours

In a grand tour(Asimov 1985) the target bases are selected randomly. The grand tour is the first and most widely known tour. It will serve as an intermediate unit of comparison which has continuity of data points in nearby frames along with the radial tour, but lacks the user control enjoyed by PCA and radial tours. This lack of control makes grand tours more of a generalist exploratory tool theoretically wouldn't excel at tasks with a pointed goal in mind.

Radial (manual) tours

The *manual* tour (Cook and Buja 1997; Spyrison and Cook 2020) defines its basis path by manipulating the basis contribution of a selected variable. A manipulation dimension is appended onto the projection plane, with a full contribution given to the selected

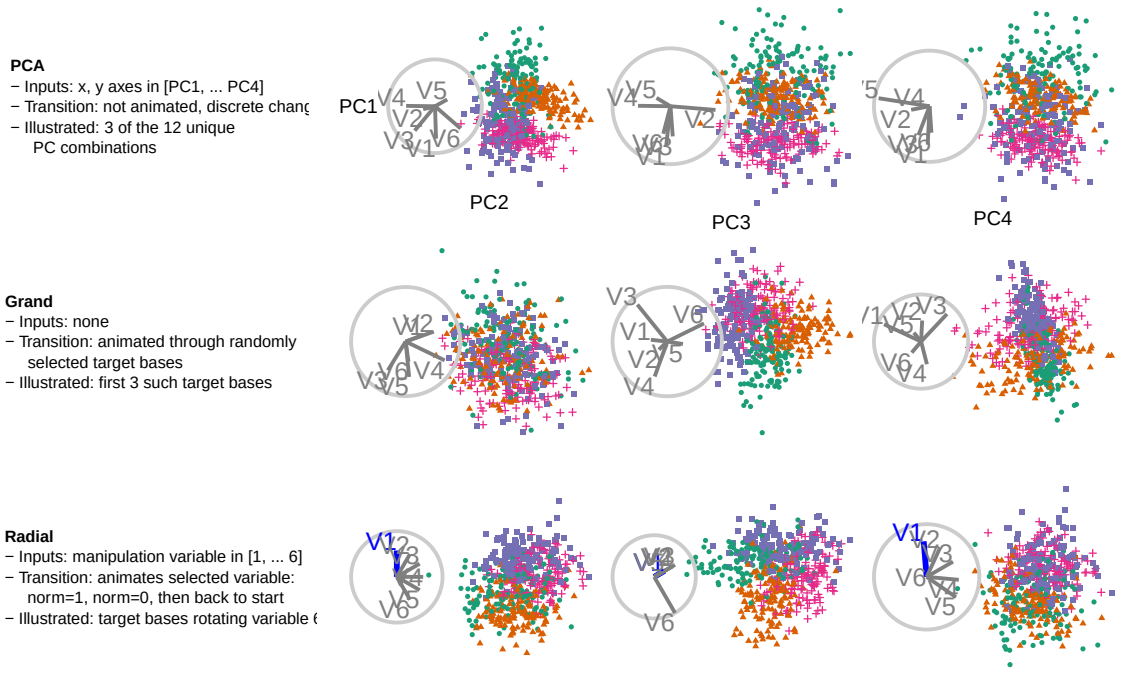

Figure 1: Example of the different visual factors. All use the same sort of biplot display to view linear projections of multivariate data. They differ in which bases are viewed which is influenced by the different factor inputs and whether or not are animated to convey the continuity of data points from 1 frame to the next.

variable. The target bases are then selected based on rotating this newly created manipulation space. The target bases are then similarly orthogonally-interpolated, data projected and rendered into an animation. In order for the variables to remain independent of each other, the contributions of the other variables must also change, *ie.* dimension space should maintain its orthonormal structure. A key feature of the manual tour is that it affords users a way to control the variable contributions of the next target basis. This means that such manipulations can be selected and queued in advance or select on the spot for human-in-the-loop analysis (Karwowski 2006). However, this navigation is relatively time-consuming due to the huge volume of p-space (an aspect of the curse of dimensionality (Bellman 1957)) and the abstract method of steering the projection basis. It is advisable to first identify a basis of particular interest and then use a manual tour as a finer, local exploration tool to observe how the contributions of the selected variable do or do not contribute to the feature of interest.

To simplify the task and keep its duration realistic, we consider a variant of the manual tour, called a *radial* tour. In a radial tour, the selected variable is allowed to change its magnitude of contribution, but not its angle; it must move along the direction of its original contribution radius. The radial tour benefits from both continuity of the data alongside grand tours, but also allows for user interaction similar to PCA.

Figure 1 illustrates the different visualization factors, the control of biplot aesthetics, and where they differ. The leftmost column discusses their differences alongside 3 of their character bases, albeit it is more difficult to see the benefit of animation in print.

3. User study

3.1. Objective

From our experience and application of linear projections we have some expectations about the factors. PCA excels at finding features that are in-plane with the first principal component. Whereas, the animated transitions of tours conveys a lot of meaningful structural information. The grand tour is good for initial exploration and explores a lot of different angles very quickly. While the manual (and radial) tours are best used as a more local exploration tool, to test the sensitivity of structure. Then for some subset of tasks we expect to find that the radial tour performs most accurately. Secondly, it maybe the case that the task can be performed faster than the alternatives. We are less certain with this there is no optimization or functional goal in the grand tour's generation of bases; it is possible that we are shown the planes that completely avoid information needed to perform the task.

We measure the accuracy and speed over 4 dimensions of block parameterizations. The null hypotheses can be stated as:

1, Accuracy)
H_0 : visualization factor does not impact task *accuracy*.
2, Speed)
H_0 : visualization factor does not impact task *speed*.

PCA will be used as a baseline for comparison as it is the most common linear embedding. The grand tour will act as a secondary control that will help evaluate the benefit of animation, where the object permanence of the data points, but without the ability to influence its path. Lastly, the radial tour should perform best as it benefits both from animation and being able to select an individual variable to change the contribution of.

3.2. Task and evaluation

The display was a 2D scatterplot with observations supervised. Cluster membership was mapped to shape and color. There were either 3 or 4 clusters each with the number of observations within each cluster. Participants were asked to 'check any/all variables that contribute more than average to the cluster separation green circles and orange triangles,' which was further explained in the explanatory video as 'mark and all variable that carry more than their fair share of the weight, or 1 quarter in the case of 4 variables.'

The instructions iterated several times in the video was: 1) Use the input controls to find a frame that contains separation between the clusters of green circles and orange triangles, 2) look at the orientation of the variable contributions in the gray circle, a visual depiction of basis, and 3) select all variables that contribute more than average in the direction of the separation in the scatterplot. Regardless of factor and block values participants were limited to 60 seconds for each evaluation of this task.

The evaluation measure of this task was designed with a couple of features in mind: 1) the sum of squares of the individual variable marks should be 1, and 2) symmetric about 0. With these in mind, we define the following measure for evaluating the task.

Let a dataset $\mathbf{X}$ be a simulation containing clusters of observations of different distributions. Let $\mathbf{X}_k$ be the subset of observations in cluster k containing the p variables.

$$\mathbf{X}_{np} = (x_1,\ ...\ x_p)$$
$$\mathbf{X}_{n_k p_k} = (x_1,\ ...\ x_p) \mid n_k \in [1, n],\ \text{is an observation subset of } \mathbf{X}$$

where

$x_{i,j,k}$ is scalar; the observation $i \in [1,\ ...\ n],$ of variable $j \in [1,\ ...\ p],$ of cluster $k \in [1,\ ...\ K]$

We define weights, W to be a vector explaining the variable-wise difference between 2 clusters. Namely the difference of each variable between clusters, as a proportion of the total difference, less $1/p$ the amount of difference each variable would hold if it were uniformly distributed. Participant responses, R are a vector of logical values, whether or not participant thinks the variable separates the two clusters more than if the difference were uniformly distributed. Then M is a vector of variable marks.

$$\begin{aligned} W_j &= \frac{(\overline{X_{j=1,k=1}} - \overline{X_{1,k=2}},\ ...\ (\overline{X_{j=p,k=1}} - \overline{X_{j=p,k=2}})}{\sum_{j=1}^{p}(|\overline{X_{j,k=1}} - \overline{X_{j,k=2}}|)} - \frac{1}{p} \\ &= (w_1,\ ...\ w_p) \end{aligned}$$

$$\begin{aligned} M &= I(r_j) * sign(w_j) * \sqrt{|w_j|} \\ &= (m_1,\ ...\ m_p) \end{aligned}$$

where I is the indicator function. Then the total marks for this task is the sum of this marks vector.

Each of the 3 periods introduced the participant to a new factor, where participants were first able to explore an untimed task with data under the simplest parameterizations. The training allows the participant to become familiar with the inputs and visual specific to the factor. Upon clicking a button to proceed text containing the correct answer displays with visual still intact to explore further. After the training, each participant performed 2 evaluation trials. After 60 seconds the display was turned off, though few participants elapsed this time. These evaluation trials were performed under different parameterizations as explained in section 3.4.

3.3. Factor application

Section 2 gives the sources and a description of the visual factors PCA, grand tours, and radial manual tours. The factors are tested within-participant, with each factor being evaluated by each participant. The order that factors are experienced in is controlled with the block assignment as illustrated below in Figure 4. Below we cover the aesthetic standardization, as well the unique input and display within each factor.

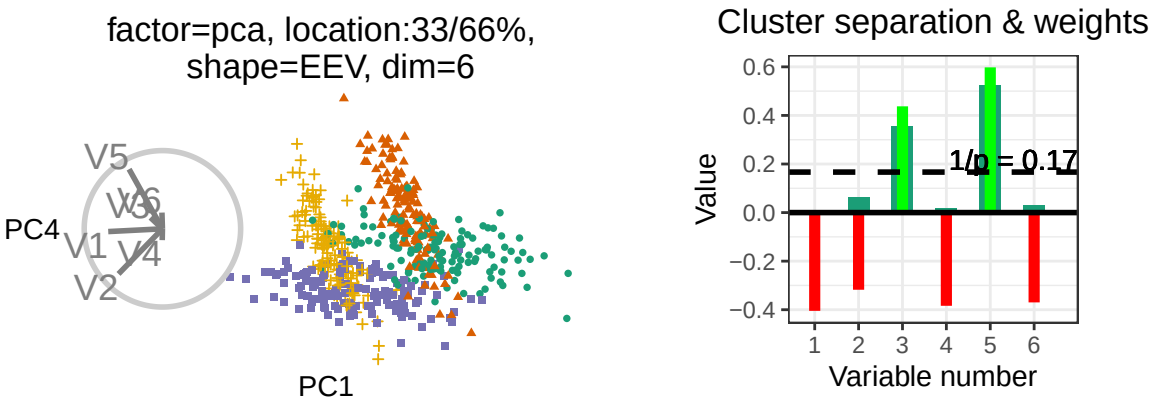

Figure 2: (L), PCA biplot of the components showing the most cluster separation with (R) illustration of the magnitude of cluster separation is for each variable (wide bar) and the weight of the marks given if a variable is selected (red/green line). The horizontal dashed line is 1 / dimensionality, the amount of separation each variable would have if evenly distributed. The weights equal the signed square of the difference between each variable value and the dashed line.

The visualization methods were standardized wherever possible. each factor was shown as a biplot, with variable contributions displayed on a unit circle. All aesthetic values (colors, shapes, sizes, absence of legend, and absence axis titles) were held consistent. Variable contributions were always shown left of the scatterplot embeddings with their aesthetic values consistent as well. What did vary between factors were their inputs which caused a discrete jump to another pair or principal components, were absent for the grand tour with target bases to animate through selected at random, or for the radial tour which variable should have its contribution animated. Key frames of each factor have been illustrated above in Figure 1.

PCA inputs allowed for users to select between the top 4 principal components for both the x and y-axis regardless of the data dimensionality (either 4 or 6). There was no user input for the grand tour, users were instead shown a 15-second animation of the same randomly selected path. Users were able to view the same clip up to 4 times within the time limit. Radial tours were also displayed at 5 frames per second within the interpolation step size of 0.1 radians. Users were able to swap between variables, upon which the display would change the start of radially increasing the contribution of the selected variable till it was full, zeroed, and then back to its initial contribution. The complete animation of any variable takes about 20 seconds and is almost fully in the projection frame at around 6 seconds. The starting basis of each is initialized to a half-clock design, where the variables were evenly distributed in half of the circle which is then orthonormalized. This design was created to be variable agnostic while maximizing the independence of the variables.

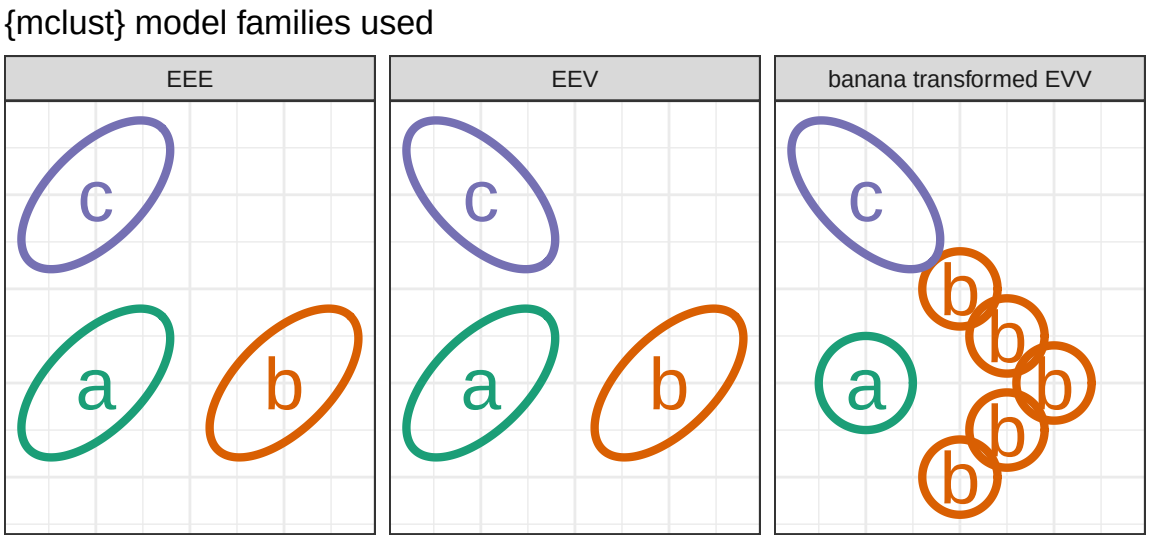

Figure 3: Ellipses of the isodensity of the model families used. Family labels are are the abbreviation for the clusters volume, shape, and orientation respectively, which are either equal or vary. We further change the EVV model by shifting fifths of the data in banana or chevron arrow shape.

3.4. Blocks and parameterization

In addition to visual factor, we vary the data across 3 aspects: 1) The location of the difference between clusters, by mixing a signal and a noise variable at different ratios, we control which variables contain cluster separation, 2) the shapes of the clusters, by changing the variance-covariance matrices, and 3) the dimensionality of the data.

Dimensionality is tested at 2 modest levels, namely, in 4 dimensions containing 3 clusters and 6 dimensions with 4 clusters. Each cluster samples 140 observations. Each dimension is originally distributed as $\mathcal{N}(2*I(signal), 1)$ | covariance Σ, before applying location mixing and standardizing by standard deviation). Signal variables have a correlation of 0.9 when they have equal orientation and -0.9 when their orientations vary. Noise variables were restricted to 0 correlation. Within each factor-period dimension is fixed with increasing difficulty, 4 then 6.

For choosing the shape of the clusters we follow the convention given by Scrucca, (Scrucca et al. 2016) who named and categorize 14 variants of model families containing 3 clusters. The name of the model family is the abbreviation of its respective volume, shape, and orientation, which are either equal or vary. We use the models EEE, EEV, and EVV, the latter is further modified by moving 4 fifths of the data out in a "V" or banana-like shape. Figure 3 shows the principal component biplot of the 3 model variants applied here. The training always uses 4 dimensions, while the 2 evaluations always contain 4 and 6 dimensions in the order of increasing difficulty. The evaluation periods use EEE, EEV, and EVV-banana respectively in increasing order of difficulty.

The separation of any pair of clusters is currently contained fully within a single variable at this point. To test the sensitivity to this we mix a noise variable with the signal-containing variable such that the difference in the clusters is mixed at the following

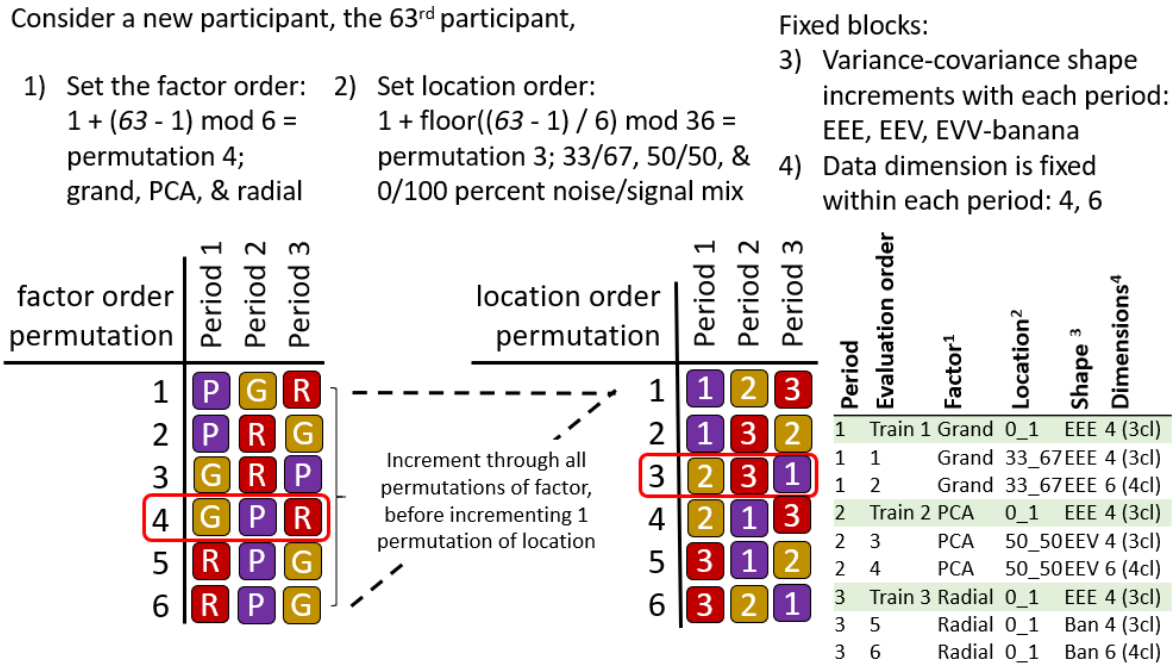

Figure 4: Illustration of how a hypothetical participant 63 is assigned factor and block parameterizations. Each of the 6-factor order permutations is exhausted before iterating to the next permutation of location order.

percentages: 0/100 (not mixed), 33/66, 50/50 (evenly mixed). The training always uses 4 dimensions, while the 2 evaluations always contain 4 and 6 dimensions in order of increasing difficulty. The training data does not mix signal Location mixing within an evaluation period is held constant and rotated through the 6 permutations of their order. Randomizing the order of the location mixing is controlled by iterating once after each of the 6-factor order permutations are evaluated. This is illustrated in Figure 4.

With this setup, we test the parameter space $dimension \in (4, 6)$, $shape \in (EEE, EEV, EVV-banana)$, $location \in (0/100, 33/66, 50/50)$ percent noise/signal mixing to evaluate the graphic display across the $factors \in (PCA, grand, radial)$. As we iterate through the possible permutations of these factors and location we perform an even evaluation of the full parameter space every 36 participants. Via pilot studies we estimate that 3 even block evaluations should be sufficient to identify the difference between the factors; we targeted for $N = 108$ participants.

In addition to the explicitly controlled block parameters, we will also be discussing each participant's evaluation order regardless of factor or location experiences. This will expose a learning effect from the repetition of being exposed to this data or problem. Keep in mind that the shape and location are always experienced in order of increasing difficulty.

3.5. Post-study survey

After the evaluation section of the study, participants were given a short survey containing questions gauging demographics, experience, and subjective evaluation of each factor on a 5-point Likert scale. The questions and possible responses are as follows:

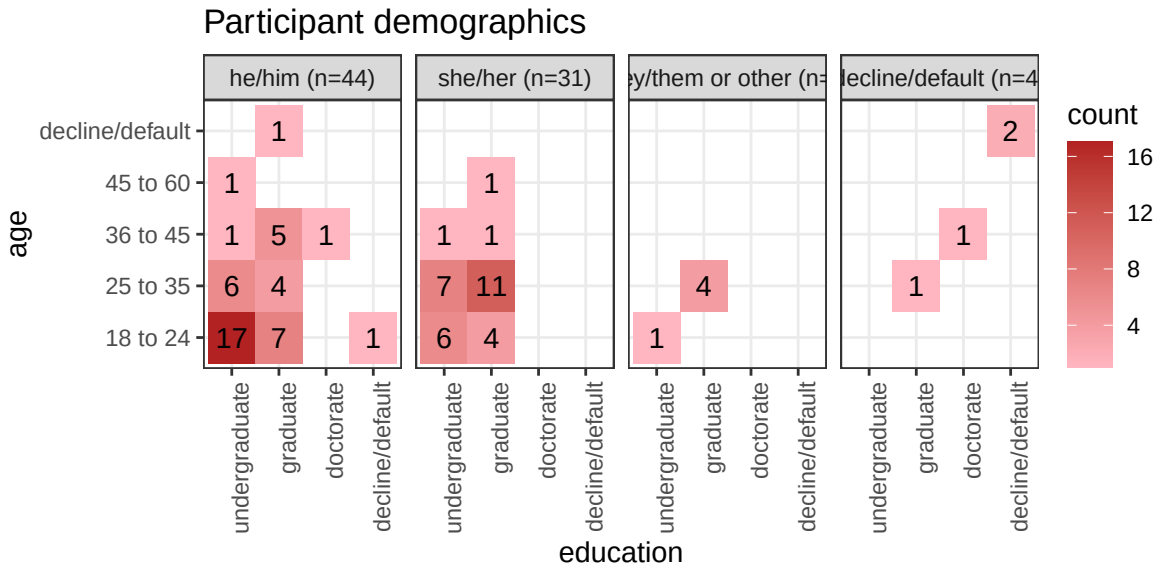

Figure 5: Heatmaps of participant demographics, counts of age group by completed education as faceted across preferred pronoun. The average participant was a young adult with an undergraduate or graduate degree.

Demographic:
- What are your preferred pronouns? [decline to answer, he/him, she/her, they/them or other] - Which age group do you belong to? [decline to answer, 18 to 24, 25 to 35, 36 to 45, 45 to 60, 60 and up] - What is your highest completed education? [decline to answer, Undergraduate degree (BA/BSc/other), Graduate degree (MA/MSc/MPhil/other), Doctorate degree (PhD/other); prolific.co participants were filtered to those stating they had at least an undergraduate degree]

Subjective by factor:
- I was already familiar with visualization. - I found this visualization easy to use. - I felt confident in my answers with this visualization. - I liked using this visualization.

3.6. Sampling population

We recruited $N = 108$ via prolific.co (Palan and Schitter 2018). We make the assumption that interpretation of biplot displays used will not be commonly used for consumption by the general population and apply a single filter on education; that participants have completed at least an undergraduate degree (some 58,700 of the 150,400 users at the time). There is also the implicit filter that Prolific participants must be at least 18 years of age. Participants were compensated for their time at £7.50 per hour, whereas the mean duration of the survey was about 16 minutes. We can't preclude previous knowledge or experience with the factors, but instead, try to control for this in the user study. Figure 5 shows distributions of age and preferred pronouns of the participants that completed the post-study survey who are relatively young and well educated.

3.7. Evenness of block evaluation

From pilot studies through a sample of convenience (consisting of information technology and statistics Ph.D. students attending Monash University) we predict that we wanted 3 even block evaluations to support differences in our factor and block parameterizations. Given that factor and location each have 6 permutations we targeted $N = 108 = 3 * (6 * 6)$ evaluations before data were collected. In data collection, we experienced several adverse conditions, primarily: limited control of application server network configuration, throughput thresholds on data read/write API, and repeat attempts from users when experiencing disconnects. To mitigate this we over-collect survey trials, exclude all partial trials, and remove the oldest attempts (mostly likely to experience adverse network conditions) from over evaluated permutations until we have our desired evaluations under each permutation.

3.8. Data capture and processing

Data was recorded by a **shiny** application and was written to a Google Sheet at the completion of each study period. Time is measured as the CPU time of the server. This could be made slightly more accurate by calculating the difference of system time at these points. Because the frequency of the participants from prolific was more then the server would handle, participants were manually accepted one or two at a time. We collected participants till we had 3 even evaluations. The processing steps were fairly minimal after formatting to tidy format and decoding values to a more human-readable state. After formatting, a flag is added to indicate if the survey had data from all 3 periods. We remove all partial studies and keep only the last 3 complete studies within each block parameterization, the studies which should have experienced the least adverse network conditions. This brings us to the 108 studies described in the paper, from which models and aggregation tables were built. The post-study surveys were similarly decoded to human-readable tidy format and then filtered to include only those 84 surveys that were associated with the final 108 studies.

The code, response files, their analyses, and the study application are publicly available at on GitHub github.com/nspyrison/spinifex_study.

4. Results

To recap, the primary response variable is task marks as defined in section 3.2, and the log of response time will be used as a secondary response variable. We have 2 primary data sets; the user study evaluations and post-study survey. The former is contains the 108 trials with explanatory variables: *factor*, *location* of the cluster separation signal, the *shape* of variance-covariance matrix, and the *dim*-ensionality of the data. Block parameterization and randomization was discussed in section 3.4. The survey was completed for 84 of these 108 trials and contains demographic information (preferred pronoun, age, and education), and subjective measures for each of the factors (*preference*, *familiarity*, *ease of use*, and *confidence*). The survey was covered in more detail in 3.5.

Below we look at the marginal performance of the block parameters and survey responses. After that, we build a battery of regression models to explore the variables and their interactions. Lastly, we look at the subjective measures between the factors.

Table 1: Model comparison of our random effect models regressing marks. Each model includes a random effect term of the participant, which explains the individual's influence on their marks. Complex models perform better in terms of R2 and RMSE, yet AIC and BIC penalize their large number of fixed effect in favor of the much simpler model containing only factor.

Fixed effects	No. levels	No. terms	AIC	BIC	R2 cond. (on RE)	R2 marg. (w/o RE)	RMSE
a	1	3	*1000*	*1027*	0.18	0.022	0.462
a+b+c+d	4	8	1026	1075	0.187	0.03	0.46
a*b+c+d	5	12	1036	1103	0.198	0.043	0.457
a*b*c+d	8	28	1069	1207	0.238	0.08	0.447
a*b*c*d	15	54	1125	1380	*0.255*	*0.115*	*0.438*

4.1. Random effect regression against marks

To more thoroughly examine explanatory variables we regress against marks. All models have a random effect term on the participant, which captures the effect of the individual participant. After we look at models of the block parameters we extend to compare against survey variables. Last, we compare how adding a random effect for data and regressing against time till last response fares against benchmark models. The matrices for models with more than a few terms quickly become rank deficient; there is not enough information in the data to explain all of the effect terms. In which case the least impactful terms are dropped.

In building a set of models to test we include all single term models, a model with all independent terms. We also include an interaction term of factor by location, allowing for the slope of each location to change across each level of the factor, which is feasible. For comparison, an overly complex model with many interaction terms is included.

Fixed effects:	**Expanded Mode:**
α	$\widehat{marks} = \mu + \alpha_i + \mathbf{Z} + \mathbf{W} + \epsilon$
$\alpha + \beta + \gamma + \delta$	$\widehat{marks} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \mathbf{Z} + \mathbf{W} + \epsilon$
$\alpha * \beta + \gamma + \delta$	$\widehat{marks} = \mu + \alpha_i * \beta_j + \mathbf{Z} + \mathbf{W} + \epsilon$
$\alpha * \beta * \gamma + \delta$	$\widehat{marks} = \mu + \alpha_i * \beta_j * \gamma_k + \mathbf{Z} + \mathbf{W} + \epsilon$
$\alpha * \beta * \gamma * \delta$	$\widehat{marks} = \mu + \alpha_i * \beta_j * \gamma_k * \delta_l + \mathbf{Z} + \mathbf{W} + \epsilon$

where μ is the intercept of the model including the mean of random effect
$\epsilon \sim \mathcal{N}(0,\ \sigma)$, the error of the model
$\mathbf{Z} \sim \mathcal{N}(0,\ \tau)$, the random effect of participant
$\mathbf{W} \sim \mathcal{N}(0,\ \upsilon)$, the random effect of simulation
α_i, fixed term for factor | $i \in$ (pca, grand, radial)
β_j, fixed term for location | $j \in$ (0_1, 33_66, 50_50) % noise/signal mixing
γ_k, fixed term for shape | $k \in$ (EEE, EEV, EVV banana) model shapes
δ_l, fixed term for dim | $l \in$ (4 variables & 3 cluster, 6 variables & 4 clusters)

Residual plots have no noticeable non-linear trends and contain striped patterns as an artifact from regressing on discrete variables. Figure 6 illustrates (T) the effect size

Table 2: The model coefficients for $\widehat{marks} = \alpha * \beta + \gamma + \delta$, with factor=pca, location=0/100, and shape=EEE held as baselines. Factor being radial is the fixed term with the strongest evidence in support of the hypothesis. When crossing factor with location radial performs worse with 33/66 percent mixing relative to the PCA with no mixing. The model fit is based on the 648 evaluations by the 108 participants.

	Estimate	Std. Error	df	t value	Pr(>|t|)	
(Intercept)	-0.12	0.08	43.9	-1.50	0.14	
factor						
fct=grand	0.15	0.09	622.4	1.74	0.08	
fct=radial	0.37	0.09	617.1	4.18	0.00	***
fixed effects						
loc=33_66	0.17	0.09	83.2	1.78	0.08	
loc=50_50	0.14	0.09	84.8	1.52	0.13	
shp=EEV	0.04	0.06	11.5	0.79	0.44	
shp=ban	-0.03	0.06	11.5	-0.48	0.64	
dim=6	-0.06	0.05	11.5	-1.39	0.19	
interactions						
fct=grand:loc=33_66	-0.06	0.13	587.3	-0.49	0.63	
fct=radial:loc=33_66	-0.34	0.13	585.2	-2.65	0.01	**
fct=grand:loc=50_50	-0.09	0.13	589.6	-0.68	0.50	
fct=radial:loc=50_50	-0.19	0.13	574.3	-1.43	0.15	

of the random terms participant and simulation, or more accurately, the 95% CI from Gelman simulation of their posterior distribution. The effect size of the participant is much larger than simulation. The most extreme participants are statistically significant at $\alpha = .95$, while none of the simulation effects significantly deviate from the null of having no effect size on the marks. In comparison, (B) 95% confidence intervals of the mean marks for participation and simulation respectively.

We also want to visually explore the conditional variables in the model. Figure 7 explores violin plots of marks by factor while faceting on the location (vertical) and shape (horizontal). Radial tends to increase the marks received, and especially so when there is no signal/noise mixing.

4.2. Time regressing models

As a secondary explanatory variable, we also want to look at time. First, we take the log transformation of time as it is right-skewed. Now we repeat the same modeling procedure, namely: 1) build a battery of all additive and multiplicative models. 2) Compare their performance, reporting some top performers. 3) Select a model to examine its coefficients.

4.3. Subjective measures

The 84 evaluations of the post-study survey also collect 4 subjective measures for each factor. Figure 9 shows the Likert plots, or stacked percentage bar plots, alongside violin plots with the same non-parametric, ranked sum tests previously used. Participants

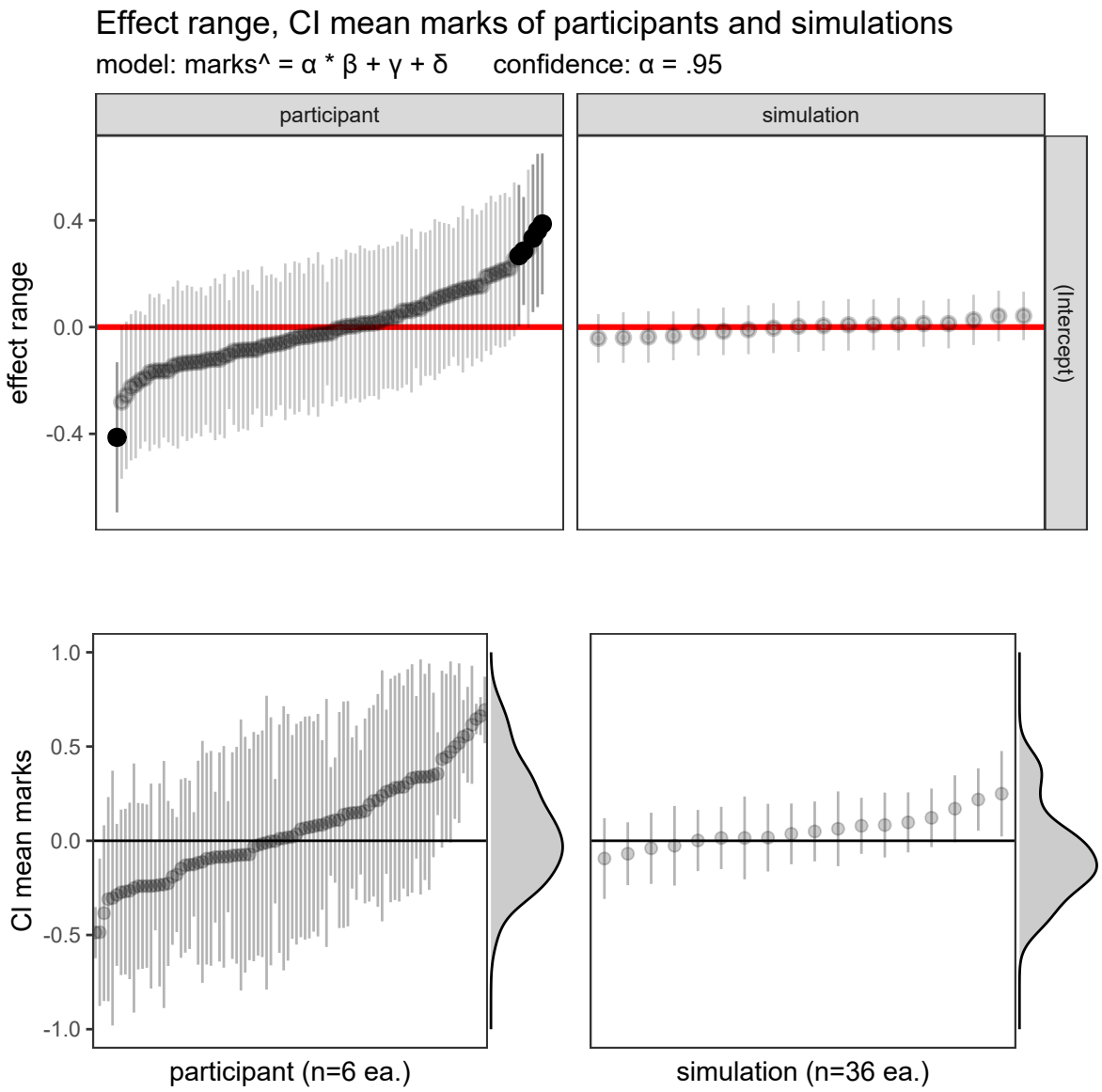

Figure 6: (T) Estimated effect ranges of the random effect terms participant and data simulation of the $\alpha * \beta + \gamma + \delta$ model. Confidence intervals are created with Gelman simulation on the effect posterior distributions. The effect size of the participant is relatively large, with several significant extrema. None of the simulations deviate significantly. (B) The ordered distributions of the CI of mean marks follow the same general pattern and give the additional context of how much variation is in the data, an upper limit to the effect range. The effect ranges capture about 2/3 of the range of the data without the model. All intervals for $\alpha = .95$ confidence.

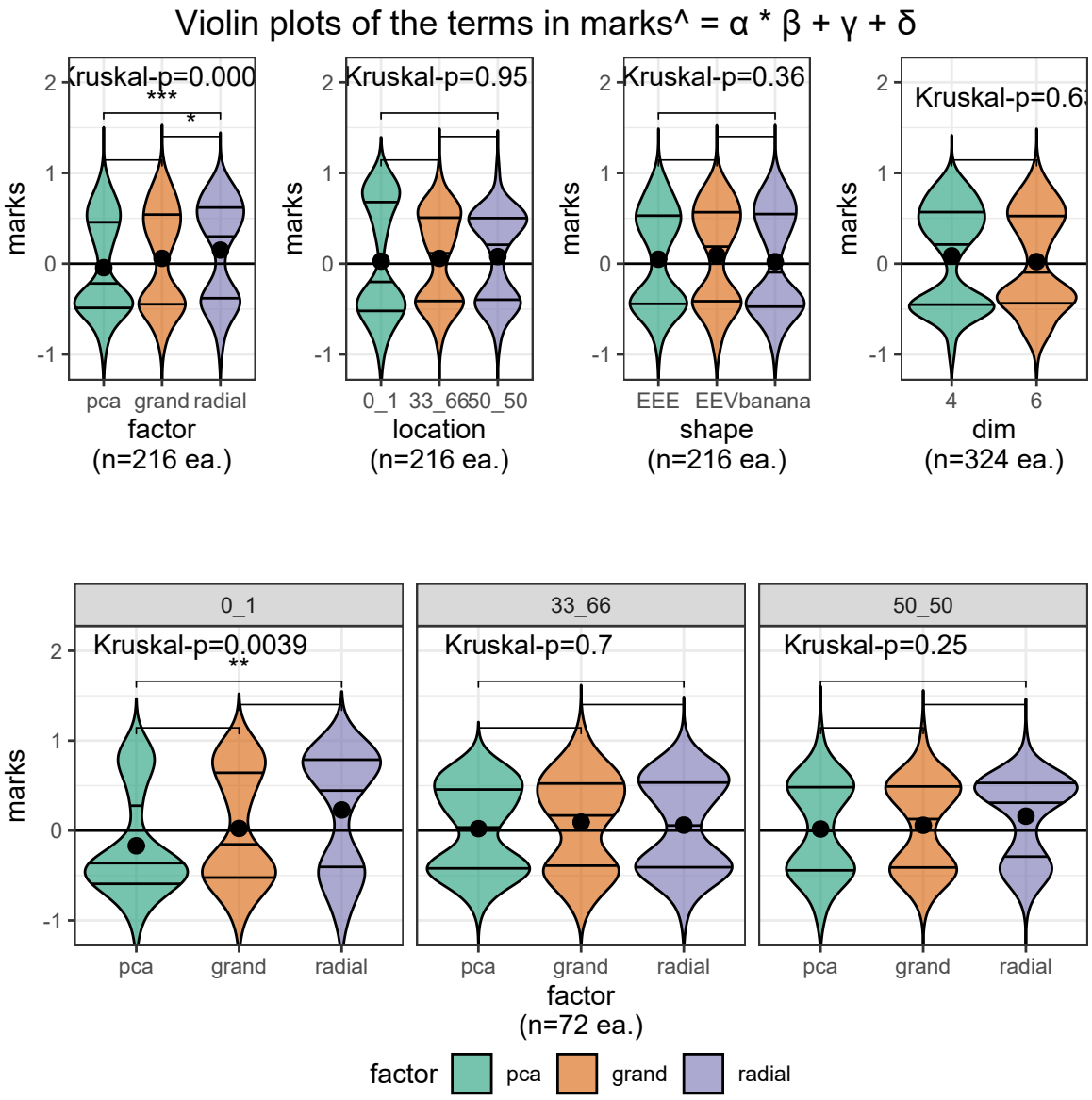

Figure 7: Violin plots of terms of the model $\widehat{marks} = \alpha * \beta + \gamma + \delta$. Overlaid with global significance from the Kruskal-Wallis test, and pairwise significance from the Wilcoxon test, both are non-parametric, ranked sum tests suitable for handling discrete data. Participants are more confident and find the radial easier to use relative to the grand tour. Participants claim low familiarity as we expect from crowdsourced participants. Radial is more preferred compared with either alternative for this task.

Table 3: Model comparisons for $log(\widehat{time})$ random effect models, where each model includes random effect terms for participants and simulations. We see the same trade-off where the simplest factor model is perferred by AIC/BIC, while R2 and RMSE perform best with the full multiplicative model. We again select the model $\alpha * \beta + \gamma + \delta$ to explore further as it has relatively high marginal R2 while having much less complexity than the full model.

Fixed effects	No. levels	No. terms	AIC	BIC	R2 cond. (on RE)	R2 marg. (w/o RE)	RMSE
a	1	3	*1000*	*1027*	0.18	0.022	0.462
a+b+c+d	4	8	1026	1075	0.187	0.03	0.46
a*b+c+d	5	12	1036	1103	0.198	0.043	0.457
a*b*c+d	8	28	1069	1207	0.238	0.08	0.447
a*b*c*d	15	54	1125	1380	*0.255*	*0.115*	*0.438*

Table 4: The model coeffients for $log(\widehat{time}) = \alpha * \beta + \gamma + \delta$, with factor:pca, locaation0/100, and shapeEEE held as baselines. location50/50 is the fixed term with the strongest evidence and takes less time. In contrast, the interaction term location=50/50:shape=EEV has the most evidence and takes much longer on average.

	Estimate	Std. Error	df	t value	Pr(>|t|)	
(Intercept)	2.71	0.14	42.6	19.06	0.00	***
factor						
fct=grand	-0.23	0.12	567.6	-1.97	0.05	*
fct=radial	0.16	0.12	573.5	1.34	0.18	
fixed effects						
loc=33_66	0.05	0.14	40.9	0.34	0.74	
loc=50_50	-0.05	0.14	42.1	-0.35	0.73	
shp=EEV	-0.15	0.09	8.3	-1.61	0.14	
shp=ban	-0.13	0.09	8.3	-1.42	0.19	
dim=6	0.14	0.08	8.3	1.90	0.09	
interactions						
fct=grand:loc=33_66	0.24	0.18	580.9	1.34	0.18	
fct=radial:loc=33_66	-0.24	0.18	582.4	-1.32	0.19	
fct=grand:loc=50_50	0.12	0.18	578.6	0.69	0.49	
fct=radial:loc=50_50	0.05	0.18	584.4	0.25	0.80	

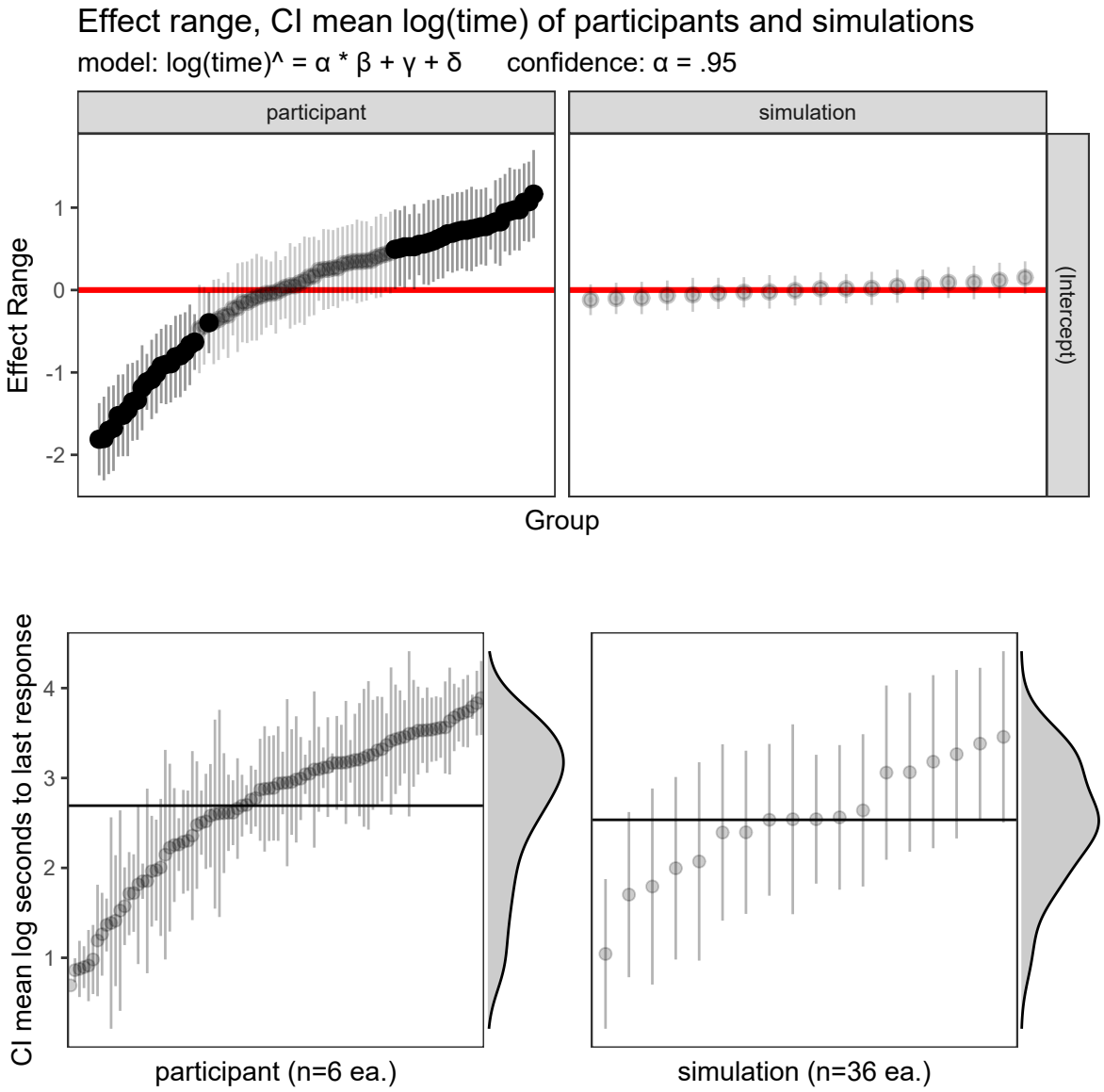

Figure 8: (T) The effect ranges of Gelman resimulation on posterior distributions. These show the magnitude and distributions of particular participants and simulations. Simulation has a relatively small effect on response time. (B) Confidence intervals for mean log(time) by participant and simulation. The marginal density shows that the response times are left-skewed after log transformation. Interpreting back to linear time there is quite the spread of response times: $e^1 = 2.7$, $e^{2.75} = 15.6$, $e^{3.75} = 42.5$ seconds. Considering simulations, on the right, the bottom has a large variation of time, relative to the effect ranges which means that the variation actually is explained in the terms of the model and not by the simulation itself.

preferred to use radial for this task. Participants were also more confident of their answers and found radial tours easier to use compared with the grand tour. All factors have reportedly low familiarity something we expect from crowdsourced participants.

5. Conclusion

Above we conducted a with-in participant user study comparing the efficacy of 3 linear projection techniques. The participants performed a supervised cluster task, specifically the identification of which variables contribute to the separation between 2 target clusters. In summary, we that use of the radial tours increases accuracy while the use of the grand tour decreases the time it takes to perform this task. These effects are large relative to the other block parameterizations, but smaller than the random effect of the participant. Radial tour was subjectively most preferred, leads to more confidence in answers, and is easier to use than alternatives.

There are a number of ways that this study could be extended. In addition to expanding the support of the block parameterizations, more interesting directions include: type of task, factors used, and experience level of the target population. It is difficult to achieve good coverage given the number of possible permutations. Be sure to step back and plan what the target support, visuals, and task are. Keep in mind the volume and quality of responses from participants especially when crowdsourcing. These planning steps are useful for navigating when the complexity of the application details.

6. Accompanying tool: spinifex application

To accompany this study we have produced a more general use tool to perform such exploratory analysis of high dimensional data. The R package, **spinifex**, (Spyrison and Cook 2020) contains a free, open-source **shiny** (Chang et al. 2020) application. The application allows users to upload, process, and interactively explore their data. Users can quickly traverse global and local extrema and then explore the nearby space with the radial tour as similarly applied in the user study. Limited implementations of grand, little, and local tours are also made available. Data can be imported in .csv and .rda format, and projections or animations can be saved as .png, .gif, and .csv formats where applicable. Run the following R code for help getting started.

7. Acknowledgments

This research was supported by an Australian Government Research Training Program (RTP) Scholarship. This article was created in R (R Core Team 2020) and **rmarkdown** (Xie, Allaire, and Grolemund 2018). Visuals were prepared with **spinifex**. All packages used are available from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/`. The source files for this article, application, data, and analysis can be found at github.com/nspyrison/spinifex_study/. The source code for the **spinifex** package and accompanying shiny application can be found at github.com/nspyrison/spinifex/.

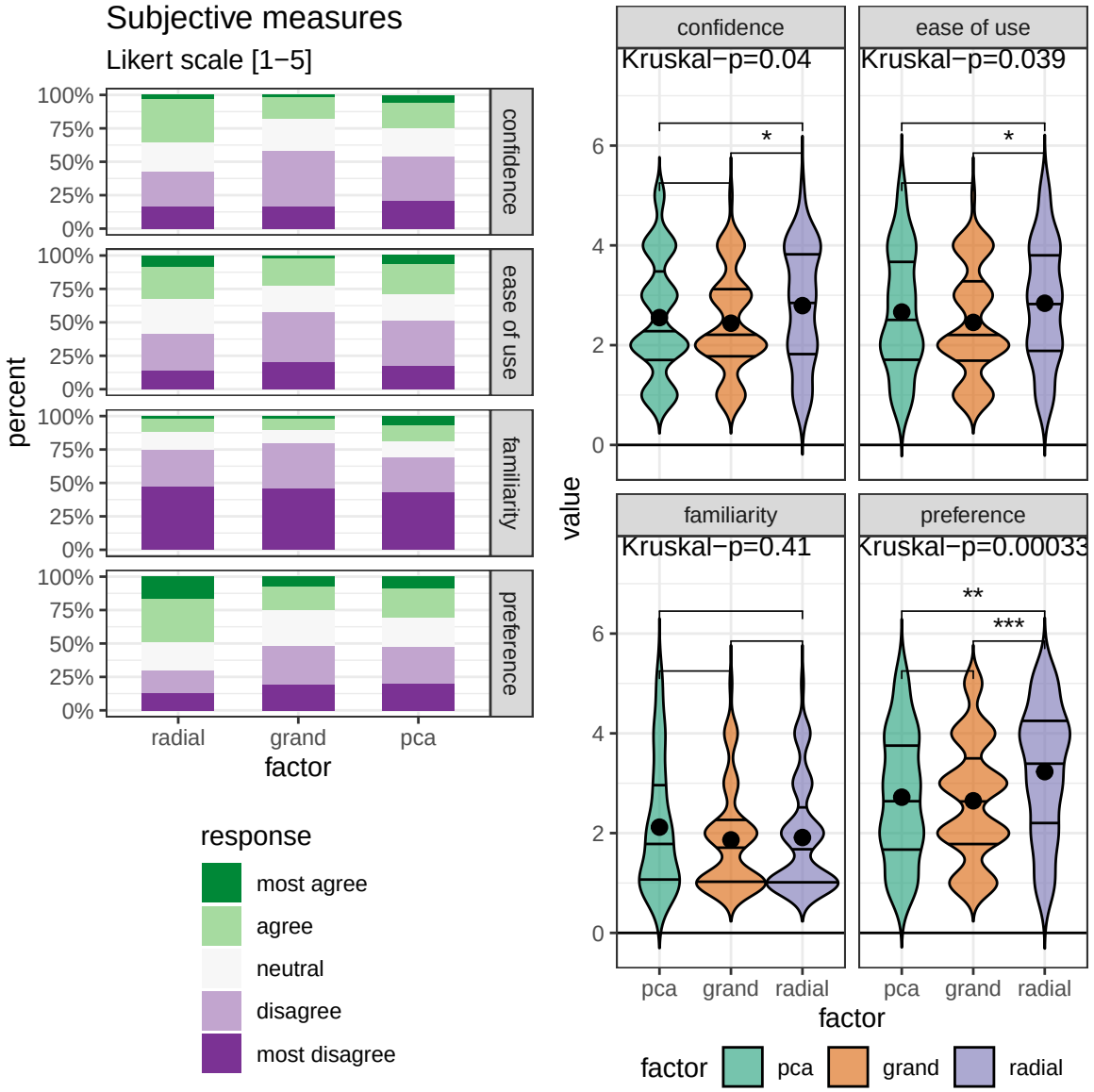

Figure 9: The subjective measures of the 84 responses of the post-study survey, 5 discrete Likert scale levels of aggrement. (L) Likert plots (stacked percent bar plots) with (R) violin plots of the same measures. Violin plots are overlaid with global significance from the Kruskal-Wallis test, and pairwise significance from the Wilcoxon test, both are non-parametric, ranked sum tests.

8. Bibliography

10 Anscombe, F. J. 1973. "Graphs in Statistical Analysis." *The American Statistician* 27 (1): 17–21. https://doi.org/10.2307/2682899.

Asimov, Daniel. 1985. "The Grand Tour: A Tool for Viewing Multidimensional Data." *SIAM Journal on Scientific and Statistical Computing* 6 (1): 128–43. https://doi.org/https://doi.org/10.1137/0906011.

Bellman, Richard Ernest. 1957. *Dynamic Programming*. Princeton University Press. https://books.google.com.au/books?id=fyVtp3EMxasC&redir_esc=y.

Bertini, Enrico, Andrada Tatu, and Daniel Keim. 2011. "Quality Metrics in High-Dimensional Data Visualization: An Overview and Systematization." *IEEE Transactions on Visualization and Computer Graphics* 17 (12): 2203–12.

Chang, Winston, Joe Cheng, J. J. Allaire, Yihui Xie, and Jonathan McPherson. 2020. *Shiny: Web Application Framework for R*. https://CRAN.R-project.org/package=shiny.

Coleman, David. 1986. "Geometric Features of Pollen Grains." In. Statistical Computing Statistical Graphics. http://stat-computing.org/dataexpo/1986.html.

Cook, Dianne, and Andreas Buja. 1997. "Manual Controls for High-Dimensional Data Projections." *Journal of Computational and Graphical Statistics* 6 (4): 464–80. https://doi.org/10.2307/1390747.

Goodman, Steven. 2008. "A Dirty Dozen: Twelve P-Value Misconceptions." *Seminars in Hematology*, Interpretation of Quantitative Research, 45 (3): 135–40. https://doi.org/10.1053/j.seminhematol.2008.04.003.

Gracia, Antonio, Santiago González, Víctor Robles, Ernestina Menasalvas, and Tatiana von Landesberger. 2016. "New Insights into the Suitability of the Third Dimension for Visualizing Multivariate/Multidimensional Data: A Study Based on Loss of Quality Quantification." *Information Visualization* 15 (1): 3–30. https://doi.org/10.1177/1473871614556393.

Karwowski, Waldemar. 2006. *International Encyclopedia of Ergonomics and Human Factors, -3 Volume Set*. CRC Press.

Liu, Shusen, Dan Maljovec, Bei Wang, Peer-Timo Bremer, and Valerio Pascucci. 2017. "Visualizing High-Dimensional Data: Advances in the Past Decade." *IEEE Transactions on Visualization and Computer Graphics* 23 (3): 1249–68. https://doi.org/10.1109/TVCG.2016.2640960.

Maaten, Laurens van der, and Geoffrey Hinton. 2008. "Visualizing Data Using t-SNE." *Journal of Machine Learning Research* 9 (Nov): 2579–2605.

Matejka, Justin, and George Fitzmaurice. 2017. "Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics Through Simulated Annealing." In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*, 1290–94. Denver, Colorado, USA: ACM Press. https://doi.org/10.1145/3025453.3025912.

Palan, Stefan, and Christian Schitter. 2018. "Prolific. Ac—A Subject Pool for Online Experiments." *Journal of Behavioral and Experimental Finance* 17: 22–27.

Pearson, Karl. 1901. "LIII. On Lines and Planes of Closest Fit to Systems of Points in Space." *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2 (11): 559–72.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Scrucca, Luca, Michael Fop, T. Brendan Murphy, and Adrian E. Raftery. 2016. "Mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models." *The R Journal* 8 (1): 289–317. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5096736/.

Sedlmair, Michael, Tamara Munzner, and Melanie Tory. 2013. "Empirical Guidance on Scatterplot and Dimension Reduction Technique Choices." *IEEE Transactions on Visualization & Computer Graphics*, no. 12: 2634–43.

Spyrison, Nicholas, and Dianne Cook. 2020. "Spinifex: An R Package for Creating a Manual Tour of Low-Dimensional Projections of Multivariate Data." *The R Journal* 12 (1): 243. https://doi.org/10.32614/RJ-2020-027.

Tukey, John W. 1977. *Exploratory Data Analysis*. Vol. 32. Pearson.

Wagner Filho, Jorge, Marina Rey, Carla Freitas, and Luciana Nedel. 2018. "Immersive Visualization of Abstract Information: An Evaluation on Dimensionally-Reduced Data Scatterplots." In.

Xie, Yihui, J. J. Allaire, and Garrett Grolemund. 2018. *R Markdown: The Definitive Guide*. Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown.

Affiliation:

Nicholas Spyrison
Monash University
Faculty of Information Technology,
Monash University
E-mail: nicholas.spyrison@monash.edu

Dianne Cook
Monash University
Department of Econometrics & Business Statistics,
Monash University

Kimbal Marriott
Monash University
Faculty of Information Technology,
Monash University

Journal of Data Science, Statistics, and Visualisation https://jdssv.org/
published by the International Association for Statistical Computing
http://iasc-isi.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:XX.XXXXX/jdssv.v000.i00 *Accepted:* yyyy-mm-dd