Journal of Data Science, Statistics, and Visualisation
MMMMMMM YYYY, Volume VV, Issue II. *doi: XX.XXXXX/jdssv.v000.i00*

A study examining the benefit of the user-controlled radial tour for understanding variable contributions to structure visible in linear projections of high-dimensional data

Nicholas Spyrison
Monash University

Dianne Cook
Monash University

Kimbal Marriott
Monash University

Abstract

Principal component analysis is a long-standing go-to method for exploring multivariate data. Data visualization *tours* are a class of linear projections animated over small changes in the projection basis. The radial tour is one instance that rotates the contribution of a select variable. Does this variable-level control help an analyst explore multivariate data? This paper describes a user study evaluating the efficacy the use of the radial tour in comparison to two existing methods, principal component analysis, and the grand tour. We devise a supervised classification task where participants evaluate variable attribution of the separation between two classes. Accuracy and response time are measured as response variables. Data were collected from 108 crowd-sourced participants, who perform 2 trials of each visual for a total of 648 trials. There is strong evidence that the radial tour increases accuracy. Participants also preferred to use the radial tour over alternatives for this task.

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

1. Introduction

Multivariate data is ubiquitous. Exploratory data analysis (EDA, Tukey 1977) is important for understanding the data and testing model assumptions. Data visualization is more robust and informative than statistical summarization alone (Anscombe 1973; Matejka and Fitzmaurice 2017). Focusing on hypothesis testing can result in tunnel vision of an analyst causing them to miss visual particularities in the data (Yanai and

Lercher 2020). We know that Data visualization is important for EDA and our comprehension of the data. But how do we know which methods to use to best visualize data?

Models are becoming increasingly complex models involving many features and terms causing an opaqueness to their interpretability. Exploratory artificial intelligence (XAI) is a new branch (Adadi and Berrada 2018; Arrieta et al. 2020) tries to make black-box models more transparent, by creating techniques to increase interpretability. Multivariate data visualization is a similarly important part of exploring features spaces and communicating interpretations of models (Biecek 2018; Biecek and Burzykowski 2021; Wickham, Cook, and Hofmann 2015).

Dimension reduction is commonly used in conjunction with visualization to provide informative low-dimensional summaries of high-dimensional data. There have been several user studies for dimension reduction comparing across embeddings and display dimensionality (Gracia et al. 2016; Wagner Filho et al. 2018). There are also 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 an absence of studies comparing techniques for assessing variable attribution across visualization methods.

This paper describes a user study conducted to assess the efficacy of different visualization methods. Cluster data is simulated under several additional experimental factors, namely: location, shape, and dimensionality. We task participants with identifying the variables attributing to separation between two clusters. We define an accuracy measure for this task and use response time as a response variable of secondary interest.

The paper is structured as follows. Section 2 discusses several visualization methods including the ones compared in the study. Section 3 describes the experimental factors, tasks, and evaluation measures used. The results of the study are discussed in Section 4. Conclusions and potential future directions are discussed in Section 5. The software used for the study is described in Section 6.

2. Background

We'll discuss common multivariate techniques including the methods we apply in the user study.

2.1. Scatterplot matrix

One could consider looking a p histograms or univariate densities. This will miss features in two or more dimensions. Similarly, all combinations of pairs of variables can be viewed as scatterplot matrix (Chambers et al. 1983). Figure 1 shows the first three components of simulated data as a scatterplot matrix. Looking at p univariate densities or many bivariate scatter plots at once quickly becomes burdensome as dimensionality increases, and only displays information containing in 2 orthogonal dimensions, that is features that require in 3 dimensions will never be resolved.

2.2. Parallel coordinates plot

Another common way to display multivariate data is with a parallel coordinates plot (Ocagne 1885), which displays observations by their quantile values for each variable with connected by lines to the quantile value in subsequent variables.

Parallel coordinates plot and other observations-based visuals such as pixel plots or Chernoff faces scales well with dimensions and but poorly with observations. These are perhaps best used when there are more variables than observations.

Observations-based visuals have a couple of issues. One is that they are asymmetric across variable ordering which can lead to different conclusions or features being focused on due to variable order. Another notable issue of observations-based visuals is the graphical channel used to convey information. Munzner suggests that position is the visual channel that is most perceptible by human perception (Munzner 2014). In the case of parallel coordinates plot the horizontal axes spans variables rather than the values of one variable. This means that we lose an axis to link information on our most discerning visual channel.

2.3. Principal component analysis

Principal component analysis is a good baseline of comparison for linear projections because of its frequent and broad use across disciplines. 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 2 or 3 components are typically used to approximate the variation multivariate data set, while the rest are discarded.

2.4. Animated linear projections, tours

A data visualization *tour* animates many linear projections over small changes in the projection basis. 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 that are distinguished by the selection of their basis paths (Lee et al. 2021; Cook et al. 2008). To contrast with the discrete orientations of PCA, we compare with continuous changes of linear projection with *grand* and *radial* tours.

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.

Radial and manual tours

The *manual* tour (Cook and Buja 1997) defines its basis path by manipulating the basis contribution of a selected variable. A manipulation dimension is appended onto

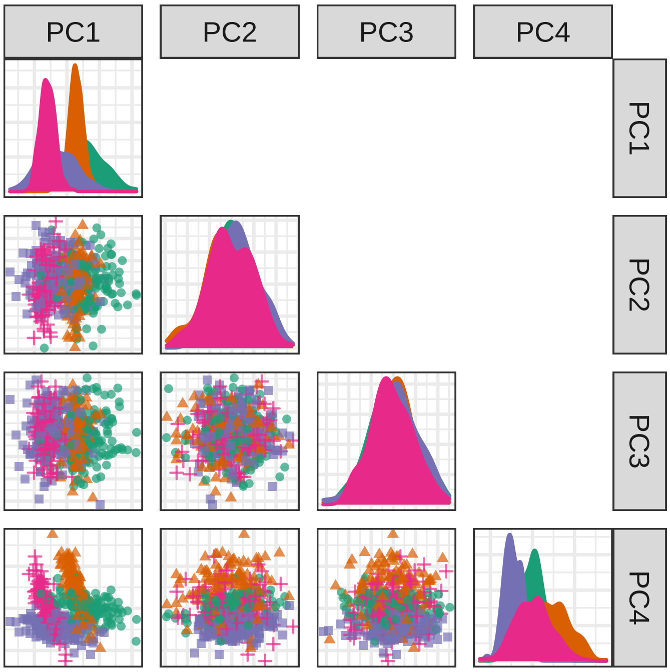


Figure 1: Scatterplot matrix of the first four principal components for simulated data. An analyst would have to view PC1 by PC2 and PC1 by PC4 be have an accuracy take on which variables attribute to the separation between clusters.

the projection plane, with a full contribution given to the selected variable. The target bases are then selected based on rotating this newly created manipulation space. The target bases are then similarly orthogonally restrained, data projected and rendered into an animation. 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 selected 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 the user to steer via choosing the variable to rotate.

The recent implementation of manual tours us the R package **spinifex** (Spyrison and Cook 2020), which facilitates manual tours (and radial variant). It is also compatible with tours made with **tourr** (Wickham et al. 2011) and facilitates exporting to .gif or .html widget, with recent graphic packages. Now that we have a readily available means to produce various tours, we want to see how they fare against traditional discrete displays commonly used with PCA.

3. User study

An experiment was constructed to assess the performance of the radial tour relative to the tour and scatterplots of principal components for interpreting the importance of variables to class separations.

The three methods were examined for three different cluster shapes, using different combinations of contributing variables, and data dimensionality. Data was collected using a specially constructed web app, through crowd-sourced with prolific.co (Palan and Schitter 2018).

3.1. Experimental factors

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 vary the number of variables and their magnitude of cluster separation, 2) the *shape* of the clusters, to reflect different distributions of the data, and 3) the *dimension*-ality of the data. Below we describe the levels within each factor, while Figure 2 gives a visual representation.

The *location* of the separation of the clusters is a crucial aspect of analysis, it is the variables or combination their of that is important to the explanation of the structure. 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 percentages: 0/100% (not mixed), 33/66%, 50/50% (evenly mixed).

In selecting the *shape* of the clusters we follow the convention given by Scrucca et al. (2016), where 14 variants of model families containing 3 clusters are defined. The name of the model family is the abbreviation of its respective volume, shape, and orientation of the cluster, 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.

Dimension-ality is tested at 2 modest levels, namely, in 4 dimensions containing 3 clusters and 6 dimensions with 4 clusters. We must do so to bound the difficulty and search space to keep the task realistic for crowdsourcing.

3.2. Objective

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, with the persistence of the data points across changes in basis, 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. Next, we cover how we expect them to perform and state the hypothesis to test.

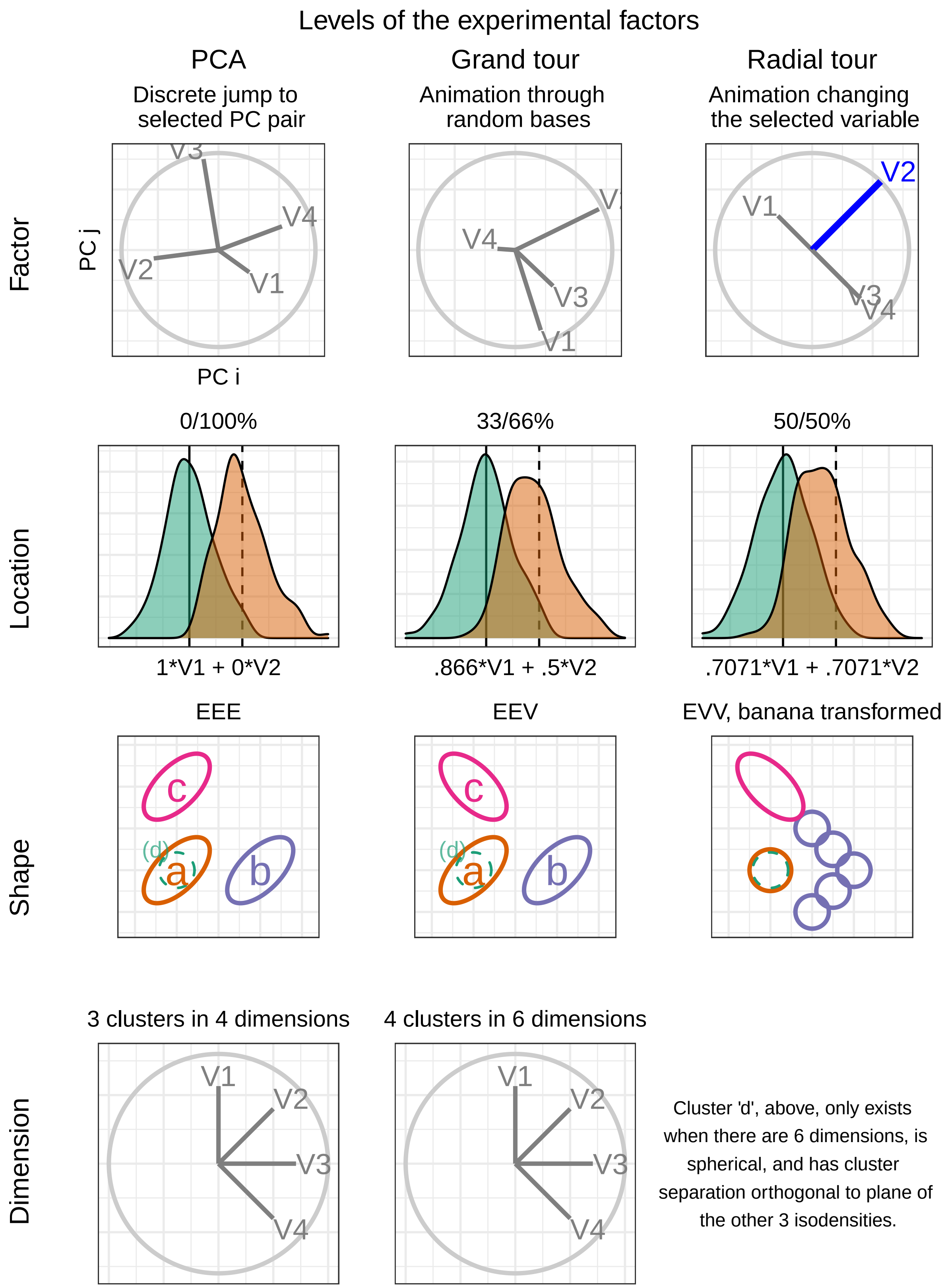


Figure 2: Illustration of the experimental factors, the parameter space of the independent variables, the mathematical support of our study.

Then for some subset of tasks, we expect to find that the radial tour performs most accurately, as it enjoys both the persistence of points and input control to explore specific variables. Secondly, it may be the case that grand performs faster than the alternatives with its absence of inputs, users can focus all of their attention on interpreting the fixed path. Conversely, we are less certain about the accuracy of such a limited grand tours as there is no objective function in the selection of the bases; it is possible that, by chance, the planes completely avoid the information needed. However, given that the data dimensionality will be modest, it seems likely that grand tour regularly crosses frames with the correct information to perform the task quickly.

We measure the accuracy and response time over the support of the discussed experimental factors. The null hypotheses can be stated as:

$$\begin{aligned}
H_0 &: y_1, \text{task accuracy does not vary with the visualization method} \\
H_\alpha &: y_1, \text{task accuracy does vary with thevisualization method} \\
H_0 &: y_2, \text{task response time does not vary with the visualization method} \\
H_\alpha &:, y_2 \text{task response time does vary with the visualization method}
\end{aligned}$$

3.3. Task and evaluation

With our hypothesis formulated let's turn our attention to the task and how to evaluate it. Regardless of visual method the elements of the display are held constant a 2D scatterplot with axis biplot to its left. Observations were supervised with the cluster level mapped to color and shape.

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 carries 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. This restriction did not impact many participants as the .25, .50, .75-quantiles of response time were about 7, 21 and 30 seconds respectively.

The evaluation measure of this task was designed with a couple of features in mind: 1) the sum of squares of the individual variable weights should be 1, and 2) symmetric about 0, that is, without preference to under- or over-guessing. 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.

$$\begin{aligned}
&\mathbf{X}_{np} = (x_1, \ ... \ x_p) \\
&\mathbf{X}_{npk} \text{ is an observation subset of } \mathbf{X} \text{ belonging to cluster } k
\end{aligned}$$

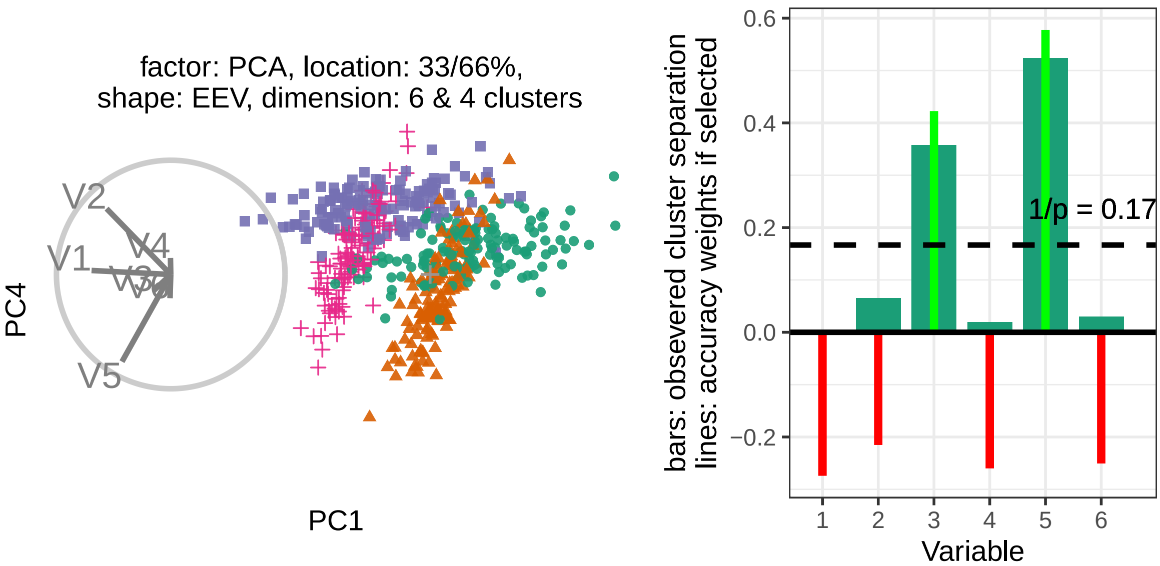


Figure 3: (L), PCA biplot of the components showing the most cluster separation with (R) illustration of the magnitude of cluster separation is for each variable (bars) and the weight of the variable accuracy if selected (red/green lines). The horizontal dashed line is 1 / p, the amount of separation each variable would have if evenly distributed. The weights equal the signed square of the difference between each variable value and the dashed line.

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 in the form of a boolean value for each variable, whether or not the participant thinks each variable separates the two clusters more than if the difference were uniformly distributed. Then Y_1 is a vector of variable accuracy.

$$W_j = \frac{(\overline{X_{j=1,k=1}} - \overline{X_{1,k=2}},\ \ldots\ (\overline{X_{j=p,k=1}} - \overline{X_{j=p,k=2}})}{\sum_{j=1}^{p}(|\overline{X_{j,k=1}} - \overline{X_{j,k=2}}|)} - \frac{1}{p}$$

$$\text{Accuracy}, Y = \sum_{j=1}^{p} I(r_j) * sign(w_j) * \sqrt{|w_j|}$$

where I is the indicator function. Then the task accuracy is the sum of this vector. We use the time till the last response as a secondary dependent variable Y_2.

3.4. Visual design standardization

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 is controlled with the block assignment as illustrated below in Figure 4. Below we cover the visual design standardization, as well the input and display within each factor.

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 of axis titles) were held constant. 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.

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.

3.5. Data simulation

Each dimension is originally distributed as $\mathcal{N}(2 * I(signal), 1) \mid$ covariance Σ, a function of the shape. 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. Each cluster is simulated with 140 observations and is offset in a variable that does not separate previous variables.

Clusters of the EVV shape are transformed to the banana-chevron shape. Then location mixing is applied by post-multiplying a (2x2) rotation matrix to the signal variable and a noise variable for the clusters in question. All variables are then standardized by standard deviation. The rows and columns are then shuffled randomly. The observation's cluster and order of shuffling are attached to the data and saved.

Each of these replications is then iterated with each level of the factor. For PCA, every pair of the top 4 principal components and saved as 12 plots. For the grand tour, we first save 2 basis paths (for 4 and 6 dimensions), each replication is then projected through the common basis path as the variable(s) containing the were previously shuffled. The resulting animations were saved as .gif files. The radial tour starts at either the 4 or 6-variable "half-clock" basis, where each variable has a uniform contribution, and no variable contributing in the opposite direction (to minimize variable dependence), a radial tour is then produced for each variable and saved as a .gif.

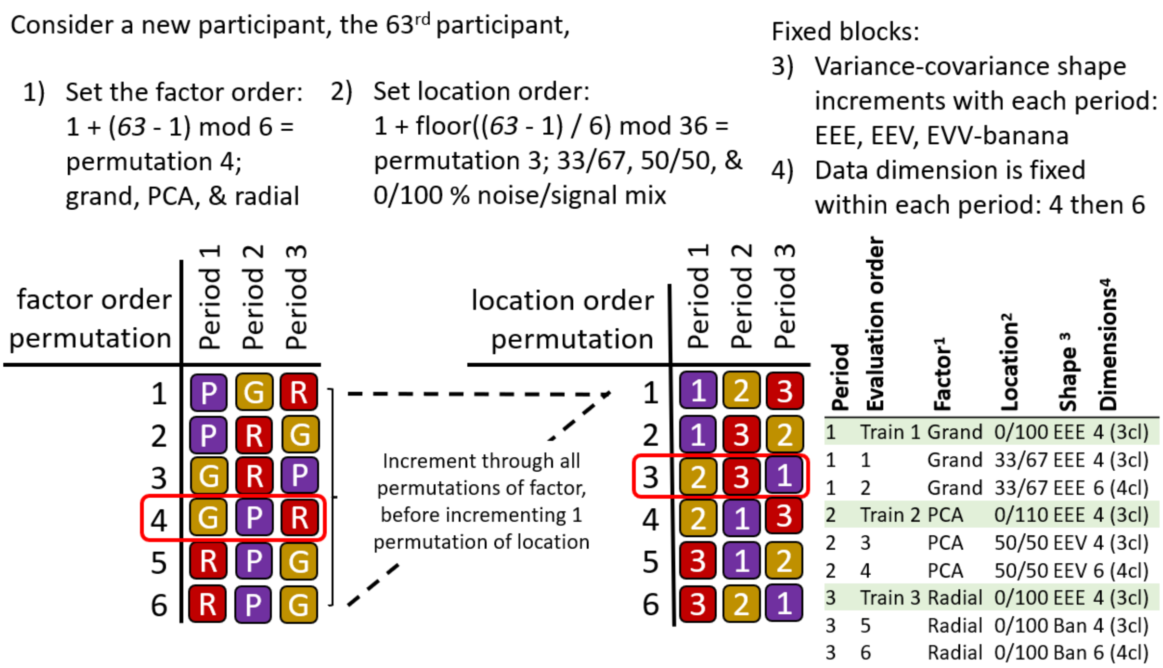


Figure 4: Illustration of how a hypothetical participant 63 is assigned factor and block parameterizations. Each of the 6-factor order permutations is exhausted before iterating to the next permutation of location order.

3.6. Randomized factor assignment

Now, with simulation and their artifacts in hand. We explain how the experimental factors are assignment, and illustrate how this is experienced from a participant's perspective.

We section the study into 3 periods, each period is linked to a randomized level of both the factor visualization and the location. The order of dimension and shape are of secondary interest and are held constant in increasing order of difficultly; 4 then 6 dimensions and EEE, EEV, then EVV-banana respectively.

The period starts with an untimed training task at the simplest remaining block parameterization; location = 0/100%, shape = EEE, and 4 dimensions with 3 clusters. This serves to introduce and familiarize participants with input and visual differences. After the training, the participant is evaluated on 2 tasks with the same factor * location level, across the increasing difficulty of dimension * shape. These evaluations removed the plot after 60 seconds, though this limit was rarely reached by participants.

The order of the levels of the factor and location is randomized with a nested Latin square where all levels of the visual factor are exhausted before advancing to the next level of location. That means we need $3!^2 = 36$ participants to perform a full block evaluation. This randomization is important to control for any potential learning effects the participant may receive. Figure 4 illustrates how an arbitrary participant experiences the experimental factors.

Through pilot studies sampled by convenience (information technology and statistics Ph.D. students attending Monash University), we predict that we need 3 full evaluations to properly power our study; we set out to crowdsource $N = 3 * 3!^2 = 108$ participants.

3.7. Participants

We recruited $N = 108$ participants via prolific.co (Palan and Schitter 2018). We filtered participants based on their claimed education requiring that they have completed at least an undergraduate degree (some 58,700 of the 150,400 users at the time); we apply this filter under the premise that linear projections and biplot displays used will not be regularly used for consumption by general audiences. There is also the implicit filter that Prolific participants must be at least 18 years of age and location/language bias associated with. 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 validate this assumption in the follow-up survey where we ask about familiarity with the factors (see Figure 6). The appendix contains a heatmap distribution of age and education paneled across preferred pronouns of the participants that completed the survey, who are relatively young and well educated.

3.8. Data collection

Data were recorded by a **shiny** application and were written to a Google Sheet after each third of the study. Especially at the start of the study, participants experienced adverse network conditions due to the volume of participants hitting the application with modest allocated resources. In addition to this, API read/write limitations further hindered data collection. To mitigate this we throttled the volume of participant and over-collect survey trials until we had received our target 3 evaluations of our 36 permutation levels.

The processing steps were minimal. First, we format to an analysis-ready form, decoding values to a more human-readable state, and add a flag to indicate if the survey had complete data. We filter to only the latest 3 complete studies of each block parameterization, those which should have experienced the least adverse network conditions. Of the studies removed the bulk were partial data and a few over-sampled permutations. 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 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 `https://github.com/nspyrison/spinifex_study`.

4. Results

To recap, the primary response variable is task accuracy as defined in section 3.3, 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: visual factor, location of the cluster separation signal, the shape of variance-covariance matrix, and the dimensionality of the data. Block parameterization and randomization were discussed in section 3.1. 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).

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.

4.1. Accuracy regression

To more thoroughly examine explanatory variables, we regress against accuracy. 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:	**Full model:**
α	$\widehat{Y_1} = \mu + \alpha_i + \mathbf{Z} + \mathbf{W} + \epsilon$
$\alpha + \beta + \gamma + \delta$	$\widehat{Y_1} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \mathbf{Z} + \mathbf{W} + \epsilon$
$\alpha * \beta + \gamma + \delta$	$\widehat{Y_1} = \mu + \alpha_i * \beta_j + \mathbf{Z} + \mathbf{W} + \epsilon$
$\alpha * \beta * \gamma + \delta$	$\widehat{Y_1} = \mu + \alpha_i * \beta_j * \gamma_k + \mathbf{Z} + \mathbf{W} + \epsilon$
$\alpha * \beta * \gamma * \delta$	$\widehat{Y_1} = \mu + \alpha_i * \beta_j * \gamma_k * \delta_l + \mathbf{Z} + \mathbf{W} + \epsilon$

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 $\mid i \in$ (pca, grand, radial)
β_j, fixed term for location $\mid j \in$ (0/100%, 33/66%, 50/50%) % noise/signal mixing
γ_k, fixed term for shape $\mid k \in$ (EEE, EEV, EVV banana) model shapes
δ_l, fixed term for dimension $\mid l \in$ (4 variables & 3 cluster, 6 variables & 4 clusters)

We also want to visually explore the conditional variables in the model. Figure 5 explores violin plots of accruacy by factor while faceting on the location (vertical) and shape (horizontal). Radial tends to increase the accruacy received, and especially so when there is no signal/noise mixing.

4.2. response time regression

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)

Table 1: Model comparison of our random effect models regressing accruacy. Each model includes a random effect term of the participant, which explains the individual's influence on their accruacy. Complex models perform better in terms of R^2 and RMSE, yet AIC and BIC penalize their large number of fixed effects in favor of the much simpler model containing only the visual 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*

Table 2: The task accuracy model coefficients for $\widehat{Y_1} = \alpha * \beta + \gamma + \delta$, 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% 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	

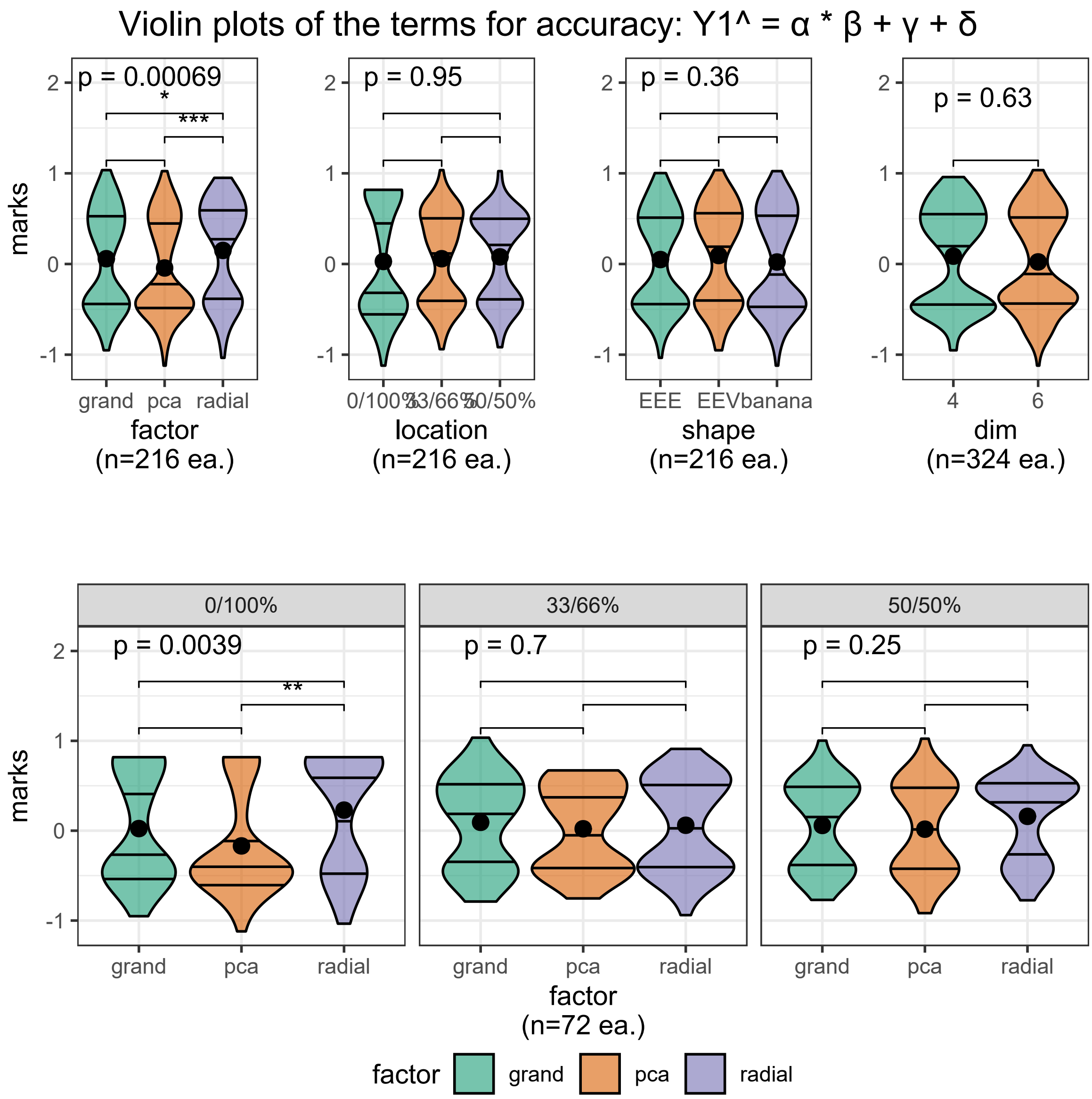


Figure 5: Violin plots of terms of the model $\widehat{Y_1} = \alpha * \beta + \gamma + \delta$. Overlaid with global significance from the Kruskal-Wallis test, and pairwise significance from the Wilcoxon test, both are non-parametric, ranked sum tests suitable for handling discrete data. Participants 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(time), $\widehat{Y_2}$ 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 preferred by AIC/BIC, while R^2 and RMSE prefers the full multiplicative model. We again select the model $\alpha * \beta + \gamma + \delta$ to explore further as it has relatively high marginal R^2 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*

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 6 shows the Likert plots, or stacked percentage bar plots, alongside violin plots with the same non-parametric, ranked sum tests previously used. Participants 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 discussed an $n = 108$, 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. This was evaluated evenly over 4 block parameterizations. In summary, we find that radial tour increases accuracy while 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. The radial tour was subjectively most preferred, lead to more confidence in answers, and is easier to use than alternatives.

There are several ways that this study could be extended. In addition to expanding the support of the block parameterizations, more interesting directions include: type of the task, visualizations 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 the target support of your block parameters. 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.

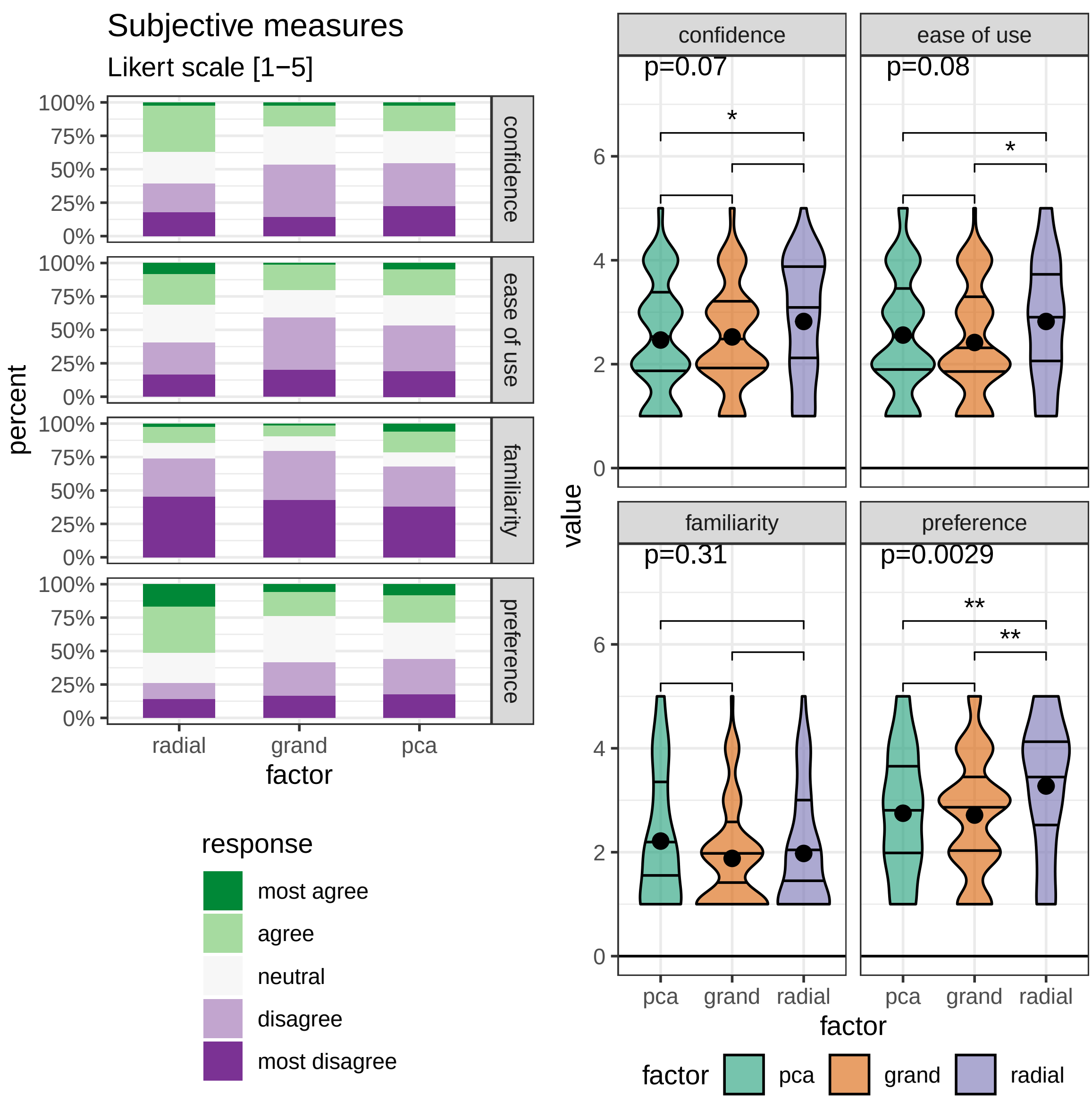


Figure 6: The subjective measures of the 84 responses of the post-study survey, 5 discrete Likert scale levels of agreement (L) Likert plots (stacked percent bar plots) with (R) violin plots of the same measures. Violin plots are overlaid with global significance from the Kruskal-Wallis test, and pairwise significance from the Wilcoxon test, both are non-parametric, ranked sum tests.

Table 4: The log(time) model coeffients for $\widehat{Y_2} = \alpha * \beta + \gamma + \delta$, with factor=pca, location=0/100%, and shape=EEE held as baselines. Location=50/50% 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	

6. Accompanying tool: radial tour 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, perform limited preprocessing, and interactively explore their data via interactive radial tour. The .html widget produced is a more interactive variant relative to the application in the user study. Screen captures and more details are provided in the appendix. Data can be imported from .csv, .rds, or .rda format, and projections. Run the following R code to run the application locally.

```
install.packages("spinifex", dependencies = TRUE)
spinifex::run_app("radial_tour")
```

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 https://github.com/nspyrison/spinifex_study/. The source code for the **spinifex** package and accompanying shiny application can be found at https://github.com/nspyrison/spinifex/.

8. References

10 Adadi, Amina, and Mohammed Berrada. 2018. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)." *IEEE Access* 6: 52138–60.

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

Arrieta, Alejandro Barredo, Natalia Díaz-Rodriguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, and Richard Benjamins. 2020. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI." *Information Fusion* 58: 82–115.

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.

Biecek, Przemyslaw. 2018. "DALEX: Explainers for Complex Predictive Models in R." *The Journal of Machine Learning Research* 19 (1): 3245–49.

Biecek, Przemyslaw, and Tomasz Burzykowski. 2021. *Explanatory Model Analysis: Explore, Explain, and Examine Predictive Models*. CRC Press.

Chambers, J. M., W. S. Cleveland, B. Kleiner, and P. A. Tukey. 1983. "Graphical Methods for Data Analysis."

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.

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.

Cook, Dianne, Andreas Buja, Eun-Kyung Lee, and Hadley Wickham. 2008. "Grand Tours, Projection Pursuit Guided Tours, and Manual Controls." In *Handbook of Data Visualization*, 295–314. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-33037-0_13.

Gracia, Antonio, Santiago GonzÃ¡lez, VÃctor 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.

Lee, Stuart, Dianne Cook, Natalia da Silva, Ursula Laa, Earo Wang, Nick Spyrison, and H. Sherry Zhang. 2021. "A Review of the State-of-the-Art on Tours for Dynamic Visualization of High-Dimensional Data." *arXiv:2104.08016 [Cs, Stat]*, April. http://arxiv.org/abs/2104.08016.

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.

Munzner, Tamara. 2014. "Visualization Analysis and Design."

Ocagne, Maurice d'. 1885. *Coordonnã©es Parallã¨les Et Axiales. MÃ©thode de Transformation gã©omã©trique Et Procã©dã© Nouveau de Calcul Graphique dã©duits de La Considã©ration Des Coordonnã©es Parallã¨les, Par Maurice d'Ocagne, ...* Paris: Gauthier-Villars.

Palan, Stefan, and Christian Schitter. 2018. "Prolific: 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.

Wickham, Hadley, Dianne Cook, and Heike Hofmann. 2015. "Visualizing Statistical Models: Removing the Blindfold." *Statistical Analysis and Data Mining: The ASA Data Science Journal* 8 (4): 203–25. https://doi.org/10.1002/sam.11271.

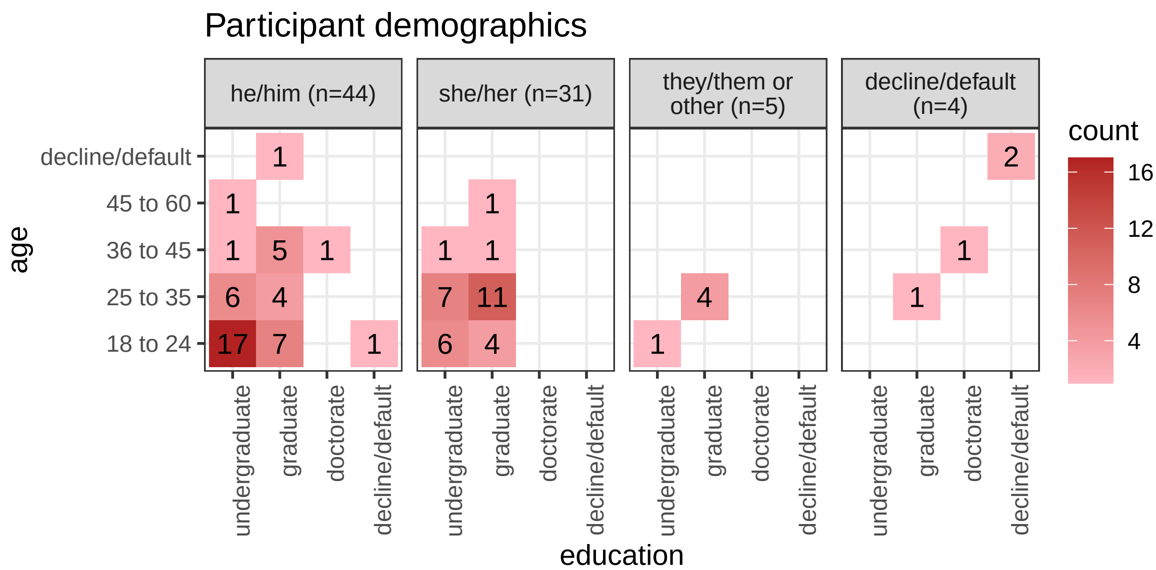


Figure 7: Heatmaps of participant demographics, counts of age group by completed education as faceted across preferred pronoun. Our sample tended to be between 18 and 35 years of age with an undergraduate or graduate degree.

Wickham, Hadley, Dianne Cook, Heike Hofmann, and Andreas Buja. 2011. "Tourr: An R Package for Exploring Multivariate Data with Projections." *Journal of Statistical Software* 40 (2). https://doi.org/10.18637/jss.v040.i02.

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.

Yanai, Itai, and Martin Lercher. 2020. "A Hypothesis Is a Liability." *Genome Biology* 21 (1): 231. https://doi.org/10.1186/s13059-020-02133-w.

9. Appendix

Survey participant demographics

The target population is relatively educated people as such linear projections may be difficult for generalized consumption. Hence we restrict Prolific.co participants to those with an undergraduate degree (58,700 of the 150,400 users at the time of the study). We found our sample of 108 participants from this group. Of these participants, 84 submitted the post-study survey, who are represented in the heatmap. All participants were compensated for their time at £7.50 per hour, with a mean time of about 16 minutes.

Random effect ranges

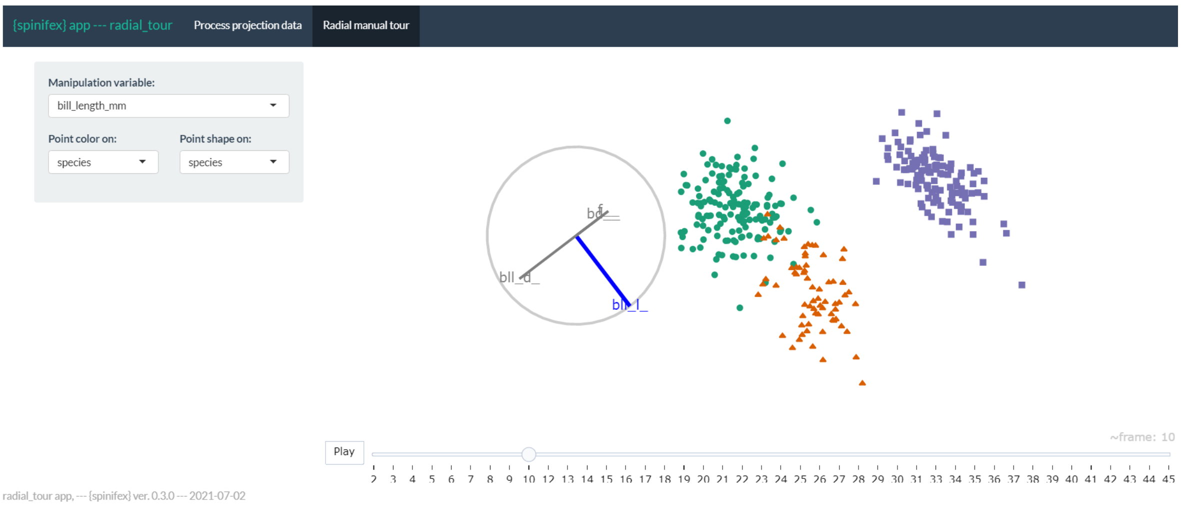
Residual plots have no noticeable non-linear trends and contain striped patterns as an artifact from regressing on discrete variables. Figure 8 illustrates (T) the effect size of the random terms participant and simulation, or more accurately, the 95% CI from Gelman simulation of their posterior distribution. 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.

Radial tour application details

Below we describe the **shiny** app made available in the **spinifex** package which is then run locally. It streamlines the process to creating and interacting with radial tours, related to, but more interactive than the application used in the user study.

In the initial tab, users upload their own (.csv, .rds, or .rda) data or select from predefined data sets. The numeric columns appear as a list of variables to include in the projection. Below that, a line displays whether or not NA rows were removed. Scaling by standard deviation is included by default, as this is a common transformation used to explore linear projections of spaces. Alternatively, the user can perform their transformations before exporting and toggling this transformation off. Summaries of the raw data, and processed numeric data to project are displayed to better illustrate how the data was read and what will be projected.

\begin{figure}

{

}

\caption{Radial tour tab, interactively creates radial tours, changing the variable of manipulation, and color or shape of the resulting manual tour. Here, the palmer penguins data is being explored, bill_length was selected to manipulate as it is the only variable separating the green cluster from the orange. By mapping shape to island of observation, we also notice that the species in green lives on all 3 islands, while the other specees live only on one of the islands.} \end{figure}

The second tab contains interaction for selecting the manipulation variable and non-numeric columns can be used to change the color and shape of the data points in the projection. The radial tour is then created in real-time animating as an interactive **plotly** .html widget.

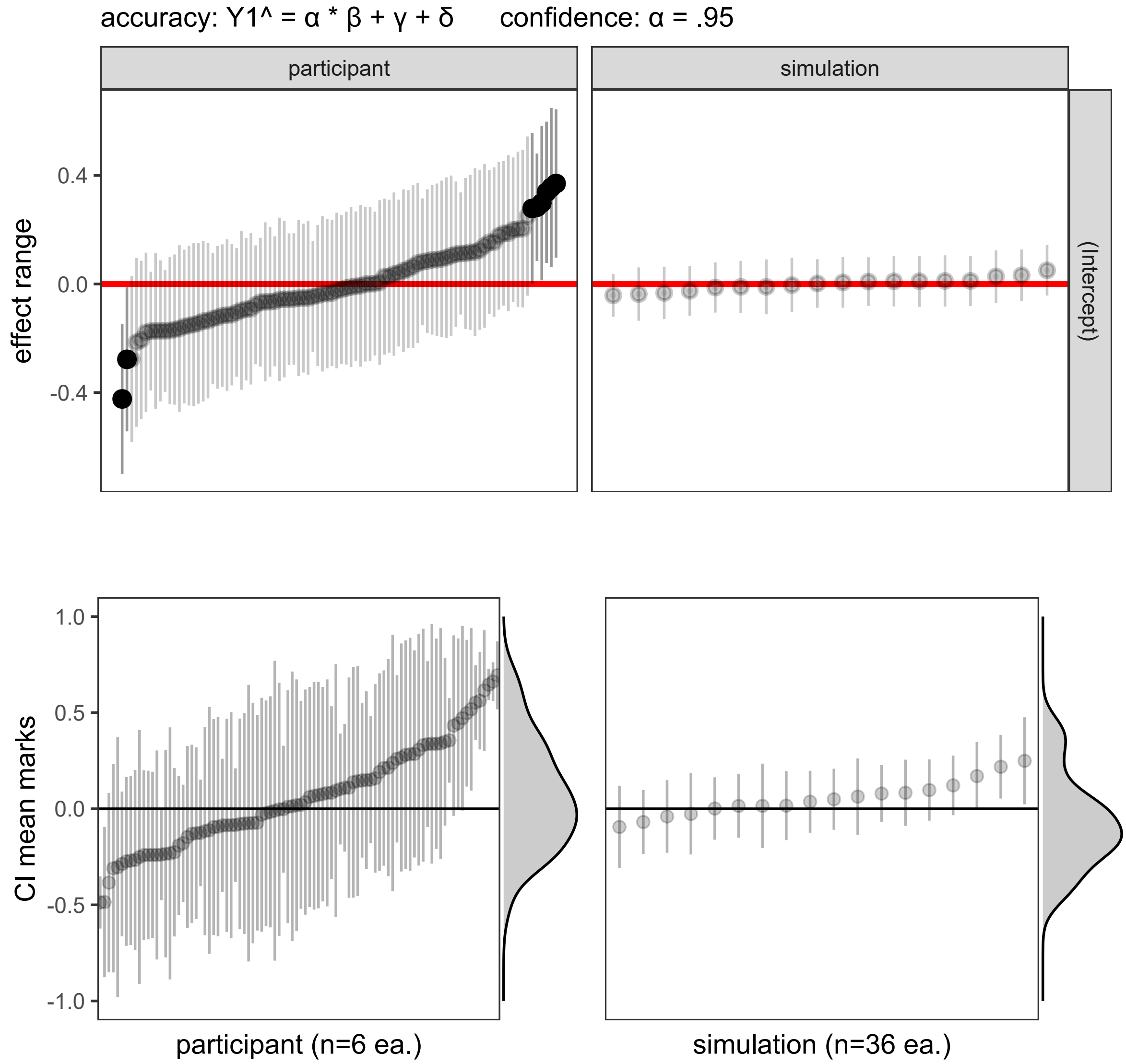


Figure 8: (T) Estimated effect ranges of the random effect terms participant and data simulation of the accuracy model, $Y_1 = \alpha * \beta + \gamma + \delta$. Confidence intervals are created with Gelman simulation on the effect posterior distributions. The effect size of the participant is relatively 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.

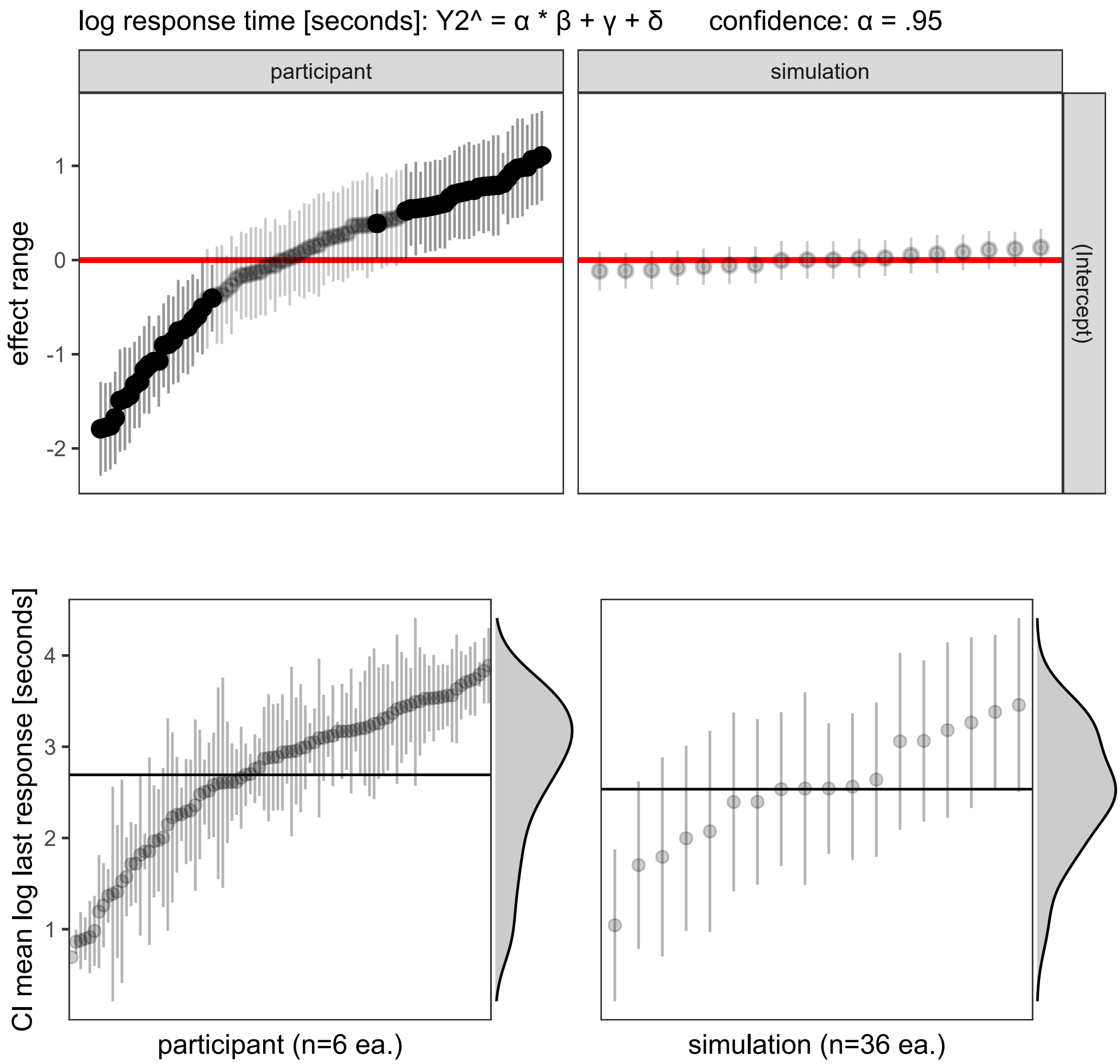


Figure 9: (T) The effect ranges of Gelman resimulation on posterior distributions for the time model, $Y_2 = \alpha * \beta + \gamma + \delta$. These show the magnitude and distributions of particular participants and simulations. Simulation has a relatively small effect on response time. (B) Confidence 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 in response time, relative to the effect ranges which means that the variation is explained in the terms of the model and not by the simulation itself.

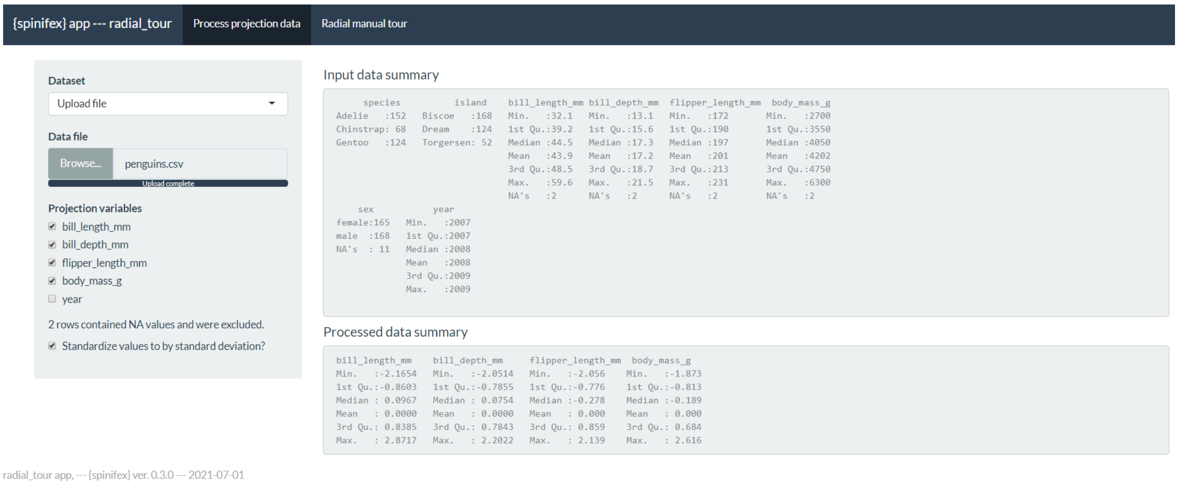


Figure 10: Process data tab, interactively loads or select data, check which variables to project, and optionally scale columns by standard deviation.

Including such an application offers users a fast, intuitive introduction of what is going on and some of the features offered. for more help getting started on applying your tours see the help documentation or follow along with our vignettes.

Affiliation:

Nicholas Spyrison
Monash University
Faculty of Information Technology
ORCiD: 0000-0002-8417-0212
E-mail: nicholas.spyrison@monash.edu
Dianne Cook
Monash University
Department of Econometrics & Business Statistics,
ORCiD: 000-0002-3813-7155
Kimbal Marriott
Monash University
Faculty of Information Technology,
ORCiD: 0000-0002-9813-0377

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