# s-CorrPlot: An Interactive Scatterplot for Exploring Correlation

Sean McKenna*, Miriah Meyer*, Christopher Gregg†, and Samuel Gerber‡

February 20, 2015

**Abstract**

The degree of correlation between variables is used in many data analysis applications as a key measure of interdependence. The most common techniques for exploratory analysis of pairwise correlation in multivariate datasets, like scatterplot matrices and clustered heatmaps, however, do not scale well to large datasets, either computationally or visually. We present a new visualization that is capable of encoding pairwise correlation between hundreds of thousands variables, called the s-CorrPlot. The s-CorrPlot encodes correlation spatially between variables as points on scatterplot using the geometric structure underlying Pearson's correlation. Furthermore, we extend the s-CorrPlot with interactive techniques that enable animation of the scatterplot to new projections of the correlation space, as illustrated in the companion video in Supplemental Materials. We provide the s-CorrPlot as an open-source R-package and validate its effectiveness through a variety of methods including a case study with a biology collaborator.

*Keywords:* Correlation, exploratory data analysis, multivariate data.

## 1 Introduction

Pearson's correlation coefficient is a basic and widely used correlation measure, which captures the degree of a linear relationship between two variables. Pearson's correlation is used in a broad range of applications, from finding genes that are involved with a specific disease network [Horvath and Dong, 2008], to finding sociological variables that interact in a complex manner [Allison, 1977]. Visualization of pairwise correlation aims to provide investigators with new hypotheses to test.

*School of Computing, University of Utah, Salt Lake City, UT, 84112. E-mail: sean@cs.utah.edu
†Department of Neurobiology and Anatomy, University of Utah, Salt Lake City, UT, 84112.
‡Department of Mathematics, Duke University, Durham, NC, 27708.

As the amount of available data continues to expand in these fields, visualizing correlation becomes challenging because standard techniques for exploring correlation, discussed in detail in Section 2, lack the capacity to deal with these increasingly large datasets. This paper develops a novel approach to visualize and explore correlation among many variables based on a spatial encoding termed the s-CorrPlot. The s-CorrPlot is shown in Figure 1(b) on an illustrative dataset; the details of this example are discussed in Section 6.1.

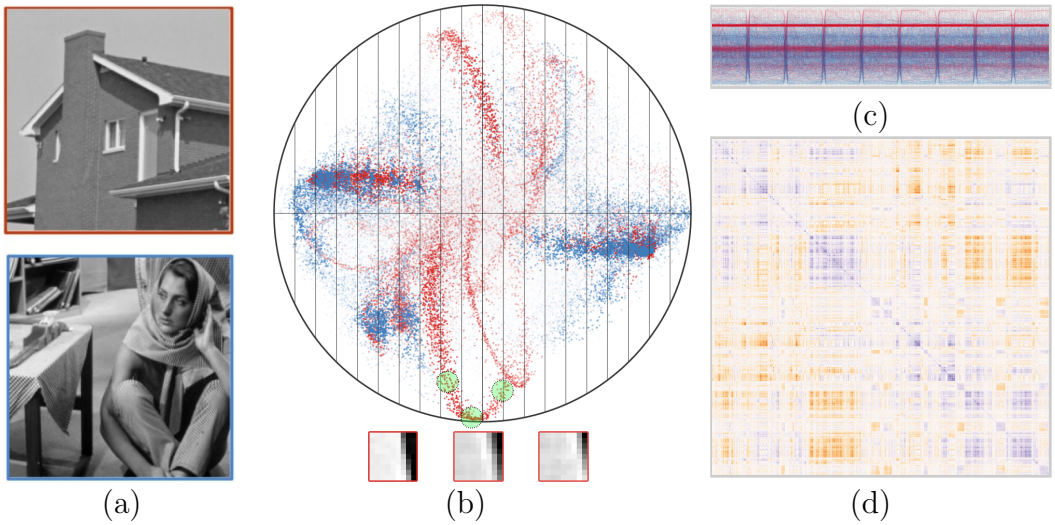



Figure 1: Multiple correlation visualizations for two datasets containing a total of over 130,000 variables, $9 \times 9$ patches of two images (a), with 81 observations (pixels) each. The proposed s-CorrPlot (b) reveals correlation structures between variables such as the highly correlated image patches shown at the bottom corresponding to a horizontal shift of a vertical edge. These continuous variations visible in the s-CorrPlot are not emphasized in a parallel coordinates plot (c) or a clustered heatmap of pairwise correlation coefficients (d). In (b) and (c), color indicates membership of the displayed variable from the images in (a). In (d), purple indicates strong positive correlation and orange strong negative correlation. Only the s-CorrPlot can plot all variables, due to computational and screenspace limitations of (c) and (d).

Scatterplot matrices [Hartigan, 1975, Cleveland and McGill, 1988] and parallel coordinates [Inselberg, 1985], shown in Figure 1(c), rely on a visual determination of correlation, but suffer from underestimation effects [Li et al., 2010] and screen space limitations with large numbers of observations [Friendly, 2002] and variables [Peng et al., 2004]. Scatterplot matrices and parallel coordinates provide a visualization of correlation that shows fine grained detail and can distinguish identical correlation coefficients arising from different causes, such as Anscombe's quartet. To illuminate structure between many variables these plots are not suitable, however.

A standard approach to visualize correlation among many variables are clustered heatmaps

[Wilkinson and Friendly, 2009, Seo and Shneiderman, 2002], as in Figure 1(d). Heatmaps directly visualize correlation between each pair of variables using a color encoding and can highlight clusters of variables reasonably well, but they also have a number of limitations: pairwise correlation computations grow quadratically with the number of variables; clustering is necessary for pattern detection, but the visual results are highly variable based on the clustering technique [Seo and Shneiderman, 2002]; and accurate evaluation of correlation values is difficult due to the relative nature of color perception [Albers, 2006, Wong, 2010].

The main contribution of this work is an alternative to heatmaps: a new visual encoding of correlation, called the s-CorrPlot, that highlights correlation structure among many variables as points in a scatterplot. This encoding scales to large datasets containing hundreds of thousands of variables and thousands of observations, both computationally and visually. The visual encoding of the s-CorrPlot is based on an 2D orthogonal projection of standardized variables. The observation that standardized variables lie on a unit hypersphere leads to a derivation of a new exact spatial encoding of correlation through orthogonal projection. The derivation of this new spatial encoding is described in Section 4, where the encoding is precise for a subset of all pairwise correlations with error bounds for the rest. We improve visual scalability of the plot with a density estimation technique that reduces visual clutter and enhances perception of structure.

As a secondary contribution to this work, we illustrate the utility of the s-CorrPlot by combining it with interactive techniques for multidimensional exploration, discussed in Section 5. We provide an implementation as an open-source R package and demonstrate the s-CorrPlot on two datasets in Section 6, including a case study with a biology collaborator.

# 2 Background & Related Work

This work focuses on Pearson's and Spearman's correlation coefficients, which measure the strength of linear and monotonic relationships between two variables, respectively. We adopt the terminology from Dempster [1969] referring to a multivariate $(n \times p)$ sample as $n$ individuals on $p$ variables. Correspondingly, we can think of an individual as a point in the **individual space** of dimension $p$ and a variable as a point in the **variable space** of dimension $n$. Methods for visualizing correlation can be categorized as operating in either of these two spaces.

## 2.1 Visualizing Correlation

Scatterplots are the most basic method for visually determining the correlation between two variables [Cleveland and McGill, 1988, Staudte, 1990, Rensink and Baldridge, 2010, Elmqvist et al., 2008]. In individual space, each axis represents a variable and a plotted point indicates the observed values of an individual for those two variables; correlation is encoded by deviation from a straight line of the resulting point cloud. For comparing more than two variables, a scatterplot matrix (SPLOM) creates a small-multiples view of the scatterplots for all possible combinations of variables [Hartigan, 1975, Cleveland and McGill, 1988]. While a very rich view, a SPLOM requires considerable screen-space and does not scale well to exploring more than several dozen variables in a dataset [Friendly, 2002].

Another method for visually estimating correlation is a parallel coordinates plot [Inselberg, 1985]. This approach offers the choice of operating in individual or variable space. In individual space, the parallel coordinate axes correspond to variables and a line across the axes to an individual; vice-versa for variable space. In individual space, perfect positive correlation between two variables is indicated through parallel lines between the two axes, and perfect negative correlation with all lines crossing between the two axes. For more than a few variables, seriation [Liiv, 2010, Hahsler et al., 2008], i.e. the ordering of the axes, becomes critical to finding meaningful patterns. A number of alternative techniques have been proposed to address seriation, such as placing similar axes near each other based on clustering results [Ankerst et al., 1998] and to reduce visual clutter [Peng et al., 2004, Ellis and Dix, 2006] for plots with more than a few dozen variables. Despite these extensions, studies of correlation between two variables advocate for scatterplots over parallel coordinate plots, though parallel coordinate plots are more perceptually effective for negative correlation values [Harrison et al., 2014]; additionally, both encodings suffer from underestimation effects [Li et al., 2010]. In variable space, positive correlation is indicated by parallel movement of two lines across all axes and negative correlation by two mirrored lines. While this approach mitigates the seriation issue of comparing multiple variables, it is susceptible to overplotting with more than several dozen variables.

A standard approach to explore correlation among many variables is to calculate the correlation coefficient between all pairs of variables and visualize the resulting correlation coefficient matrix directly. The canonical technique used to visualize this matrix is a heatmap [Wilkinson and

Friendly, 2009], where each correlation coefficient in the matrix is encoded with color. While this data-dense visualization scales to hundreds of variables, seriation or ordering of variables becomes crucial [Liiv, 2010, Hahsler et al., 2008], often implemented as a clustering of the rows and columns in order to enhance the perception of trends in the data [Eisen et al., 1998, Seo and Shneiderman, 2002]. The observable patterns in a clustered heatmap, however, vary greatly depending on the method of seriation [Wilkinson and Friendly, 2009]. Furthermore, it is difficult to scale heatmaps to thousands of variables, both visually due to display limitations [Seo and Shneiderman, 2002], and computationally due to the quadratic cost, in the number of variables, for both memory and processing time of the correlation coefficient matrix [Bohn et al., 2009].

The color encoding of correlation coefficient values in a heatmap has two additional visualization drawbacks. First, the relative nature of color perception makes accurate interpretation of the correlation values difficult [Albers, 2006, Wong, 2010]. And second, in many settings it is of interest to annotate the variables with additional information, such as categories or set membership of variables. Without the use of color, visualizing this additional information is cumbersome to do, particularly for such large datasets.

To address the limitations of existing correlation visualization methods, the s-CorrPlot uses a novel spatial encoding of correlation rooted in the geometric interpretation of correlation in variable space, discussed in Section 3. There are several other encodings that stem from this geometric interpretation in variable space [Corsten and Gabriel, 1976, Falissard, 1999, Trosset, 2005]; we compare the s-CorrPlot against these encodings in Section 4.5. These spatial encodings lead to more accurate visual inferences [Bertin, 1983, Cleveland and McGill, 1984], while the color channel is free to encode additional information.

# 3 Geometric Interpretation of Correlation

In this section, we describe the geometric interpretation of Pearson's correlation [Dempster, 1969, Rodgers et al., 1984] that underpins the s-CorrPlot. The geometrical interpretation represents each **variable** as a vector in $\mathbb{R}^n$, where $n$ is the number of **observations** per variable. In this interpretation, Pearson's correlation is the cosine of the angle between the mean centered variables. Thus, correlation can be spatially represented as $p$ points on a $(n-2)$-sphere. In statistical language, the points on this sphere are termed standardized variables. We provide the mathematical

details of this interpretation in the rest of this section.

We use the notation of a $k$-**flat** to refer to a $k$-dimensional linear subspace of $\mathbb{R}^n$ that does not necessarily contain the origin. Thus, a plane is a 2-flat and a hyperplane a $(n-1)$-flat. Similarly, we refer to a $k$-**sphere** as the generalization of a unit sphere to $k$ dimensions.

Pearson's correlation coefficient, $\hat{r}$, for any two variables $\mathbf{x} = \{x_1, \ \dots \ , x_n\}$ and $\mathbf{y} = \{y_1, \ \dots \ , y_n\}$ is

$$\hat{r}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{1}$$

where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ are the means of $\mathbf{x}$ and $\mathbf{y}$, respectively.

By letting $\tilde{\mathbf{x}} = \{(x_1 - \overline{x}), \ \dots \ , (x_n - \overline{x})\}$ and $\tilde{\mathbf{y}} = \{(y_1 - \overline{y}), \ \dots \ , (y_n - \overline{y})\}$, Equation 1 can be written as

$$\hat{r}(\mathbf{x}, \mathbf{y}) = \frac{\tilde{\mathbf{x}} \cdot \tilde{\mathbf{y}}}{\|\tilde{\mathbf{x}}\| \ \|\tilde{\mathbf{y}}\|}, \tag{2}$$

which highlights the geometrical nature of the correlation coefficient. In this geometric view, the standardization of a variable can be viewed as a projection to the **correlation sphere** — a specific, $(n-2)$-sphere embedded in $\mathbb{R}^n$. To standardize a variable, first, the mean of the variable is subtracted from each observation, and second, the variable is scaled to unit length. The first step corresponds to projecting the variable onto a $(n-1)$-flat orthogonal to $\mathbf{1}$, the all ones vector. More formally, the subtraction of the mean for any variable $\mathbf{x}$ can be written as $\tilde{\mathbf{x}} = \mathbf{x} - \frac{\mathbf{x}\cdot\mathbf{1}}{n}\mathbf{1}$. This is a projection onto the hyperplane $\mathbf{x} \cdot \mathbf{1} = 0$, thus, the mean subtracted variables reside in a $(n-1)$-flat. The second step, scaling to unit length, results in the projection of the variables onto the aforementioned $(n-2)$-sphere. For any two variables, their correlation is now directly encoded through the relative positions of their **standardized variables**. For any two standardized variables close to each other on the sphere, their dot product, and thus their correlation coefficient, is close to 1, and, for those that lie on opposite sides, it will be $-1$.

Spearman's correlation coefficient is Pearson's correlation but on ranked observations. Pearson's correlation coefficient describes the strength of linear relationships between variables, while Spearman's correlation coefficient describes monotonic ones. The ranked and standardized variables correspond to a subset of the correlation sphere. For the reader inclined to combinatorics, the ranking leads to standardized variables that can be identified with vertices of the permutohedron of order $(n-2)$. Both the geometric interpretation and the s-CorrPlot apply to either measure of correlation.

# 4 Spatial Correlation Scatterplot

In this section, we build on the existing geometric interpretation of correlation in variable space to present the description and analysis behind the s-CorrPlot. The s-CorrPlot represents each variable as a point on a scatterplot. The scatterplot results from an orthogonal projection of the multidimensional correlation sphere. From the geometrical description in Section 3, we derive a new and precise encoding of correlation through the projection of standardized variables.

In the rest of this section, we first derive the novel spatial encoding of correlation: the s-CorrPlot. We show that a projection can be used to display the correlation exactly for a subset of the dataset and then quantify and derive error bounds for determining correlation between any two variables in the resulting scatterplot. Then, we highlight a density estimation technique that enhances the perception of patterns and structure in large datasets.

The s-CorrPlot adds a novel approach to encode and read correlation from the resulting 2D scatterplot. We compare the s-CorrPlot to heatmaps in Section 4.4 and other spatial encodings of variable space in Section 4.5.

## 4.1 The s-CorrPlot

Equation 2 can be used to compute a standard correlation coefficient matrix, which can then be visualized with a heatmap display. The geometric interpretation of correlation, however, supports another line of reasoning: we can project the standardized variables that lie on the correlation sphere onto a plane through the origin. After the projection step, the variables can be displayed as points on a scatterplot. The rest of this section describes how correlation can be directly read from the projected standardized variables.

We can define a plane through the origin by selecting two noncollinear standardized variables, $\mathbf{p}$ and $\mathbf{s}$ on the correlation sphere, as illustrated by $U$ in Figure 2. Projecting all variables onto this plane collapses them to points on the **s-CorrPlot**. The projection plane forms a circular intersection with the selected points. For any variable on this circular boundary, the correlation to any other projected variable in the plane is encoded exactly. Furthermore, error bounds in Section 4.2 show that the approximation error increases slowly as one moves away from the boundary.

To graphically illustrate how the s-CorrPlot encodes correlation coefficients, we will use a simple example of three variables with observations for four individuals. Thus, the variables can

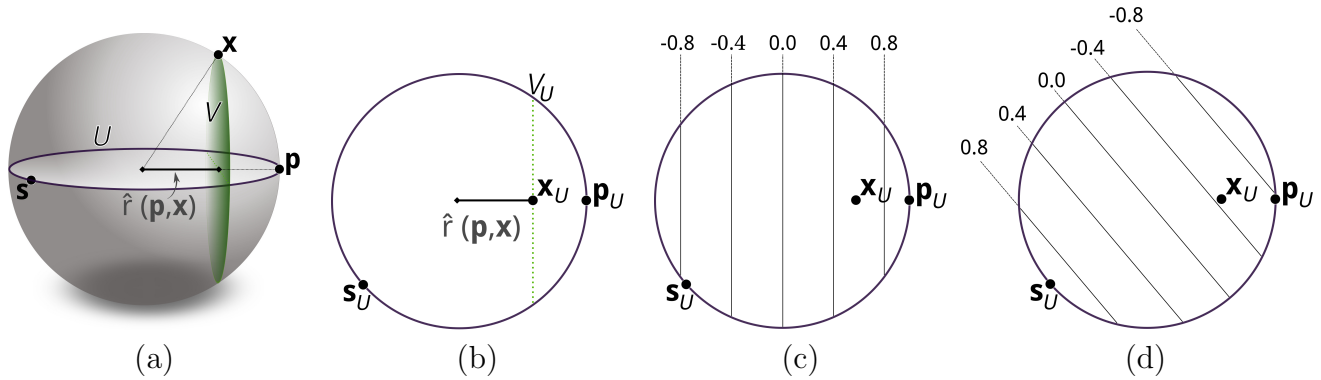



Figure 2: As an example, we show how three variables ($\mathbf{p}$, $\mathbf{s}$, and $\mathbf{x}$) with four observations each project onto the s-CorrPlot. For (a), we can illustrate our variables as standardized vectors on the correlation sphere, directly shown in 3D here. The correlation coefficient between any two variables is the dot product between their standardized vectors, such as with $\mathbf{p}$ and $\mathbf{x}$. With these two standardized variables, a $(n-2)$-flat $V$ is defined. The s-CorrPlot is defined by the projection plane $U$, containing both $\mathbf{p}$ and $\mathbf{s}$. Projection onto $U$ results in the s-CorrPlot as shown in (b), preserving correlation coefficients to both $\mathbf{p}$ and $\mathbf{s}$. In the s-CorrPlot, $V$ projects to a vertical line $V_U$ of equal correlation to $\mathbf{p}$. As such, (c) these vertical lines can be generalized as grid lines along $U$, denoting sets of equidistant correlation values to $\mathbf{p}$. Similarly, (d) grid lines to $\mathbf{s}$ can be shown.

be represented as vectors in $\mathbb{R}^4$, as with the three standardized variables: $\mathbf{p}$, $\mathbf{s}$, and $\mathbf{x}$. Because these standardized variables effectively reside in a 3D subspace of $\mathbb{R}^4$, we can directly illustrate them in Figure 2(a). Next, the correlation sphere is intersected with a projection plane, $U$, which is defined as going through the origin and containing any two noncollinear variables $\mathbf{p}$ and $\mathbf{s}$. $U$ is represented by a [2 x $n$] matrix, defined as:

$$U = [\mathbf{p}, o(\mathbf{p}, \mathbf{s})]^\top \tag{3}$$

with column vectors $\mathbf{p}$ and $\mathbf{s}$, where:

$$o(\mathbf{p}, \mathbf{s}) = \frac{\mathbf{s} - (\mathbf{s} \cdot \mathbf{p})\mathbf{p}}{\|\mathbf{s} - (\mathbf{s} \cdot \mathbf{p})\mathbf{p}\|} \,. \tag{4}$$

Any standardized variable, $\mathbf{x}$, can be projected onto the plane $U$, producing the 2D coordinates of the s-CorrPlot, $\mathbf{x}_U = U\,\mathbf{x}$. This projection results in the s-CorrPlot, shown in Figure 2(b).

Based on Equation 2, the correlation coefficient for two variables is equal to the dot product between their vectors, such as for $\mathbf{p}$ and $\mathbf{x}$ as illustrated in Figure 2(a) and reflected in Figure 2(b). Thus since $\mathbf{x}_U = U\,\mathbf{x}$, it follows that, for any vector $\mathbf{x}$, the correlation to $\mathbf{p}$ is directly encoded in the first component of the vector $\mathbf{x}_U$. In fact, any vector that projects onto the line $V_U$, shown in Figure 2(b), has the same first component value, and thus the same correlation to $\mathbf{p}$. The line

$V_U$ corresponds to the projection of a $(n-2)$-flat, $V$, onto $U$, where $V$ is orthogonal to $\mathbf{p}$ and contains $\mathbf{x}$. Thus, any vector that lies on $V$ is at the same distance from $\mathbf{p}$, and thus has the same correlation value. We illustrate this $(n-2)$-flat $V$ in Figure 2(a). Moving $V$ along the vector $\mathbf{p}$ produces grid lines as shown in Figure 2(c). These grid lines specify values of equal correlation to $\mathbf{p}$ for any location in the s-CorrPlot.

Similarly, we can establish the correlation to $\mathbf{s}$ for any location in the s-CorrPlot by defining the flat $V$ perpendicular to the vector $\mathbf{s}$, illustrated in Figure 2(d). By projecting all the standardized vectors with respect to $\mathbf{p}$ and $\mathbf{s}$, the s-CorrPlot spatially encodes the correlation coefficients with respect to these two variables. The projection, for $p$ variables, results in a $O(p)$ algorithm for generating any single projection on the s-CorrPlot.

This spatial encoding of correlation affords several advantages. First, the geometric interpretation of correlation underlying the s-CorrPlot enables multidimensional exploration techniques. Also, categorical information for groups of variables can be encoded using color or shape. Lastly, linear computation for each scatterplot enables scaling to large datasets at interactive frame rates.

## 4.2 Bounds on the s-CorrPlot

Although the s-CorrPlot *exactly* encodes the correlation of variables that lie on the circular boundary to any other variable, the pairwise comparisons become less exact when both variables fall closer to the center. Thus, there is some uncertainty between any two variables displayed in the s-CorrPlot. Note, this uncertainty is not to be confused with any statistical uncertainty arising from the estimation of the correlation coefficient between the two variables. The uncertainty is solely due to the loss of information in the projection step of the s-CorrPlot.

For any single plot, the correlation value $\hat{r}(\mathbf{x}, \mathbf{y})$ is within the range:

$$\hat{r}_{\max}(\mathbf{x}, \mathbf{y}) = \frac{\|\mathbf{x}_U\|^2 + \|\mathbf{y}_U\|^2 - \|\mathbf{x}_U - \mathbf{y}_U\|^2}{2} + d_x d_y \tag{5}$$

$$\hat{r}_{\min}(\mathbf{x}, \mathbf{y}) = \frac{\|\mathbf{x}_U\|^2 + \|\mathbf{y}_U\|^2 - \|\mathbf{x}_U - \mathbf{y}_U\|^2}{2} - d_x d_y \tag{6}$$

with $d_{\mathbf{x}} = \sqrt{1 - \|\mathbf{x}_U\|^2}$ and $d_{\mathbf{y}} = \sqrt{1 - \|\mathbf{y}_U\|^2}$, where $\mathbf{x}_U$ and $\mathbf{y}_U$ are the coordinates of the projected $\mathbf{x}$ and $\mathbf{y}$ standardized variables on the s-CorrPlot. As either $\mathbf{x}_U$ or $\mathbf{y}_U$ approaches the boundary of the s-CorrPlot, $(\hat{r}_{max}(\mathbf{x}, \mathbf{y}) - \hat{r}_{min}(\mathbf{x}, \mathbf{y})) \to 0$. On the other extreme, as both $\mathbf{x}_U$ or $\mathbf{y}_U$ approach $\mathbf{0}$, the center of the s-CorrPlot, the bounds range from $-1$ to 1, and the correlation

between the two is completely undetermined by the spatial encoding. We provide a derivation of Equations 5 and 6 in the rest of this section.

For two standardized variables, their squared Euclidean distance determines their correlation:

$$\hat{r}(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{y} = 1 - \frac{\|\mathbf{x} - \mathbf{y}\|^2}{2}, \tag{7}$$

which is derived by rewriting the squared Euclidean distance in terms of the inner product $(\mathbf{x} - \mathbf{y}) \cdot (\mathbf{x} - \mathbf{y}) = 2 - 2\mathbf{x} \cdot \mathbf{y}$.

The correlation coefficient between two projected variables $\mathbf{x}_U$ and $\mathbf{y}_U$ onto $U$ is bounded by the minimal and maximal squared distance between $\mathbf{x}$ and $\mathbf{y}$ on the correlation sphere, such that $\mathbf{x}$ is projected to $\mathbf{x}_U$ and $\mathbf{y}$ is projected to $\mathbf{y}_U$. To derive the minimal and maximal squared distances, let $N_\mathbf{x} = \{\mathbf{z} | U\mathbf{z} = \mathbf{x}_U\}$ and $N_\mathbf{y} = \{\mathbf{z} | U\mathbf{z} = \mathbf{y}_U\}$, *i.e.* the flats orthogonal to $U$ (the nullspace of $U$) centered at $\mathbf{x}_U$ and $\mathbf{y}_U$, respectively. All variables $\mathbf{x} \in N_\mathbf{x}$ and $\mathbf{y} \in N_\mathbf{y}$ are by definition projected to $\mathbf{x}_U$ and $\mathbf{y}_U$, respectively. The flats $N_\mathbf{x}$ and $N_\mathbf{y}$ are parallel at a distance of $\|\mathbf{x}_U - \mathbf{y}_U\|$ and intersect the correlation sphere at distances $d_\mathbf{x}$ and $d_\mathbf{y}$ from $U^\top \mathbf{x}_U$ and $U^\top \mathbf{y}_U$, respectively. The intersections of $N_\mathbf{x}$ and $N_\mathbf{y}$ with the correlation spheres are $(n-4)$-spheres with radius $d_\mathbf{x}$ and $d_\mathbf{y}$.

The minimal and maximal distances between two variables on these $(n-4)$-spheres are achieved when the variables are located on the closest and farthest points between the two $(n-4)$-spheres, which in turn are located at the poles with respect to the coordinate system induced $N_\mathbf{x}$ or $N_\mathbf{y}$. Together with the distance between the spheres, this yields a minimal squared distance of:

$$\|\mathbf{x}_U - \mathbf{y}_U\|^2 + (d_\mathbf{x} - d_\mathbf{y})^2 \tag{8}$$

and a maximal squared distance of:

$$\|\mathbf{x}_U - \mathbf{y}_U\|^2 + (d_\mathbf{x} + d_\mathbf{y})^2 \ . \tag{9}$$

These squared Euclidean distances combined with Equation 7 yield the upper bound

$$\hat{r}_{\max}(\mathbf{x}, \mathbf{y}) = 1 - \frac{\|\mathbf{x}_U - \mathbf{y}_U\|^2 + (d_\mathbf{x} - d_\mathbf{y})^2}{2} \tag{10}$$

and lower bound

$$\hat{r}_{\min}(\mathbf{x}, \mathbf{y}) = 1 - \frac{\|\mathbf{x}_U - \mathbf{y}_U\|^2 + (d_\mathbf{x} + d_\mathbf{y})^2}{2} \tag{11}$$

which result in Equation 5 and Equation 6.

Note that the minimal distance corresponds to an upper limit, and the maximal distance corresponds to a lower limit on the correlation coefficient. For $n \leq 4$, these are the only correlation values possible between two variables, while for $n > 4$ there exists a continuous path between the locations that achieve the minimal and maximal distance, and thus any correlation between the upper and lower bound is possible. The upper and lower bounds, here geometrically deduced, can be formally derived as the solutions of a constrained minimization and maximization problem.

## 4.3 Density Estimation

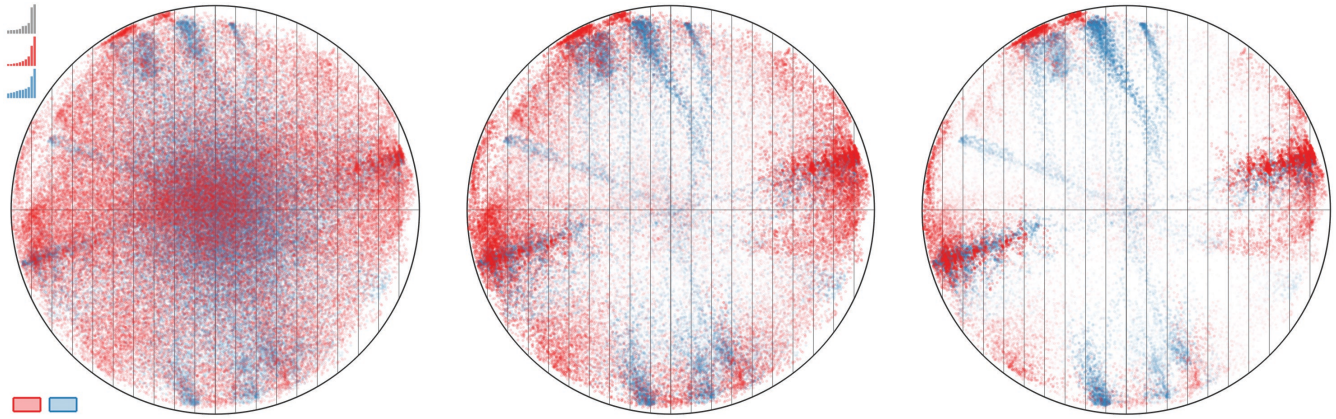
One side-effect of the projection in the s-CorrPlot is that a much larger area of the correlation sphere projects to locations in the center of the s-CorrPlot, rather than the outside boundaries. This can result in significant overplotting for large datasets with many variables. One method to combat overplotting is alpha-blending, i.e. adjusting the transparency of the plotted points, like in Figure 3(a). However, this method tends to emphasize dense regions towards the center of the plot, which are not necessarily reflective of structures in the dataset. As shown in Section 4.2 for points near the center, variables with arbitrary correlations to each other can map to the same spot. Since alpha-blending is in essence an approach to density estimation on the projected points, it can highlight artifacts not representative of structure in the data.

To more effectively highlight patterns and structure as it exists on the correlation sphere, we incorporate a density estimation technique to the s-CorrPlot. The density estimation measures the density of standardized variables on the correlation sphere as a preprocessing step. The density value is stored per standardized variable and encoded with the transparency in the s-CorrPlot, examples of which are shown in Figures 3(b-c).

The density at any location, $\mathbf{x}$, on the correlation sphere can be estimated with a kernel density estimator [Parzen, 1962, Silverman, 1986] on the set of standardized vectors $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$:

$$f(\mathbf{x}) = \frac{1}{p} \sum_{i=1}^{p} K_{n-2}(\mathbf{x} \cdot \mathbf{y}_i, h) \tag{12}$$

with $K_{n-2}$ the kernel and $h$ the kernel bandwidth parameter. Because the density we are estimating lies on a $(n-2)$-sphere, we use the $(n-2)$-dimensional von Mises-Fisher distribution [Sra, 2012, Banerjee and Dhillon, 2005], $K_{n-2}(z, h) = c_{n-2}(h)\, e^{hz}$, where $c_{n-2}(h)$ is the appropriate normalizing constant. The kernel bandwidth adjusts the emphasis on which features are highlighted.



(a) (b) (c)

Figure 3: This is the s-CorrPlot showing a different projection of the image patch dataset in 1. In (a), each variable is a point in the s-CorrPlot with some transparency, a prime example of overplotting in the center of the scatterplot. With density estimation in (b), we significantly reduce this overplotting and more effectively highlight clusters in the scatterplot. This plot uses a large bandwidth emphasizing spread-out clusters, but we can also use a smaller bandwidth as in (c) to more effectively highlight tightly clustered variables.

In Figure 3(b-c), we illustrate the effect of adjusting the kernel bandwidth.

Naive calculation of the density estimation results in a $O(p^2)$ computation, where $p$ is the number of variables. Using compact kernel functions, it is possible to speed this up to $O(p \log(k))$, where $k$ is the approximate number of neighbors required to contain most of the mass of the kernel for a given bandwidth $h$. In our implementation of the s-CorrPlot, we calculate the density estimation as a preprocessing step, storing for each variable its $k = 200$ nearest neighbors to enable interactive adjustment of the bandwidth. This leaves the core $O(p)$ projection algorithm unaffected during runtime, allowing multidimensional exploration to be interactive.

## 4.4 Comparison to Heatmaps

For a dataset with $n > 3$ individuals and $p > 3$ variables, all pairwise correlations cannot generally be encoded in two spatial dimensions. Heatmaps forgo a spatial encoding in favor of a color encoding for all pairwise relationships. This can lead to visually conflicting information since strongly correlated variables can be spatially far apart in the heatmap, depending on the method of seriation. The s-CorrPlot tackles distortion differently; instead of requiring all pairwise correlations be displayed, only a subset of the pairwise correlations are displayed exactly. The s-CorrPlot

encodes correlation exactly for any variable that lies on the projection plane and with very little loss of accuracy to all variables that are sufficiently close to the projection plane. Computationally, this spatial encoding reduces the cost from $n^2p$ to $2np$.

A major challenge for exploring correlation structure in heatmaps is seriation. Methods of seriation for heatmaps can be seen as an attempt to restore some of the lost spatial information by grouping strongly correlated variables close together in the heatmap (*i.e.* in a sequence of rows or columns). The heatmap depends on a good seriation for visual inference, but seriation is an expensive computation and for many datasets no single method will highlight all correlation structure. The s-CorrPlot shifts the challenge of seriation from computation to one of human pattern recognition. As we explain in Section 5, the s-CorrPlot can be combined with interaction and animation in order to view different projections of the data, allowing structure and shape to be seen within the data that is not present within a heatmap. For more details, please read Section 5 and watch the companion video in Supplemental Materials.

### 4.5 Comparison to Spatial Encodings of Variable Space

The geometric interpretation of Pearson's correlation results in a multidimensional representation of the data. A standard approach to visualize multidimensional datasets is to apply dimensionality reduction methods [Jolliffe, 1986, Cox and Cox, 1994, Tenenbaum et al., 2000, Belkin and Niyogi, 2003]. Dimensionality reduction methods aim to capture a lower-dimensional representation that preserves the geometric structure or statistical content of the multidimensional data. If the data is reduced to 2D, the data can be directly visualized through a scatterplot or combined with more sophisticated visualization systems [Seo and Shneiderman, 2005, Yang et al., 2007, Tatu et al., 2009, Turkay et al., 2012]. However, it is unclear if and how correlation is spatially encoded after applying such dimensionality reduction methods. Furthermore, the spherical geometry underlying Pearson's correlation can induce large distortions when *flattened* into 2D.

There are several other encodings which use this geometric interpretation of variable space to spatially encode correlation. Similar to the s-CorrPlot, these methods perform 2D projections of the data, but they differ in how they are interpreted and thus their resulting visual encoding. Each encoding suffers from different drawbacks which motivated the new design of the s-CorrPlot.

The *h*-plots approach [Corsten and Gabriel, 1976] encodes correlation strength based on an-

gular separation of points on the correlation sphere. In *h*-plots, the variables are geometrically transformed but not scaled to unit length. These variables are projected onto the first two principal components of the dataset and rendered as lines from the origin. A line's deviation from unit length indicates how well the angular separation represents the correlation between variables. For points close to the origin, the angle is not representative of correlation strength alone. A recent modification of this approach creates correlation diagrams with the angular separation of lines more directly representing correlation [Trosset, 2005]. However, for both approaches, comparing angles between the lines is difficult and possibly inconsistent [Talbot et al., 2012], and scaling to more than several dozen variables produces significant visual clutter.

The focused principal components approach (FPCA) [Falissard, 1999] uses radial distance to encode correlation between a variable of interest and all other variables. In contrast to the s-CorrPlot, this approach places a projection plane orthogonal to the variable of interest, which results in that variable projecting to the center of the FPCA plot. Consequently, the variable of interest is defined as the origin in the FPCA plot, with strongly correlated, *and* anticorrelated, variables projected near the origin, and variables with zero correlation projected at a radius of one. For datasets with $n > 4$, however, the FPCA method breaks down because the correlation sphere has dimensionality $> 2$. This results in multiple orthogonal directions to any 2D projection plane; in this case, two variables can project exactly to the center regardless of their correlation coefficient to each other, *i.e.* for $n > 4$ any value between the bounds on the correlation given by Equations 5 and 6 can be achieved. Thus, for $n > 4$, the FPCA interpretation only provides exact information about perfectly uncorrelated variables.

# 5 Interactive Exploration with the s-CorrPlot

As a secondary contribution, we designed the s-CorrPlot to incorporate both interaction and animation, unlike previous static correlation encodings discussed in Section 4.5. In doing so, we illustrate how the s-CorrPlot can be paired with multidimensional exploration techniques, in the spirit of existing systems that employ user-driven exploration [Swayne et al., 1998, 2003, Elmqvist et al., 2008].

We encourage our readers to watch the short companion video in Supplemental Materials to more easily understand the interactive exploration aspect. Below, we first outline related work in

multidimensional exploration. Next, we discuss the interactions employed for selecting different projections of the data. Lastly, we detail how we implemented the animation between projections.

## 5.1 Existing Multidimensional Exploration Techniques

To better grasp multidimensional spaces, methods such as projection pursuit [Huber, 1985, Friedman, 1987], the grand tour method [Asimov, 1985], and combinations of both [Cook et al., 1995], explore multidimensional space through sequences of 2D projections. Several mature systems that implement these techniques include xgobi [Swayne et al., 1998] and ggobi [Swayne et al., 2003], which provide animations and interactions to let the user explore the complete space of projections. The spatial encoding of correlation in the s-CorrPlot is also applicable within any of these existing systems. Our implementation of the s-CorrPlot was motivated by these exploratory techniques in order to create simple user-guided tours, or animations between projections.

## 5.2 User-Driven Exploration

The error bounds in Section 4.2 show that getting a complete and exact view of all pairwise correlation measures requires multiple projections; at most, this involves a projection for every variable. To mitigate this loss of information, our implementation of the s-CorrPlotemploys several simple aspects of user-driven exploration to help examine the space of possible projections. These interactions increase the effectiveness of the underlying spatial encoding of the s-CorrPlot.

For interaction, users drive the exploration of the multidimensional correlation sphere by selecting the variables $\mathbf{p}$ and $\mathbf{s}$ of interest. Note that these variables do not have to correspond to any of the variables in the dataset but can correspond to any point on the correlation sphere. After selecting a new variable, the s-CorrPlot is re-oriented through a continuous animation of a rotation between the current projection and a newly selected one, as described in Section 5.3.

In the s-CorrPlot, users can select $\mathbf{p}$ and $\mathbf{s}$ by clicking on data points in the scatter plot or from principal components of the data. The principal components [Jolliffe, 1986] help to find initial interesting projections. We display the principal components as bar charts for both the complete dataset and for categorical subsets of the data, such as the two groups in Figure 3(a). Each bar corresponds to a principal component and the height represents its corresponding eigenvalue, or the variance of the data in that direction, see top-left of Figure 3(a). By clicking on the individual

bars, the user can select to orient either **p** or **s** in the direction of that principal component.

Determination of a projection can be extended to include other data-driven measures of interesting projections, such as scagnostics [Tukey and Tukey, 1985, Wilkinson et al., 2005, Sips et al., 2009, Tatu et al., 2009]. Scagnostics establishes mathematical definitions for interesting features of multidimensional scatterplots. Traditionally, it is applied to a SPLOM for finding interesting structures in the individual scatterplots, but, for the s-CorrPlot, scagnostics would run on all of the resulting projections. Just like principal components, these scagnostic measures can then be used to define a data-driven projection.

### 5.3 Animating Between Projections

When a new projection is defined or selected, the current projection plane is swept across the correlation sphere to the new orientation, preserving the spatial encoding of correlation. There are a variety of methods for interpolating between projection planes [Wickham et al., 2011].

As with multidimensional exploration between projections, animating between planes results in structures, such as clusters of correlated variables, moving together (or apart) in 3D, giving the user a partial sense of the relationship of the standardized variables in the multidimensional space. Perceptually, the animation results in seeing "shape from motion" [Ullman, 1979].

In the s-CorrPlot, we interpolate across the vectors chosen for the projection. In addition, we orient the viewer by projecting the primary vector to a fixed location on the far-right of the s-CorrPlot and draw the gridlines vertically with respect to this primary vector. It is, however, more common in multidimensional exploration systems to interpolate between planes using fixed angular increments, thus providing constant speed of motion between planes. Cook et al. [2008] describe one such algorithm, with some cost to the preservation of the spatial encoding throughout the animation. An avenue for future work is to implement the s-CorrPlot within an existing multidimensional exploration system, while balancing a consistent spatial encoding.

## 6 Validation

In this section, we demonstrate and validate the capabilities of the s-CorrPlot with an example dataset and a case study with a biology collaborator. The collaborator is also a co-author on this

paper. An additional use-case, not described in this paper, is included in the companion video in Supplemental Materials. These datasets are contained within the open-source R-package, *scorr*.

## 6.1 Demonstrative Example: Image Patch Data

To compare the s-CorrPlot to existing methods and highlight visualization of correlation structures in large datasets, we construct example datasets with many variables that contain interdependencies between the variables. The datasets consist of all $9 \times 9$ patches of two images. Each image patch represents a variable containing 81 observations, *i.e.* the pixel values. The two test images we use are the *Barbara* and *House* images, shown in Figure 1(a), which are commonly used test images in image processing [Portilla, 2013]. Together, these two images produce over 130,000 variables. We visualize both datasets using the s-CorrPlot in Figure 1(b), where each variable is assigned a color corresponding to which image it comes from. We also show this dataset in the companion video in Supplemental Materials.

The s-CorrPlot can easily find continuous correlation structures, like the visible red bands corresponding to shifts, rotations, and gradients, which other visualizations cannot. In real datasets, this can highlight latent nonlinear dependencies between a set of variables. For example, the *House* image data at the bottom of Figure 1(b) shows three image patches from such a correlation band. In these image patches, there is a decreasing dark edge in the image patch as we move along the curve in the s-CorrPlot. Visual clutter and seriation would make such structures next to impossible to find in parallel coordinate or heatmap visualizations.

In Figure 1(c), we visualize the image datasets using the *ggplot2* implementation of parallel coordinates [Wickham, 2009], in variable space. Memory issues in this implementation prevented us from visualizing the complete datasets; so only 50k variables are plotted across the axes. The parallel coordinates plot is helpful in highlighting clusters, such as patches containing flat features as well as patches exhibiting few specific texture patterns, especially in the heavily textured *Barbara* image, but it does not assist with continuous correlation structures seen in the s-CorrPlot. Lastly, the visual clutter in parallel coordinates makes fine-grained analysis extremely difficult.

This simple example also exposes three major challenges for heatmaps. First, generating the 130k x 130k pairwise correlation matrix does not fit into the memory of a typical desktop computer, nor does a heatmap display of the matrix fit on a standard display. Second, interesting correlation

structures are difficult to interpret in a heatmap as a result of seriation. In Figure 1(d), we take a 22k subset of the *House* image patch data and for seriation choose to cluster via complete linkage; structures immediately visible in the s-CorrPlot are all but lost in the heatmap. And third, while the s-CorrPlot supports direct comparison of multiple datasets through color encoding, as in Figure 1(b), it is unclear how to perform this sort of comparison in a single heatmap display.

## 6.2 Case Study: Gene Expression in the Brain

Biologists often analyze the correlation of **gene expression** — how much a gene is turned on or off in a cell — across datasets to gain insights into gene functions and to infer novel relationships between genes [Seo and Shneiderman, 2002]. This analysis seeks to answer questions pertaining to the relationship of correlation between genes, especially how these relationships change over time, across species, or in the presence of disease.

We are working with a biologist at the University of Utah who is tackling similar questions, by studying genes that work together in the brain in order to uncover genetic influences on brain function, behavior, and disease. Using high-throughput sequencing, he measures the expression level of genes in specific brain regions, even to the detail of expression of **exons**, which are subparts of genes. These measurements are taken in different strains of mice, which form the observations in his dataset. The genes and exons are the variables he wants to correlate and study.

His typical study involves several dozen observations, and approximately 10,000 to 100,000 variables. The state-of-the-art approach for studying the correlation of gene expression is weighted gene co-expression network analysis (WGCNA) [Langfelder and Horvath, 2008, Oldham et al., 2008, Winden et al., 2009]. WGCNA uses the correlation or similarity of genes to construct a weighted network among all genes, and, using this network, genes that have a high degree of topological overlap are grouped together into gene modules. However, WGCNA was designed to support only 10,000 to 20,000 genes so does not scale to the size of datasets that our collaborator struggles to analyze.

At first, our collaborator explored 38,365 genes in two regions of the brain, with 22 observations, using the s-CorrPlot, shown in the top of Figure 4. Since each gene can exist in either brain region, this results in a combined total of 76,730 variables. The gene expression levels measured in brain region 1 are shown in red, and those in brain region 2 are shown in blue. He first looked at

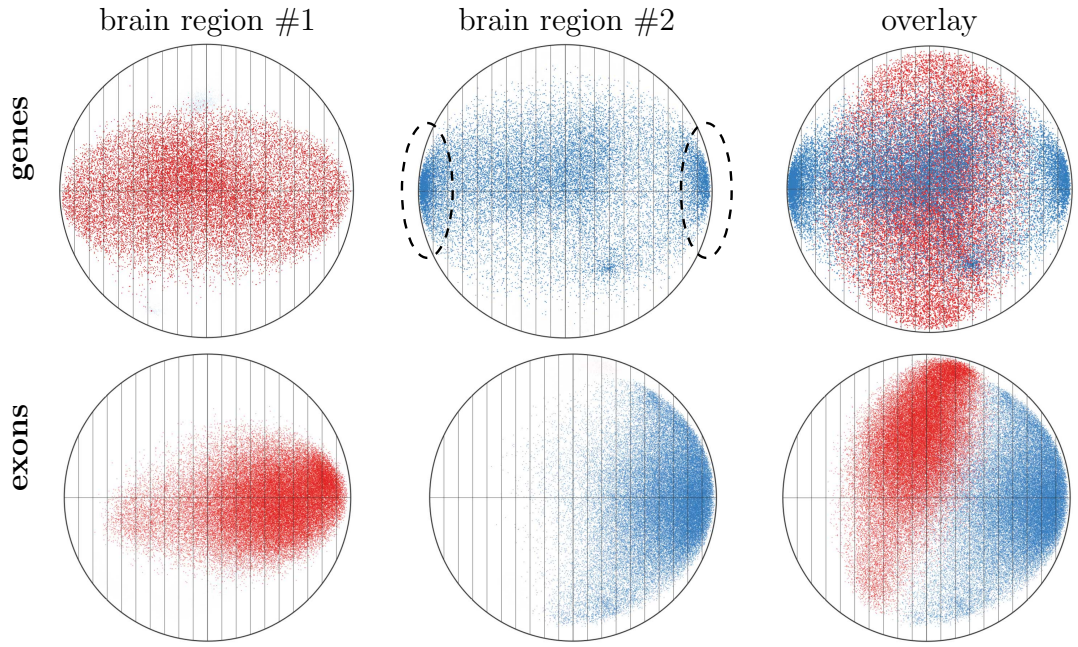



Figure 4: Two different biological datasets containing 76,730 (genes) and 120,000 (exons) variables, with 22 and 37 observations, respectively. For each dataset, genes and exons have been colored according to two different brain regions in which the expression levels were measured, resulting in separate and combined overlay visualizations. The s-CorrPlot highlights different patterns of correlation in each of these brain regions, indicating potentially significant differences in their biological processes.

just brain region 1 (red), orienting the s-CorrPlot using the first principal component for these variables — he noted that no strong clusters emerged. He then did the same for just brain region 2 (blue), and saw a significant grouping of correlated and anticorrelated points, shown in the dashed ovals. Overlaying the two brain regions confirmed interesting differences across the correlation of all genes between these two regions. The differences in the correlation structure of the data are anticipated to reflect differences in the cell types and mechanisms that regulate gene expression and the function of the two brain regions.

Using a different dataset, our collaborator visualized the expression levels of different exons in the same two brain regions, as shown in the bottom-half of Figure 4. This particular dataset contains 60,000 exons in each brain region, for a total of 120,000 variables, with each variable containing 37 observations. This is the first analysis of correlation at the exon level that our collaborator is aware of, perhaps due in part to the inability of existing tools to handle these large datasets. With the s-CorrPlot, our collaborator was able to interactively explore the many

exons and deduce that there are also region specific patterns at the exon level. He noted that the patterns in the exon data are significantly different than that for the data at the gene-level, indicating that differences in these brain regions could be described at a smaller scale than genes.

Taken as a whole, the differences in the patterns between the two regions of the brain are completely unknown and unexplored in our collaborator's field. These observations have prompted him to design follow-up computational studies and wet-lab experiments, fueled by hypotheses, which are formed by his use of the s-CorrPlot for correlation analysis. He commented: *"This is revealing new brain-region specific patterns in the data that we were completely unaware of. It offers the potential for deriving entirely new hypotheses about the functional relationships between genes in different brain regions that we can test experimentally."*

# 7 Conclusions and Future Work

Through a careful examination of the geometrical representation of Pearson's correlation, this paper derives a simple yet powerful approach to visualize correlation for large datasets. The s-CorrPlot presents an alternative to the commonly employed clustered heatmap for exploratory analysis of correlation; our proposed encoding scales better with increasing data size and provides finer details on the multidimensional correlation structure between many variables.

The spatial encoding in the s-CorrPlot suggest several interesting directions for future research. Exploring the cause of the difference between heatmaps, seriation, and the s-CorrPlot could lead to new insights on perception research for correlation and multidimensional structure. The spatial encoding makes it possible to visually overlay additional statistical information about variables, such as confidence intervals or other uncertainty measures, in the plot. The ability to encode additional information about the variables, using visual channels like color or shape, supports the integration of more sophisticated set-membership visualizations on the points.

Furthermore, by connecting the s-CorrPlot with a mature multidimensional exploration system, like ggobi, more techniques could be used to explore correlation structure, like the ground tour, projection pursuit, and scagnostics. Lastly, the use cases shown in the companion video in Supplemental Materials demonstrate that the s-CorrPlot is potentially useful in a wide variety of applications. In particular, an avenue to explore is to adapt the s-CorrPlot for outlier detection.

## ACKNOWLEDGEMENTS

The authors wish to thank Kristi Potter, Wei-Chao Huang, Tom Fletcher, and Kris Zygmunt for their feedback on this work. This work is sponsored in part by the Air Force Research Laboratory, the DARPA XDATA program, and the Office of Naval Research award N00014-12-1-0601. The content of the information does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred.

## SUPPLEMENTAL MATERIALS

**Code & Data:** An open-source R-package, *scorr*, implementing the s-CorrPlot. Includes the datasets shown in the paper, and a static version of the s-CorrPlot for those unable to compile the interactive tool. For more information, see `http://mckennapsean.com/scorrplot`.

**Video:** Demonstration of exploring correlation with the interactive s-CorrPlot visualization. (.mp4)

# References

J. Albers. *Interaction of Color*. Yale University Press, 2006. ISBN 9780300115956. URL `http://books.google.com/books?id=wN9o0OULXjIC`.

P. D. Allison. Testing for interaction in multiple regression. *American Journal of Sociology*, pages 144–153, 1977.

M. Ankerst, S. Berchtold, and D. Keim. Similarity clustering of dimensions for an enhanced visualization of multidimensional data. In *Information Visualization, 1998. Proceedings. IEEE Symposium on*, pages 52 –60, 153, oct 1998. doi: 10.1109/INFVIS.1998.729559.

D. Asimov. The grand tour: a tool for viewing multidimensional data. *SIAM J. Sci. Stat. Comput.*, 6(1):128–143, Jan. 1985. ISSN 0196-5204. doi: 10.1137/0906011. URL `http://dx.doi.org/10.1137/0906011`.

A. Banerjee and I. Dhillon. Clustering on the unit hypersphere using von Mises-Fisher distributions. *Journal of Machine Learning Research*, pages 1345–1382, 2005. URL `http://machinelearning.wustl.edu/mlpapers/paper_files/BanerjeeDGS05.pdf`.

M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15(6):1373–1396, 2003. ISSN 0899-7667.

J. Bertin. *Semiology of graphics*. University of Wisconsin Press, 1983. ISBN 0299090604.

S. J. Bohn, D. Payne, G. Nakamura, and D. Love. Analytics for massive heat maps. In *Proceedings of SPIE, the International Society for Optical Engineering*. Society of Photo-Optical Instrumentation Engineers, 2009.

W. S. Cleveland and M. E. McGill, editors. *Dynamic Graphics for Statistics*. Wadsworth, 1988.

W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984. ISSN 01621459. URL `http://www.jstor.org/stable/2288400`.

D. Cook, A. Buja, J. Cabrera, and C. Hurley. Grand tour and projection pursuit. *Journal of Computational and Graphical Statistics*, 4(3):155–172, 1995.

D. Cook, A. Buja, E.-K. Lee, and H. Wickham. Grand tours, projection pursuit guided tours, and manual controls. In *Handbook of data visualization*, pages 295–314. Springer, 2008.

L. Corsten and K. Gabriel. Graphical exploration in comparing variance matrices. *Biometrics*, 32(4):851–863, Dec. 1976.

T. F. Cox and M. A. A. Cox. *Multidimensional Scaling*. Monographs on Statistics and Applied Probability 59. Chapman and Hall, London, 1994.

A. P. Dempster. *Elements of continuous multivariate analysis*, volume 388. Addison-Wesley Reading, Mass., 1969.

M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. National Academy of Sciences*, 95(25):14863–14868, 1998. URL `http://www.pnas.org/cgi/content/abstract/95/25/14863`.

G. Ellis and A. Dix. Enabling automatic clutter reduction in parallel coordinate plots. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):717–724, Sept. 2006. ISSN 1077-2626. doi: 10.1109/TVCG.2006.138. URL `http://dx.doi.org/10.1109/TVCG.2006.138`.

N. Elmqvist, P. Dragicevic, and J.-D. Fekete. Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. *IEEE Transactions on Visualization and Computer Graphics (Proc. InfoVis 2008)*, 14(6): 1141–1148, 2008.

B. Falissard. Focused principal component analysis: Looking at a correlation matrix with a particular interest in a given variable. *Journal of Computational and Graphical Statistics*, 8(4):pp. 906–912, 1999.

J. H. Friedman. Exploratory projection pursuit. *Journal of the American Statistical Association*, 82(397):249–266, 1987. URL `http://yaroslav.hopto.org/papers/friedman-exploratory.pdf`.

M. Friendly. Corrgrams. *The American Statistician*, 56(4):316–324, 2002.

M. Hahsler, K. Hornik, and C. Buchta. Getting things in order: An introduction to the r package seriation. *Journal of Statistical Software*, 25(3):1–34, 3 2008. ISSN 1548-7660. URL http://www.jstatsoft.org/v25/i03.

L. Harrison, F. Yang, S. Franconeri, and R. Chang. Ranking Visualizations of Correlation Using Weber's Law. *IEEE Transactions on Visualization and Computer Graphics*, 2014. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6875978.

J. A. Hartigan. Printer graphics for clustering. *Journal of Statistical Computation and Simulation*, 4(3):187–213, 1975. URL http://www.informaworld.com/10.1080/00949657508810123.

S. Horvath and J. Dong. Geometric interpretation of gene coexpression network analysis. *PLoS computational biology*, 4(8):e1000117, 2008.

P. J. Huber. Projection pursuit. *The annals of Statistics*, pages 435–475, 1985.

A. Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1:69–91, 1985. ISSN 0178-2789. doi: 10.1007/BF01898350. URL http://dx.doi.org/10.1007/BF01898350.

I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 1986.

P. Langfelder and S. Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):559, 2008.

J. Li, J.-B. Martens, and J. J. van Wijk. Judging correlation from scatterplots and parallel coordinate plots. *Information Visualization*, 9(1):13–30, 2010. URL http://dblp.uni-trier.de/db/journals/ivs/ivs9.html#LiMW10.

I. Liiv. Seriation and matrix reordering methods: An historical overview. *Statistical analysis and data mining*, 3 (2):70–91, 2010.

M. C. Oldham, G. Konopka, K. Iwamoto, P. Langfelder, T. Kato, S. Horvath, and D. H. Geschwind. Functional organization of the transcriptome in human brain. *Nature neuroscience*, 11(11):1271–1282, 2008.

E. Parzen. On the estimation of a probability density function and the mode. *AMS*, 33:1065–1076, 1962.

W. Peng, M. O. Ward, and E. A. Rundensteiner. Clutter reduction in multi-dimensional data visualization using dimension reordering. In *Proceedings of the IEEE Symposium on Information Visualization*, INFOVIS '04, pages 89–96, Washington, DC, USA, 2004. IEEE Computer Society. ISBN 0-7803-8779-3. doi: 10.1109/INFOVIS.2004.15. URL http://dx.doi.org/10.1109/INFOVIS.2004.15.

J. Portilla. Test images. http://decsai.ugr.es/~javier/denoise/test_images/index.htm, 2013.

R. A. Rensink and G. Baldridge. The perception of correlation in scatterplots. *Comput. Graph. Forum*, 29(3): 1203–1210, 2010. URL `http://dblp.uni-trier.de/db/journals/cgf/cgf29.html#RensinkB10`.

J. L. Rodgers, W. A. Nicewander, and L. Toothaker. Linearly independent, orthogonal, and uncorrelated variables. *The American Statistician*, 38(2):133–134, 1984.

J. Seo and B. Shneiderman. Interactively exploring hierarchical clustering results [gene identification]. *Computer*, 35(7):80–86, 2002.

J. Seo and B. Shneiderman. A rank-by-feature framework for interactive exploration of multidimensional data. *Information Visualization*, 4(2):96–113, 2005.

B. W. Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

M. Sips, B. Neubert, J. P. Lewis, and P. Hanrahan. Selecting good views of high-dimensional data using class consistency. *Comput. Graph. Forum*, 28(3):831–838, 2009. URL `http://dblp.uni-trier.de/db/journals/cgf/cgf28.html#SipsNLH09`.

S. Sra. A short note on parameter approximation for von mises-fisher distributions: and a fast implementation of $i_s(x)$. *Computational Statistics*, 27:177–190, 2012. ISSN 0943-4062.

R. Staudte. *Seeing, Through Statistics: An Introductory Text with MINITAB Problems*. Prentice Hall, 1990. ISBN 9780724810796. URL `http://books.google.com/books?id=obCAAAAACAAJ`.

D. F. Swayne, D. Cook, and A. Buja. Xgobi: Interactive dynamic data visualization in the x window system. *Journal of Computational and Graphical Statistics*, 7(1):113–130, 1998.

D. F. Swayne, D. T. Lang, A. Buja, and D. Cook. Ggobi: Evolving from xgobi into an extensible framework for interactive data visualization. *Computational Statistics & Data Analysis*, 43(4):423–444, 2003.

J. Talbot, J. Gerth, and P. Hanrahan. An empirical model of slope ratio comparisons. *Visualization and Computer Graphics, IEEE Transactions on*, 18(12):2613–2620, 2012.

A. Tatu, G. Albuquerque, M. Eisemann, J. Schneidewind, H. Theisel, M. Magnor, and D. Keim. Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. In *Proceedings of IEEE Symposium on Visual Analytics Science and Technology (IEEE VAST)*, Atlantic City, New Jersey, USA, 10 2009.

J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(550):2319–2323, 2000.

M. W. Trosset. Visualizing correlation. *Journal of Computational and Graphical Statistics*, 14(1):1–19, 2005.

J. Tukey and P. Tukey. Computing graphics and exploratory data analysis: An introduction. In *Proc. Sixth Ann. Conf. and Exposition Computer Graphics*, pages 773–785. National Computer Graphics Association, 1985.

C. Turkay, A. Lundervold, A. Lundervold, and H. Hauser. Representative Factor Generation for the Interactive Visual Analysis of High-Dimensional Data. *IEEE Transactions on Visualization and Computer Graphics*, 2012. URL `http://www.uib.no/med/avd/ii/nyhetsbrev/MedViz/Publicationofthemonth/Turkay12Representative,Helwigdes2012.pdf`.

S. Ullman. *The interpretation of visual motion.* Massachusetts Inst of Technology Pr, 1979.

H. Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009. ISBN 978-0-387-98140-6. URL `http://had.co.nz/ggplot2/book`.

H. Wickham, D. Cook, H. Hofmann, and A. Buja. tourr: An r package for exploring multivariate data with projections. *Journal of Statistical Software*, 40(2):1–18, 2011.

L. Wilkinson and M. Friendly. The history of the cluster heat map. *The American Statistician*, 63(2):179–184, 2009.

L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics. In *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on*, pages 157–164. IEEE, 2005.

K. D. Winden, M. C. Oldham, K. Mirnics, P. J. Ebert, C. H. Swan, P. Levitt, J. L. Rubenstein, S. Horvath, and D. H. Geschwind. The organization of the transcriptional network in specific neuronal classes. *Molecular systems biology*, 5(1), 2009.

B. Wong. Points of view: Color coding. *Nat Methods*, 7(8):573, Aug. 2010. doi: 10.1038/nmeth0810-573. URL `http://dx.doi.org/10.1038/nmeth0810-573`.

J. Yang, D. Hubball, and M. Ward. Value and Relation Display: Interactive Visual Exploration of Large Datasets with Hundreds of Dimensions. *IEEE Transactions on Visualization and Computer Graphics*, 2007. URL `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.127.2545&rep=rep1&type=pdf`.