s-CorrPlot: Encoding and Exploring Correlation

Sean McKenna*
University of Utah

Miriah Meyer
University of Utah

Christopher Gregg
University of Utah

Samuel Gerber
Duke University

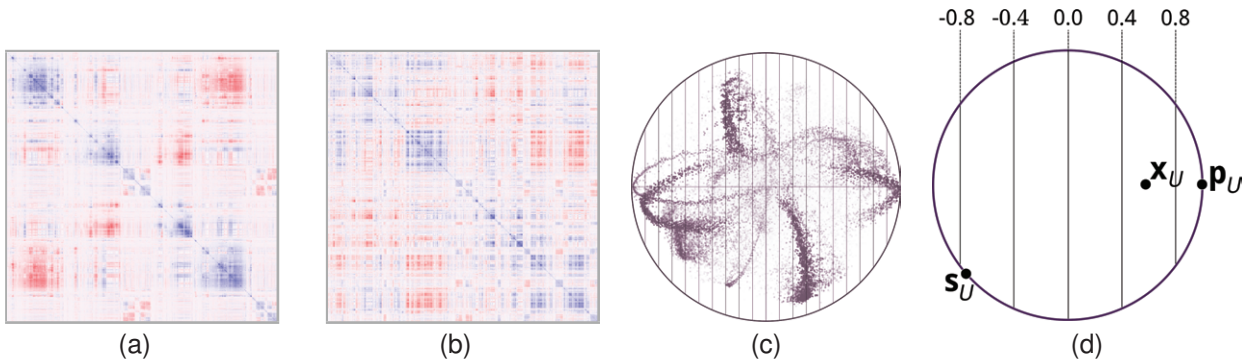


Figure 1: Visualizations of correlation for a dataset containing 22,000 variables. The left two images show the correlation coefficients using a heatmap, clustered with (a) average linkage, and (b) complete linkage. The visible patterns in the heatmap are highly dependent on the clustering algorithm. In (c), our novel s-CorrPlot spatially encodes correlation coefficients, highlighting very different multidimensional structures in the data. The s-CorrPlot effectively encodes correlation through gridlines as in (d) for a specific variable of interest, **p**.

1 Introduction

The degree of correlation between variables in a dataset is a key quantity in many data analysis applications. Correlation establishes the similarity between many variables and is often used by analysts as an exploratory tool to help form new hypotheses and sift through vast amounts of data in a variety of different domains.

As the amount of available data continues to expand in these fields, visualizing correlation becomes challenging because standard visualization techniques for exploring correlation lack the capacity to deal with these increasingly large data sets. Both scatterplot matrices (SPLOMs) [1] and parallel coordinates [4] rely on a visual determination of correlation, but have been shown to suffer underestimation effects [6] and screen space limitations [7]. Clustered heatmaps [12], on the other hand, visualize correlation through pairs of variables using color, as shown in Figure 1(a) and (b), but they have a number of limitations: pair-wise correlation computations grow quadratically with the number of variables; and accurate evaluation of correlation values is all but impossible due to the relative nature of color perception [13].

These limitations are addressed by our novel technique: the *spatial correlation scatterplot*, or the **s-CorrPlot**, as shown in Figure 1(c). The *s-CorrPlot* scales to hundreds of thousands of variables, both computationally and visually. The s-CorrPlot encodes correlation exactly for selected variables. The encoding is derived from the geometric interpretation of correlation as variables on a multidimensional sphere [5]. Projecting this sphere into two dimensions permits encoding correlation between variables exactly for a subset of the pairwise correlation coefficients. This exact encoding combined with interactive selection of different projections enables exploration of correlation structures in large data sets.

Two other methods exploit the geometric nature of Pearson's correlation for a spatial encoding. However, these methods only provide static visualizations. The first method, *h*-plots [2], encodes correlation through angular separation; as such, it suffers from the known issue of comparing angles between slopes of lines [10]. The second method is FPCA [3], which uses radial distance to encode correlation. While very similar to the s-CorrPlot, FPCA suffers from issues in multidimensional datasets, since then it only provides exact information about perfectly uncorrelated variables.

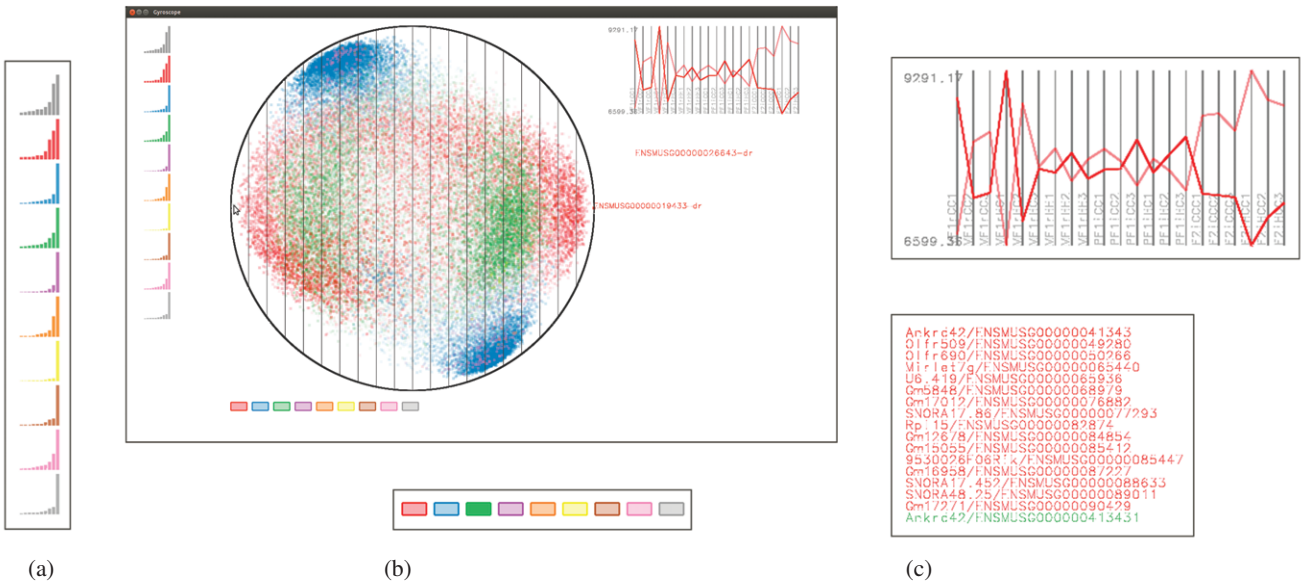
*e-mail: sean@cs.utah.edu

2 s-CorrPlot: Proof-of-Concept

We built a prototype of the s-CorrPlot as part of a linked-view proof-of-concept visualization, as shown in Figure 2. This prototype uses the overview + detail design pattern [8], with the s-CorrPlot providing a high-level view of the underlying data. In the s-CorrPlot, the color of the points represents an optional categorical label associated with each variable. Label buttons also serve as interactive legends, as shown in Figure 2(b).

The user can interactively select any variable, causing that variable to become the new primary variable of interest, **p**. This selection defines a new orientation of the projection plane, and thus the s-CorrPlot animates to the new projection. Then, the selected variable is placed to the far-right side of the projection plot. Vertical grid lines indicate the correlation values to that primary variable of interest, as illustrated in Figure 1(d). Selected secondary variables of interest, **s**, are placed in the lower hemisphere of the plot, and together these variables, **p** and **s**, define the projection. Alternatively, the s-CorrPlot supports selection of principal components. This aids the user both in finding potentially interesting projections, and to spread out variables in the plot so they can be more easily selected. Shown in Figure 2(a), interactive barcharts showing the eigenvalues from a PCA for each label let the user select associated principal components in the dataset.

The s-CorrPlot is linked to two detail views, as in Figure 2(c). At the top is a parallel coordinates plot where each axis represents an observation in the dataset, and the plotted lines represent variables. Variables are plotted in this view when they are hovered over or selected in the s-CorrPlot, which enables fine-grained analysis of multiple variables across all of their observations. Beneath this is a list, providing further details for currently selected variables.



(a) (b) (c)

Figure 2: The s-CorrPlot is linked to several other views for exploration: (a) interactive bar graphs of the eigenvalues from a PCA on the labeled subsets of data, which allows the user to select associated principal components for defining these projections; (b) label buttons for enabling or disabling subsets of variables; (c) and a parallel coordinates plot, with observations as the axes, where user-selected variables of interest get plotted as lines, coupled with a list of variables found in the projection plot near the cursor.

3 Open-Source Code and Evaluation

The proof-of-concept is implemented within the statistical computing environment R through an open-source, package *gyroscope*, available at `http://mckennapsean.com/scorrplot`. The website contains details on how to install and use the package, which includes both example data and scripts.

We have evaluated the correlation encoding through several use-cases and case studies. One of these use-cases, seen in our three-minute video, highlights critical outliers from a New York City subway dataset. In a hackathon, we used the s-CorrPlot to replicate several key findings from a recent *Nature* paper that explored clustering regions of the brain based on their correlation of gene expression. Lastly, we have conducted a case study on the tool by deploying it in a biology lab, where our collaborators study the correlation of gene expression to uncover genetic influences on brain function, behavior, and disease.

4 Conclusions and Future Work

To overcome the challenges of existing methods for visualization of correlation, we have illustrated the use of a new, interactive, spatial correlation scatterplot that utilizes a geometric interpretation of correlation. We found that, through a combination of data-driven and user-driven selections, collaborators were able to find interesting projections and form new hypotheses from their data. We are currently working on explicating the mathematical details behind the s-CorrPlot and the details of our evaluation approaches through a future publication.

There are also many potential directions for the s-CorrPlot. The s-CorrPlot can be enriched with various multidimensional techniques, such as scagnostics [11] as additional data-driven projections for the s-CorrPlot. This concept of data-guided exploration is by no means new. The de facto tool for doing such a multidimensional analysis is ggobi [9], and the s-CorrPlot encoding could also be embedded inside ggobi. Nevertheless, some of the initial interactions and features we incorporated in our proof-of-concept could also prove useful in these existing frameworks for aiding the exploration of correlation for many variables.

References

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

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

[3] 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.

[4] A. Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1:69–91, 1985.

[5] J. Lee Rodgers and W. A. Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, 1988.

[6] 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.

[7] 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.

[8] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proc. IEEE Symp. Visual Languages (VL)*, page 336, Washington, DC, USA, 1996. IEEE Computer Society.

[9] 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.

[10] 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.

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

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

[13] B. Wong. Points of view: Color coding. *Nat Methods*, 7(8):573, Aug. 2010.