SCALE-INVARIANT GEOMETRIC DATA ANALYSIS

Marina Girgis | Max Robinson | Institute for Systems Biology (ISB)

DEFINTION

SIGDA is a data analysis method designed to identify and to visualize the relation within a complex set of quantitative data.

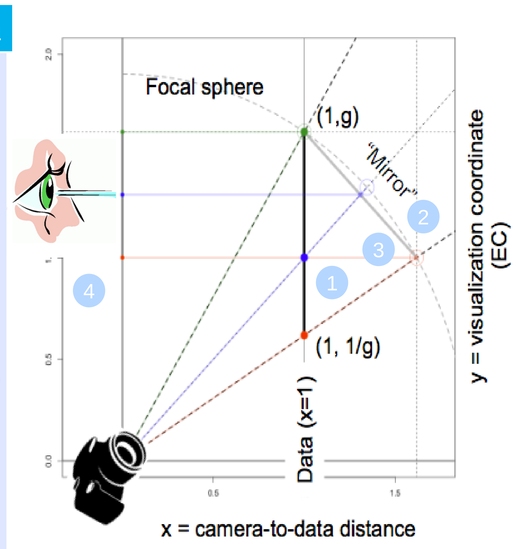
Procedure

- SIGDA preserves a complete statistical model for the data utilizing mathematical techniques called singular value decomposition and conformal transformations to view projections of the data on a new set of dimensions.

Graphic Insight

The Data Camera

1. SIGDA arranges the data to fit within the view of the camera.
2. It then projects it onto a spherical mirror so that all data points correspond to a place on a sphere.
3. It flattens the sphere
4. It reflects the overlapping row and column points back to the view plane.



Computation

- SIGDA takes the data matrix 'A' and decomposes it as follows:
- $A = D_r W D_c$ where D_x is a notation for a diagonal matrix with diagonal matrix x. Now, r is scaling vector of the rows and c is a scaling vector of the columns. Finally, W is a rotational matrix.
- It then applies singular value decomposition (SVD) to matrix W:
- $W = SD_s V^T$ where S contains the row eigenvectors and V^T contains the column eigenvectors. D_s is a diagonal matrix with the singular values of W.

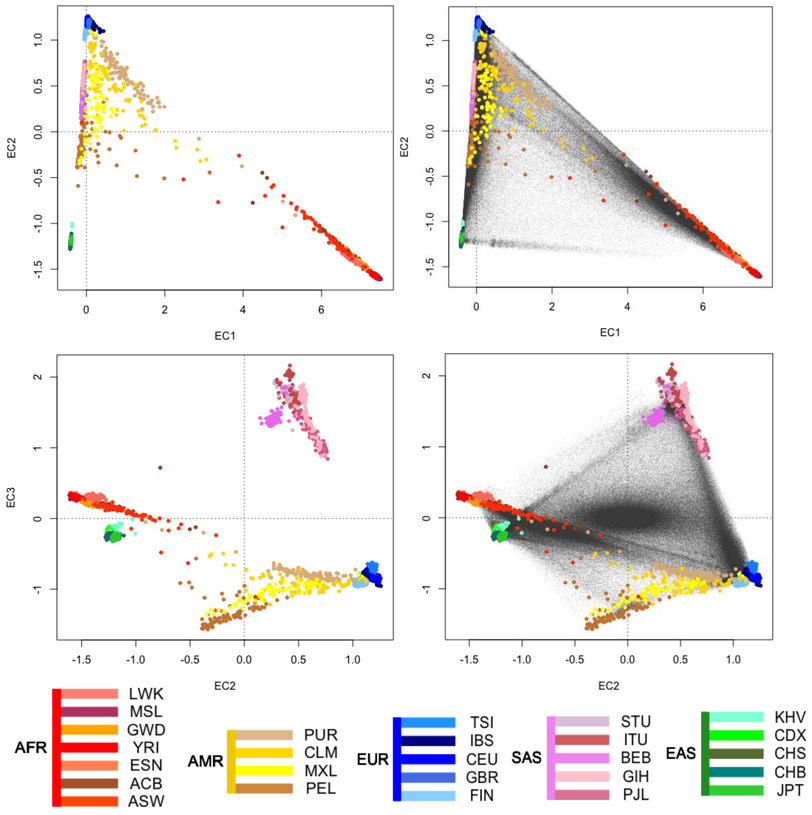
$$\underset{m \times n}{A} = \underset{m \times m}{\begin{bmatrix} r_1 & & \\ & \ddots & \\ & & r_m \end{bmatrix}} \underset{m \times n}{W} \underset{n \times n}{\begin{bmatrix} c_1 & & \\ & \ddots & \\ & & c_n \end{bmatrix}} \quad W = \underset{m \times m}{S} \underset{m \times n}{\begin{bmatrix} s_1 & & \\ & \ddots & \\ & & s_m \end{bmatrix}} \underset{n \times n}{V^T}$$

Experiment

The 1000 Genomes Project

- A famous project started in 2008 to identify genetic differences between individuals. It looks at the genomes of 2500 people from 26 different population groups.
- We applied SIGDA to a matrix of 2504 rows representing people, and 1247408 columns representing single nucleotide variants (SNVs). In other words, we analyzed about 3 billion variants.

Results



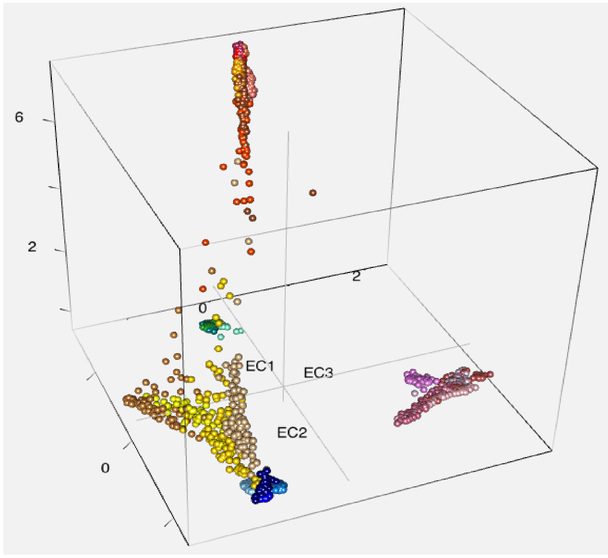
Data / Observations

- SIGDA reflects the linear patterns in the data accurately
- The correlations within and between rows and columns are visible

Institute for Systems Biology

STAR

3D Visualization

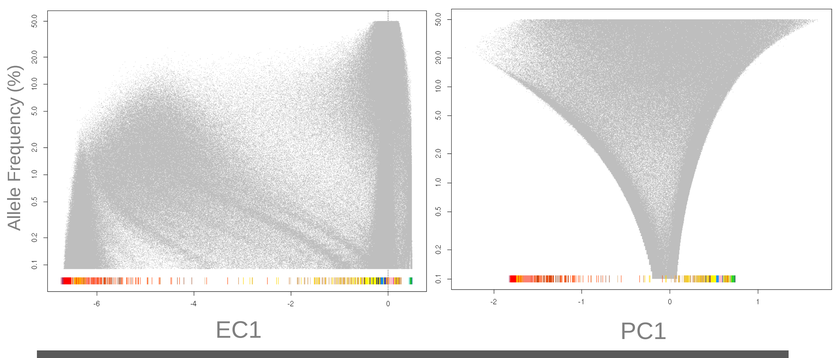


Conclusion

- SIGDA has shown to be more informative than other data analysis methods:

1. Below, we plot the first axis of variation against scale (the percent frequency of the SNVs) with SIGDA and Principal Component Analysis (PCA). We notice SIGDA finds independent population axes that are consistent at all allele frequency (throughout human history) while PCA finds one that depends on allele frequency.
2. We can isolate a group of nucleotide variants specific to a group of individuals using SIGDA.

SIGDA PCA



Acknowledgments

The 2018 STEM Teacher and Researcher Program and this project have been made possible through support from Chevron (www.chevron.com), the National Marine Sanctuary Foundation (www.marinesanctuary.org), the California State University Office of the Chancellor, and California Polytechnic State University, in partnership with Institute for Systems Biology.

CAL POLY SAN LUIS OBISPO CSU Chevron NATIONAL MARINE SANCTUARY FOUNDATION