Images of BoHeMIa: Optimizing Sample Size Using Visualization

Miller Christen[1,2], Leslie Spahr[1], and Neil Calkin[1]

[1] Department of Mathematical and Statistical Sciences, Clemson University, SC
[2] Department of Physics and Astronomy, Clemson University, SC

What are Bohemian Matrices?

- BOunded HEight Matrices of Integers are matrices whose entries are sampled from a fixed set of integers called a **population**.
- The integers can be **Gaussian**, having the form a + bi, where b is the imaginary component.
- Matrices sampled from the same population are collectively called a **family**.
- **Unit Upper Hessenberg (UUH)** matrices have the additional constraint that the first subdiagonal consists of all 1's, and the entries below it are all 0's. (Fig. 1)
- It is noted that in this study, we will focus primarily on the behavior of UUH matrix families. This is merely one of several Bohemian Matrix structures.[1]

$$\begin{bmatrix} n_1 & n_2 & n_3 & n_4 & n_5 \\ 1 & n_6 & n_7 & n_8 & n_9 \\ 0 & 1 & n_{10} & n_{11} & n_{12} \\ 0 & 0 & 1 & n_{13} & n_{14} \\ 0 & 0 & 0 & 1 & n_{15} \end{bmatrix}$$

$$\text{for } n_\alpha \in \{population\}$$

Fig. 1: Unit Upper Hessenberg matrix structure

Methods

Iterating through sample sizes of the same dimension from the known full dataset image…

- Controls: Compute the full dataset image at standard resolution and generate images of random sample sizes (Fig. 4).
- Test Similarity: We perform singular value decomposition (SVD) and compare angles, θ, between singular values of each sample plot and the full image to measure similarity.

$$A = U\Sigma V^T \qquad \theta = \cos^{-1}\left(\frac{\Sigma_s \cdot \Sigma_{full}}{\|\Sigma_s\| \cdot \|\Sigma_{full}\|}\right)$$

- Angle Behavior: This terminates once the angle is within a certain threshold. This threshold was chosen by looking at the convergence of the angle over our given samples to be 2.5 degrees (Fig. 5).

Determining whether our randomly generated samples were consistent….

- Bootstrapping: One hundred subsamples of various sizes were taken from a pool of 10^8, ranging from 1,000 to 1,000,000.
- Comparison: These subsamples were compared using the same SVD method, and the mean and standard deviation were plotted for each sample size (Fig. 3).
- Validation: This indicates samples above 10,000 matrices show < 1-degree deviation, implying any random sample of this size represents the population with equal accuracy.

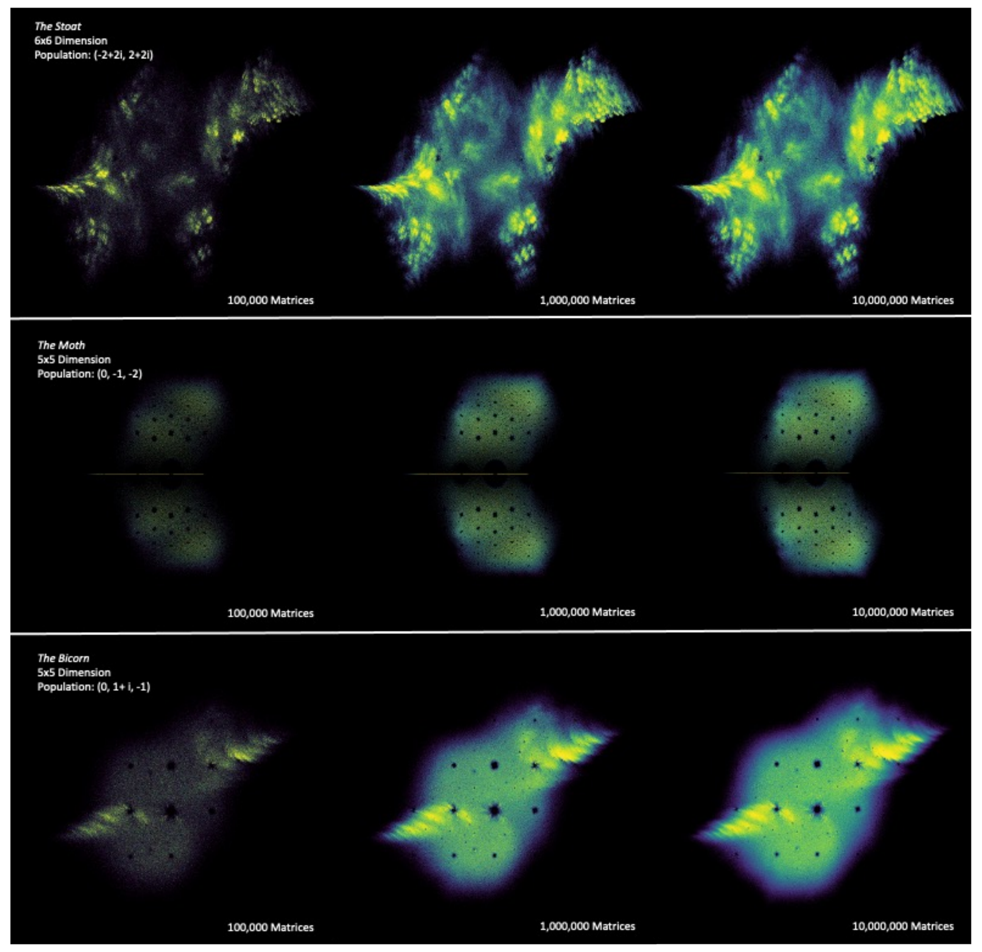

Fig. 4: A progression of various sample sizes from our controlled matrices

Visualizing Bohemian Matrix Families

- Matrix families are visualized by calculating the **eigenvalues** of every possible matrix in the family and then creating a **density plot** of those values in the **complex plane**.
- These plots reveal interesting patterns in the data, like spirals, black holes (usually around integer values), and fractal-like structures. It is noted that at higher dimensions, some of these significant structure patterns disappear or blur out. The reasoning behind this is an ongoing question we welcome onlookers to explore.

The Problem: Realistic Computation Time

- *The Stoat* (Fig. 2) is a UUH family with population {-2+2i, 2+2i}. In these examples, the highest-density eigenvalues are shown in yellow. As the dimension expands, the computation time for eigenvalue densities quickly grows beyond realistic time. This poses the issue that full datasets of possible eigenvalues for larger matrices are too expensive to compute.
- To address this issue, our research has focused on discovering a sample size that reliably preserves the structures and other visual information of the full dataset plots.

Dimension:	5x5	6x6	7x7
Computation time:	4.68 seconds	4.22 minutes	9.95 hours

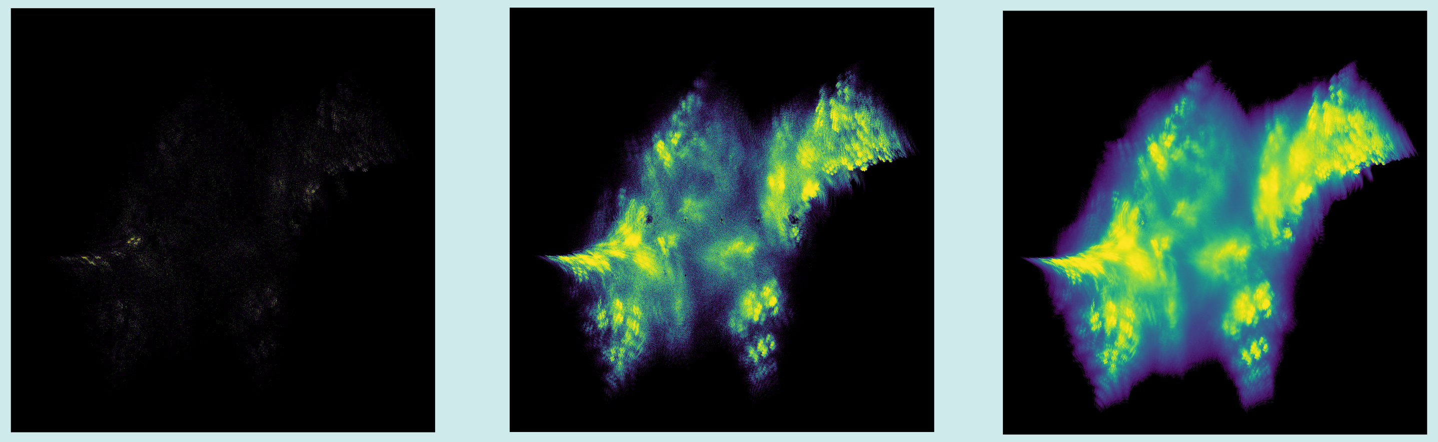
Fig. 2: A series of density plots of the UUH family with entry population {-2+2i, 2+2i}in different dimensions.

Results

Initial findings suggest that a sample size of 10^7 matrices suffices to approximate the true density plot across families with dimensions 5 or 6.

We use the confidence found in our methods above that our random sample is representative of all samples of that given size (Fig. 3). We also have reason to believe our angle threshold is within a good range.

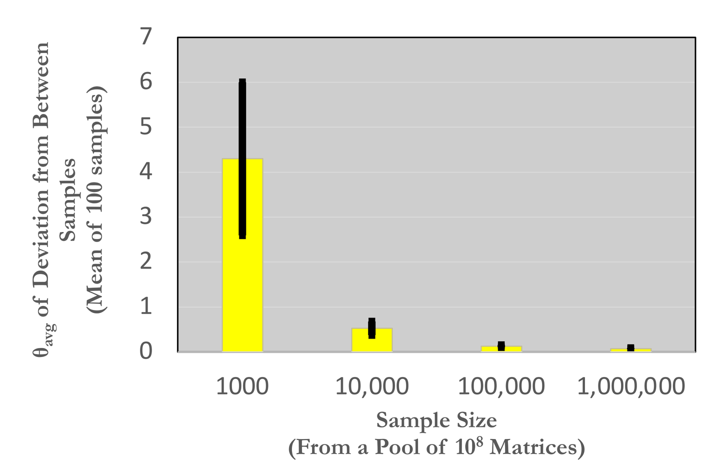

Fig. 3: Graph of the average angle of SVD deviation between images at different sample sizes.

We then apply these methods to larger dimensions where the full dataset is not computationally feasible, comparing each sample to the one prior. The angle threshold was adjusted to account for larger numbers and converged within 10 degrees (Fig. 6).

Fig. 5

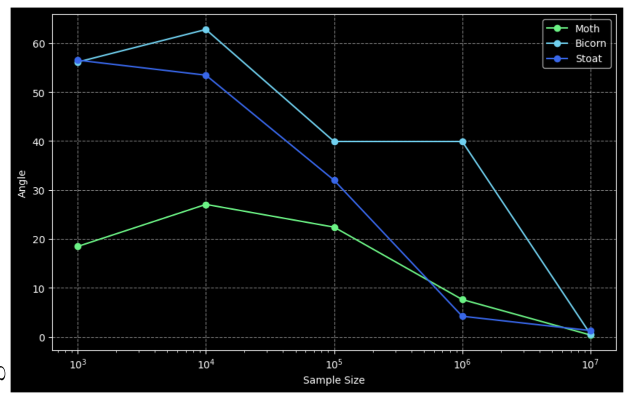

Fig. 6

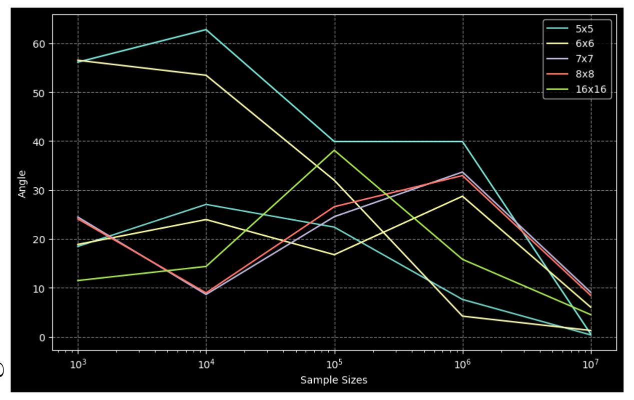

Fig. 5-6: Plot of angles between various sample sizes and the full dataset (left) and prior sample (right). The angle converges as the dominant patterns of the sample image stabilize.

Conclusion & Discussion

The sample size we concluded to be sufficient for higher dimensions of UUH matrices was consistent among a range of dimensions. It remains that 10^7 preserves the patterns and structure of the eigenvalue density plots. This is an important step in the exploration of bohemian matrices, as it opens the door to larger dimensions at a reasonable computational cost. We can now confidently compute eigenvalue density plots of sample sizes and study their structures as an accurate representation of the true dataset.

As this is a relatively new field, several questions remain. For example, why do some patterns disappear at higher dimensions? The image of the moth seems to lose the holes in its wings past dimension 10. We speculate that with the UUH structure, numerical artifacts may diminish as the upper triangular section grows in size.

Additional ongoing research questions can be asked:
- What might this look like mapped onto different structures?
- What might this method look like in higher-resolution images?

References

[1] Calkin, N. J., Chan, E. Y. S., & Corless, R. M. (2023). Bounded Height Matrices of Integers (Bohemian Matrices). In *Computational Discovery on Jupyter* (pp. 105–150). Philadelphia, PA: Society for Industrial and Applied Mathematics.

[2] Corless, R., & Thornton, S. (n.d.). Retrieved from http://www.bohemianmatrices.com/

Acknowledgements

We thank our advisor, Dr. Neil Calkin (Clemson University), as well as Dr. Robert M. Corless (University of Western Ontario) and Dr. Eunice Y.S. Chan (Chinese University of Hong Kong), who co-authored Computational Discovery on Jupyter.