Filtering Multivariate Data Through Convex Floating Bodies

Saitheeraj Thatigotla[1], Joseph Anderson[2]
[1]University of Tennessee, Knoxville, [2]Salisbury University

Abstract

Heavy-tailed data presents issues with unsupervised learning algorithms such as Principal Component Analysis (PCA) and Independent Component Analysis (ICA) that depends on having finite lower moments. Heavy-tailed data, by nature, does not necessitate finite first (mean) and second (variance) moments. In this project, we looked at applying the convex floating body as a filter on a Cauchy distribution-skewed data for PCA and ICA and the effectiveness of that filter.

Background

Heavy-tailed data has many applications in real world scenarios and result naturally in many complex systems. The central limit theorem, which states as the number random variables increases, the mean approaches a fixed value, is violated as the data could have a non-finite or infinite variance or mean (Markovich). Thus, relying on algorithms of methods that depend on finite lower moments would not be using accurate representations of the data.

PCA is a dimensionality reduction technique to find the change-of-basis vectors that will change data to a lower dimension that represents the most important dimensions to least important dimensions as the axes and with each dimension orthogonal to each other. There are two main approaches: One way is to use eigenvalue decomposition, EVD, of the covariance matrix to get the eigenvectors and the other is to use singular value decomposition, SVD, of the data matrix to get the eigenvectors (Shlens).

ICA is a solution to the Blind Source Separation (BSS) problem. A classic example is the Cocktail Party Problem (CPP) which is based on the following scenario: suppose there is a cocktail party with multiple people having conversation. If there are multiple microphones placed around the room, the microphones will pick up a time series signal that is an amalgamation of the difference speakers (or source signals) which themselves vary based on the distance of the speakers to the microphones. If these are the time signals:

$$x_1(t) = a_{11}s_1 + a_{12}s_2$$
$$x_2(t) = a_{21}s_1 + a_{22}s_2$$

We can represent the the mixed result and source signals as column vectors and the a terms as a mixing matrix, the previous equations can be represented by the following:

x = As

The goal of ICA is then to find a demixing matrix, the inverse of **A,** to recover **s.** An important note to make is that the sources are assumed to be independent and at most only one source is Gaussian. That is due to the fact that the joint density of gaussian sources would be symmetric and will not include any information about the **A** mixing matrix (Hyvärinen).

A convex body is any body where if a line segment is drawn between any points within the body, the line segment itself will also be in the body.

Definition. If we let **K** be a convex body in $\mathbb{R}$ and $\delta \geq 0$, the floating body is the intersection of halfspaces whose hyperplanes cut off a set of volume δ of **K**. The floating body exists for a convex set as long as $\delta > \text{vol}_d(K)/2$. Otherwise the convex floating body would be empty (Nagy et. al).

Goals

The initial goals of our project for this summer was to compare the floating bodies of different types of distributions including heavy-tailed Cauchy-skewed distributions, look at the effectiveness of the floating body for PCA, ICA, and potentially other algorithms, and parallelizing code written in the Julia programming language.

Methods

We used the cocktail party dataset from the Ravel Corpora \cite{Alameda-Pineda et. al.). We made various different sizes of the data to make it quicker to run and test on the computers (sizes of 10,100,1000,10000,100000,etc.) and reduced the dimensions from 4 (4 columns) to 2 (2 columns). We also generated random data from a Cauchy distribution and added it to the dataset to skew it. Then, we generated a square mixing matrix **A** by generating random column vectors from a Gaussian distribution which all have a unit length of 1 and multiplied it with my sample matrix to generate the resultant matrix **x**.

Next, we identify what points are inside and outside the polar body of the original data. From there, we also had a function that we used to generate points on the boundary of the polar body for data visualization.

In order to generate the primal body, we first generate random samples of points within the polar body. After that, we generate the supremum of the points within the polar body for every direction. By inverting the supremum, we get the boundary of the primal boundary.
After finding the primal body, we then applied to the PCA and ICA algorithms in different ways. For ICA, we used the fastICA algorithm using the pow3 contrast function (g function is x^3). And for PCA, we found the eigenvectors of the covariance matrix, from data within the floating body and uniform data from the floating body, and multiplied them with the **x** resultant matrix.

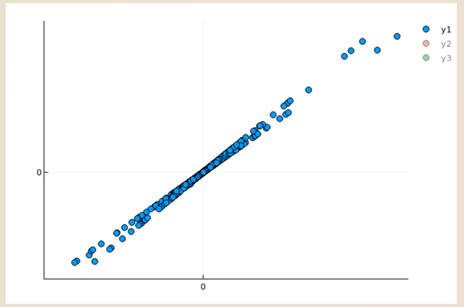

Fig 1. This is what the resultant matrix **x** looked like.

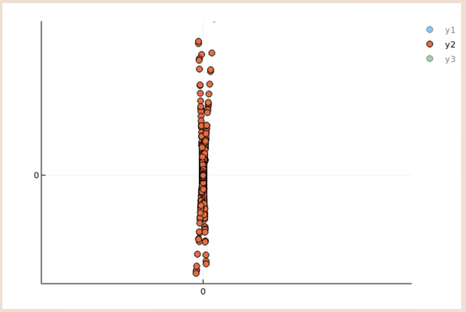

Fig 2. Data inside floating body

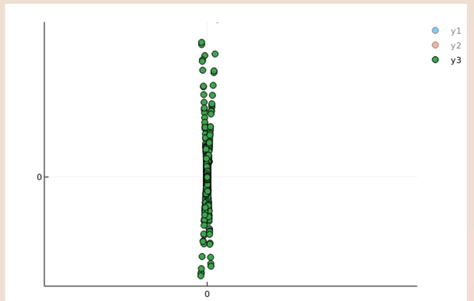

Fig 3. Uniform data from floating body

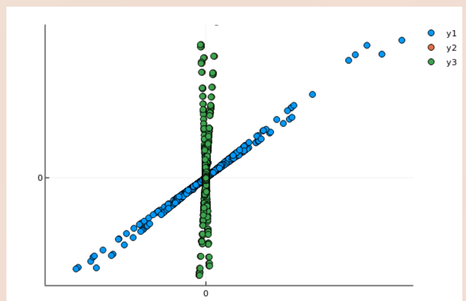

Fig 4. All 3 of the previous figures superimposed

Conclusions

In the end, the current results are inconclusive. It seems the floating body has transformed the data for PCA, but more work still needs to be done to verify the results so far. In regards to ICA, we ran into issues with the fastICA implementation having trouble converging for the cost function. More time and work would be useful in correcting the error and being to test if the original signals can be recovered. We also were not able to test other distribution besides Cauchy and did not get to the parallelization of the code. We still plan on continuing working on the project after the REU program ends to achieve more conclusive results.

References

Xavier Alameda-Pineda et al. “RAVEL: An Annotated Corpus for Training Robots with Audiovisual Abilities”. In:Journal on Multimodal User Interfaces 7.1-2 (2013), pp. 79–91. URL : http://hal.inria.fr/hal-00720734/en.

Joseph Anderson et al. “Heavy-Tailed Analogues of the Covariance Matrix for ICA”. In: CoRR abs/1702.06976 (2017). arXiv: 1702.06976. URL : http://arxiv.org/abs/1702.06976.

Joseph Anderson and Luis Rademacher. “Efficiency of the floating body as a robust measure of dispersion”. In: 2019.

Gari Clifford. “Chapter 15 - BLIND SOURCE SEPARATION: Principal & Independent Component Analysis”. In: Biomedical Signal and Image Processing—HST-582J/6.555J/16.456J. MIT OpenCourseWare. Cambridge MA, 2008.

Aapo Hyvärinen and Erkki Oja.Independent component analysis: a tutorial Tech. rep. 1999.

N. Markovich. Nonparametric Analysis of Univariate Heavy-Tailed Data: Research and Practice. Wiley Series in Probability and Statistics. Wiley, 2008. ISBN : 9780470723593. URL : https://books.google.com/books?id=j8V4DBHfnzcC.

Stanislav Nagy, Carsten Schuett, and Elisabeth M Werner. “Data depth and floating body”. In: arXiv preprint arXiv:1809.10925 (2018).

Jonathon Shlens. “A Tutorial on Principal Component Analysis”. In: CoRR abs/1404.1100 (2014). arXiv: 1404.1100. URL :http://arxiv.org/abs/1404.1100.