Effects of Lossy Compression on Correlation Workflows

Cooper Sanders and Jon Calhoun (Advisor)

Holcombe Department of Electrical and Computer Engineering - Clemson University

Introduction

A correlation matrix is a matrix representation of a network in which each node is a scalar variable, and each edge is a correlation coefficient of the two connected nodes. Workflows such as Knowledge Independent Network Construction (KINC) heavily rely on the computation of such matrices.

Hardware Limitations

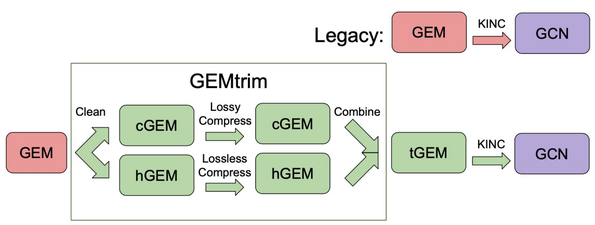
Downstream matrix operations of correlation workflows take too much memory to run all at once.

Recent technological advances provide massive amounts of data.

Compression will provide a solution to memory challenges.

$$\frac{(A-\bar{A})\cdot(A-\bar{A})^{\mathrm{T}}}{\sqrt{\sum(A-\bar{A})^2}\cdot\sqrt{\sum(A-\bar{A})^2}^{\mathrm{T}}}$$

Example KINC workflow with compression



Goal: *Explore how compressing input matrices effects the average percent error in the output matrices for correlation testing.*

Experimental Setup

Compressor Configurations

sz_abs [1e-1, 1e-2, 1e-3, 1e-4, 1e-5]
sz_rel [1e-1, 1e-2, 1e-3, 1e-4, 1e-5]
sz_psnr [70, 80, 90, 100, 120]
zfp [1e-1, 1e-2, 1e-3, 1-e4, 1e-5]

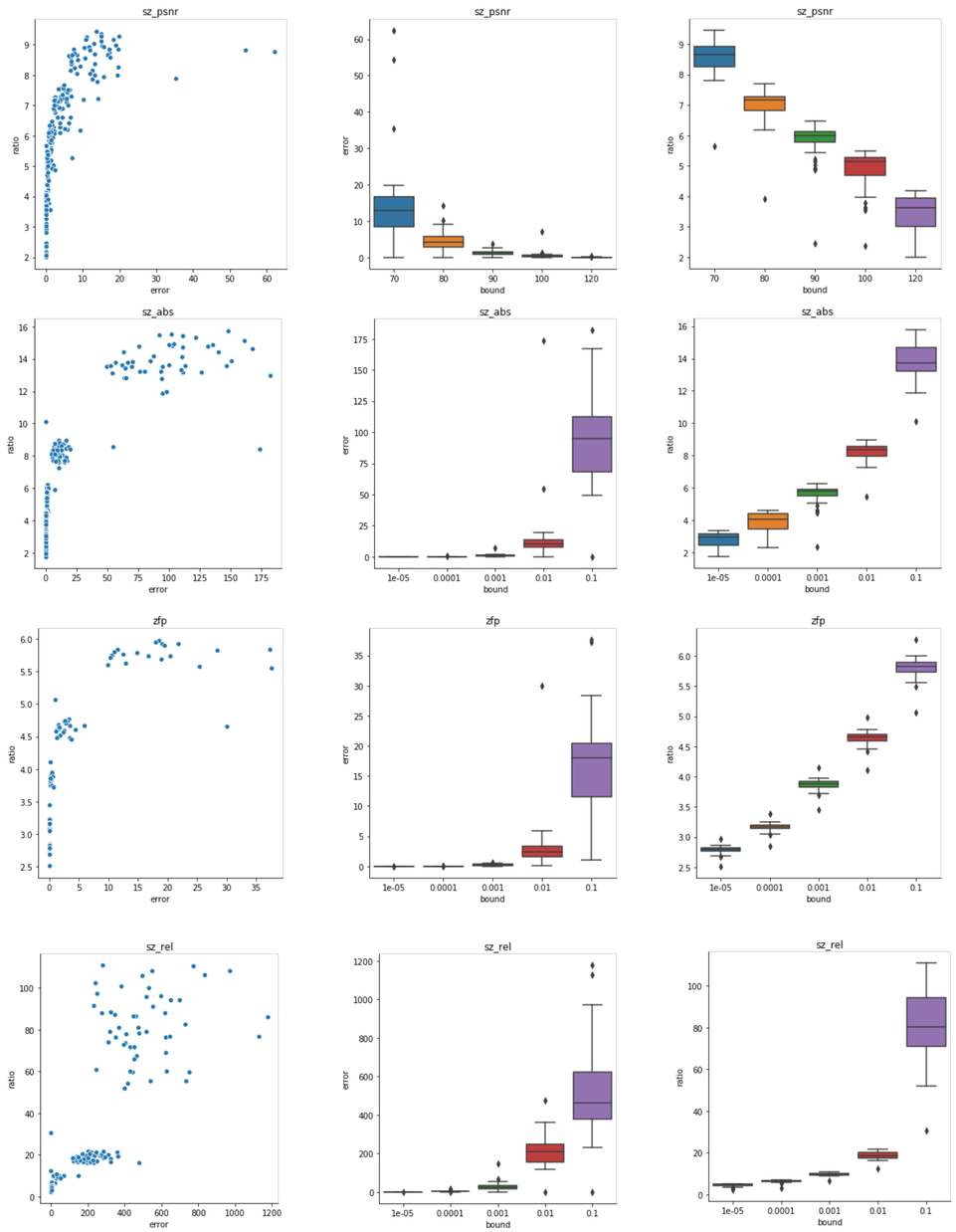
Methods

52 Gene Expression Matrices (GEMs) were used, which are commonly used as input for the KINC workflow. For each compressor config, 5 error bounds were chosen, yielding about 250 total GEMs to compress for each config. Compression Ratio and Average Percent Error were recorded, parametrized on error bound, and then plotted against each other.

NAN values were replaced with zero for compression and then reinserted after compression. They were ignored in error computation.

Experimental Results

How do different compressor configurations affect correlation accuracy and compression ratio?



Future Work

Run similarity workflows such as KINC on compressed data

Optimize KINC even for uncompressed data