Weekly update 04/12/2019-11/12/2019

Stephen Coleman

10/12/2019

Consensus clustering

I have discovered that the command line version of MDI does not handle sparse categorical data well. I am trying to understand the priors he uses (the most likely source of the problem according to Paul), but a lack of comments apart from such cases as "TODO: add an explicit "sample from prior" call in here" are making it an uphill battle.

Comparison command line to MATLAB

Let's compare the posterior similarity matrices (PSMs) produced by the different methods. Our methods are:

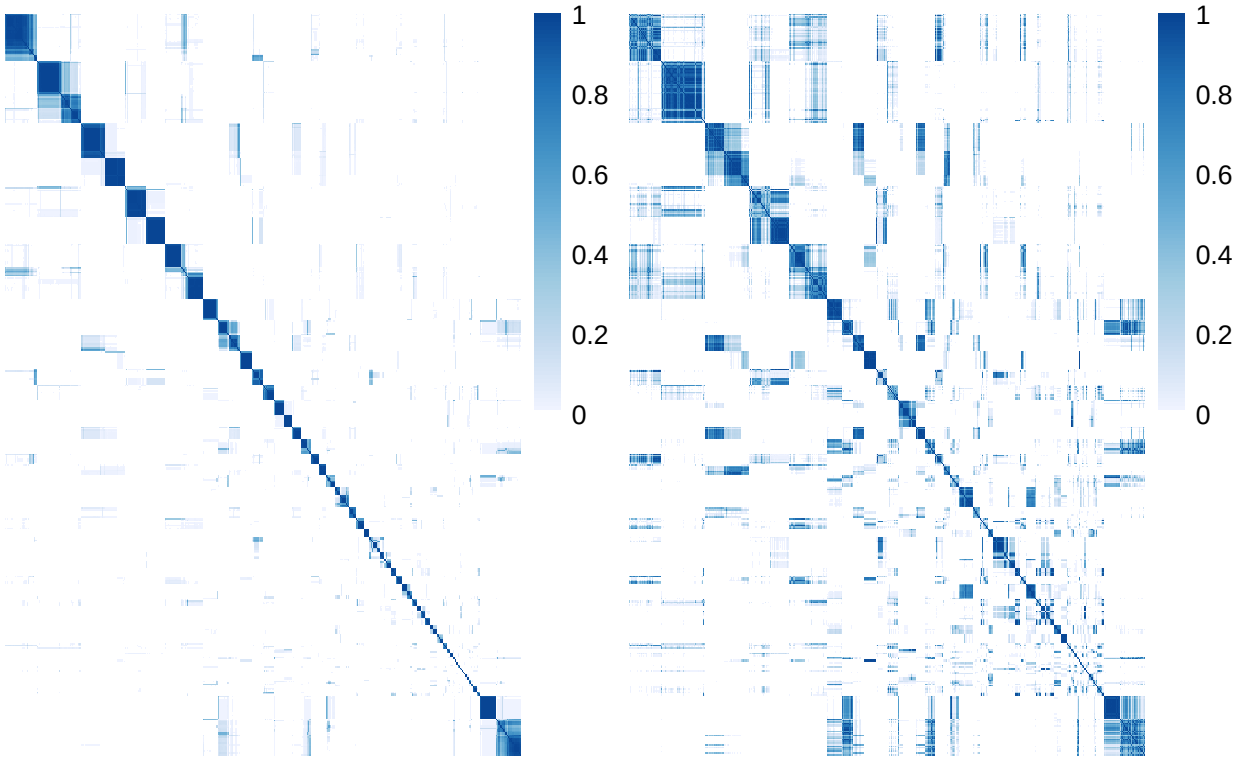
1. Command line MDI (358,500);
2. Command line consensus clustering (1,000 seeds for 500 iterations);
3. MATLAB MDI (8,080 iterations); and
4. MATLAB consensus clustering (1,000 seeds for 100 iterations).

The number of iterations is due to constraints on the

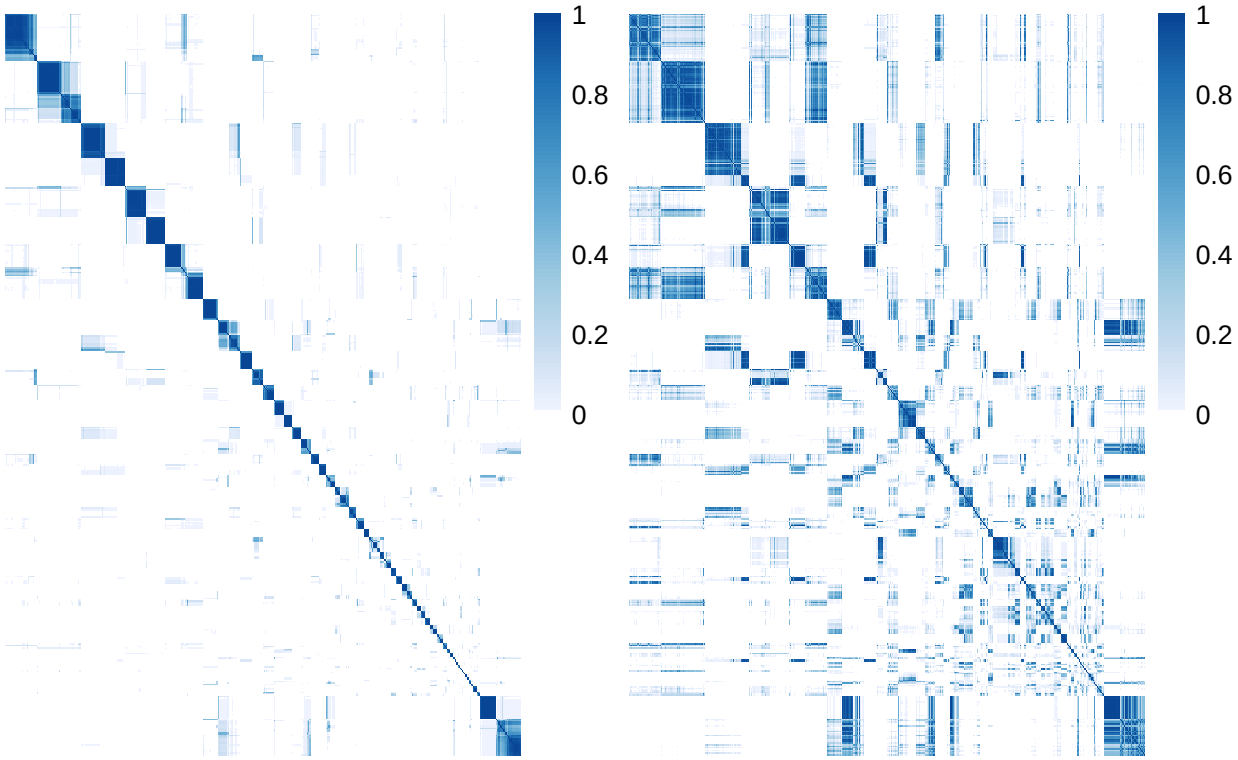
In each case the MATLAB MDI is our gold standard.

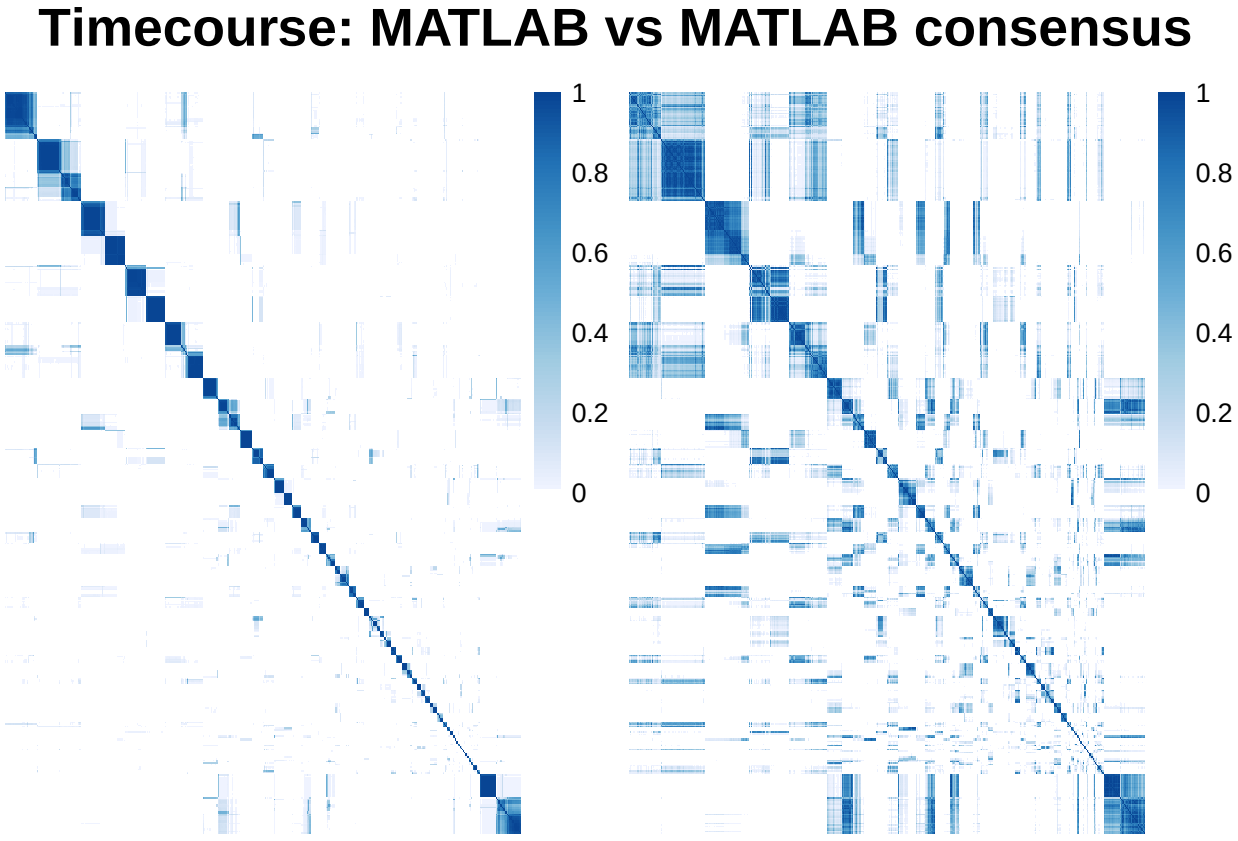
We have three datasets, the time course data (which seems most consistent across methods), which is run using the Gaussian Process setting, and then two sparse categorical datasets, the Chip-chip transcription factor data from the Harbison paper and the protein-protein interaction data (Harbison and PPI resepective). Let's look at the comparison of the timecourse data for the MATLAB MDI and Command line MDI:

Timecourse: MATLAB vs CMD Line



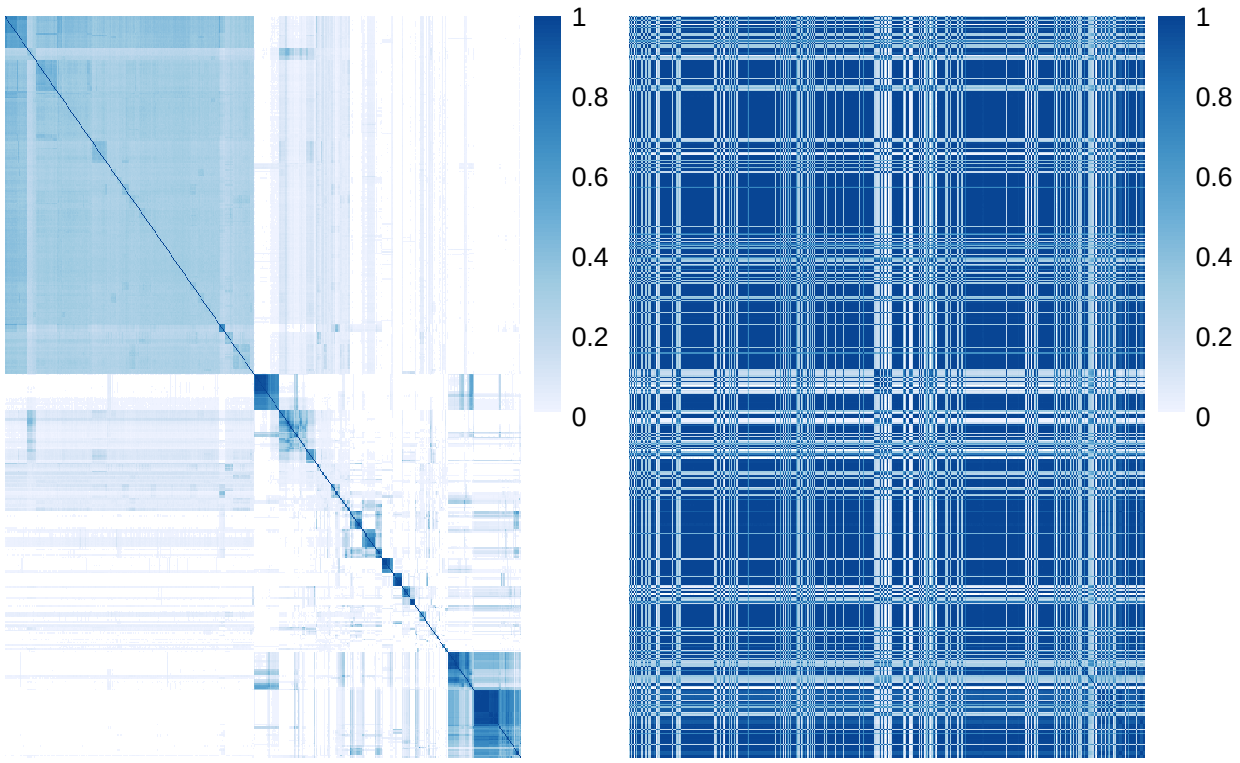
Comparing the MATLAB to both forms of consensus clustering:

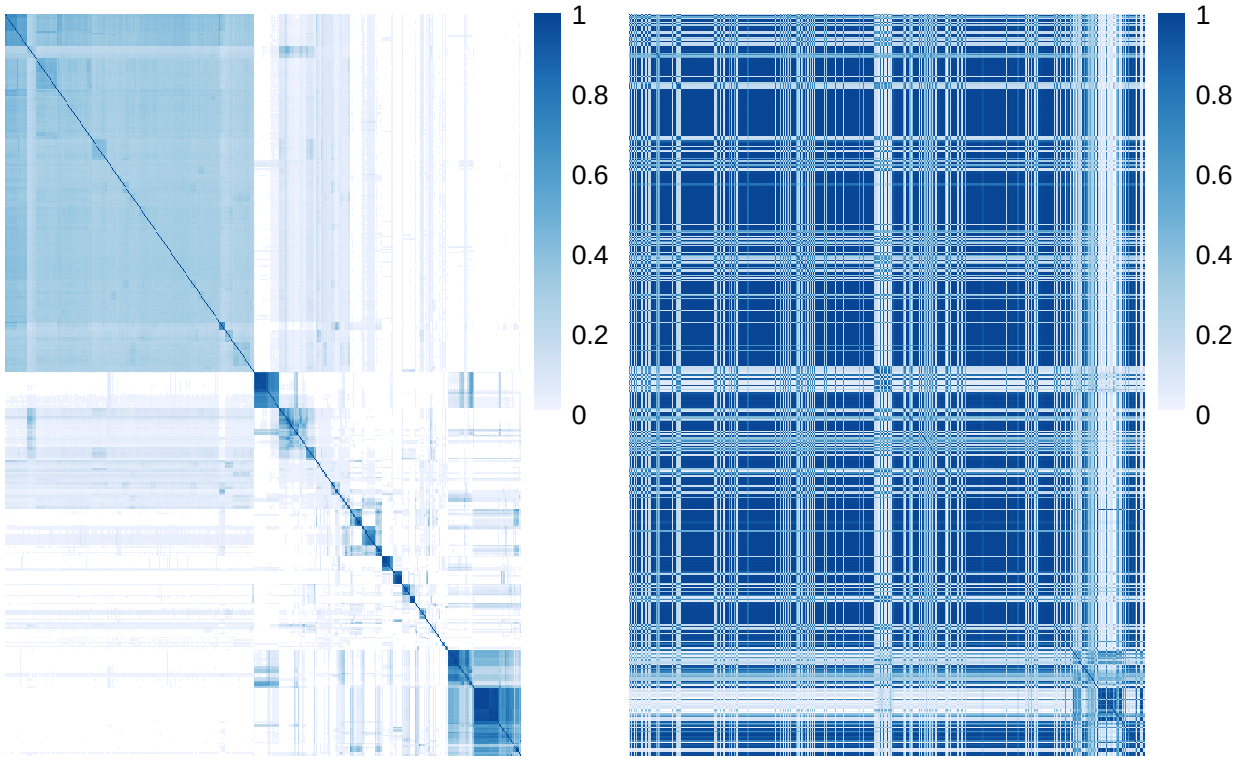




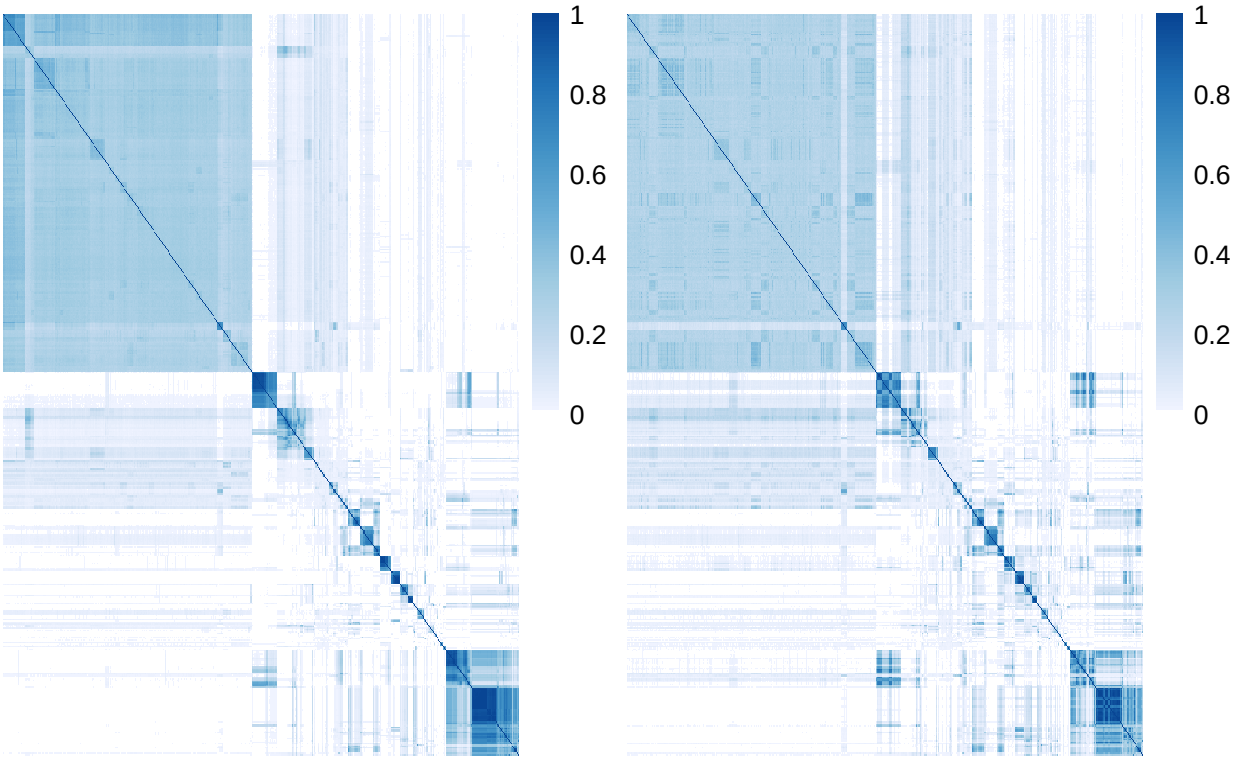
This is not too bad. There's slightly different stories but quite a lot of agreement too. The real difference is

within the categorical datasets:

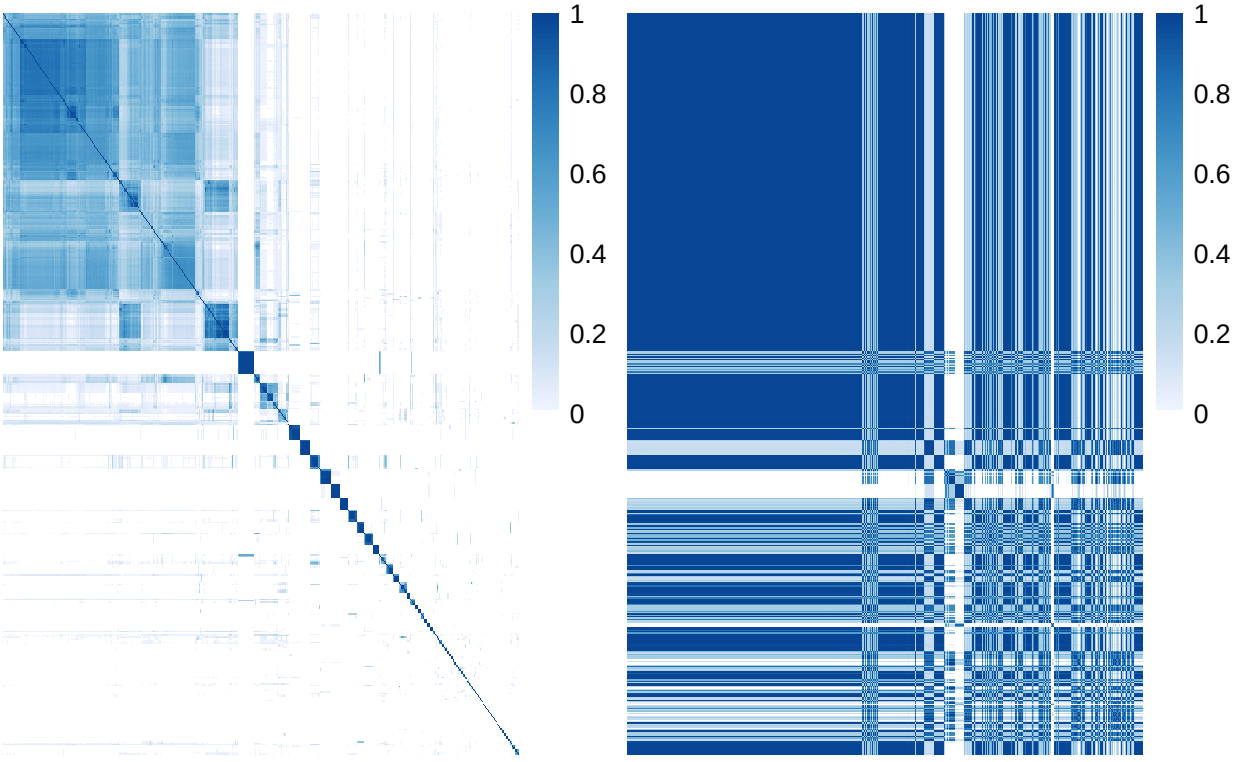


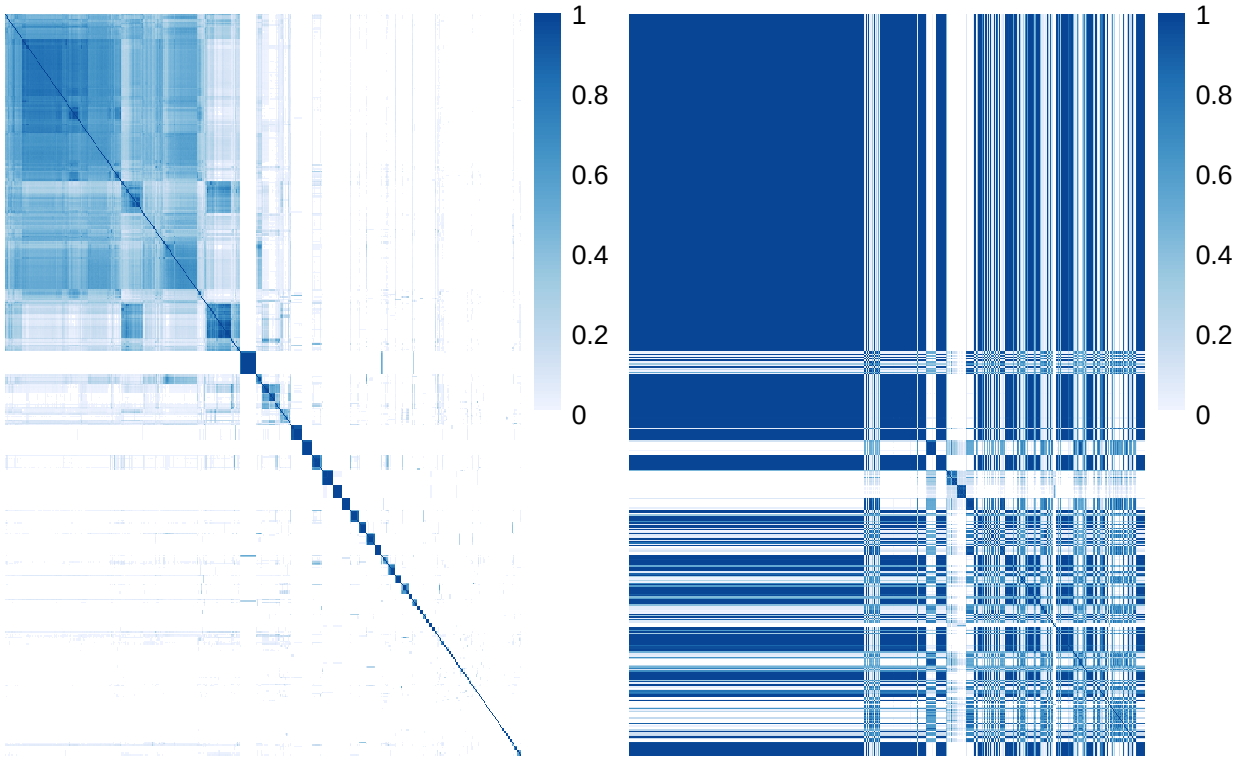


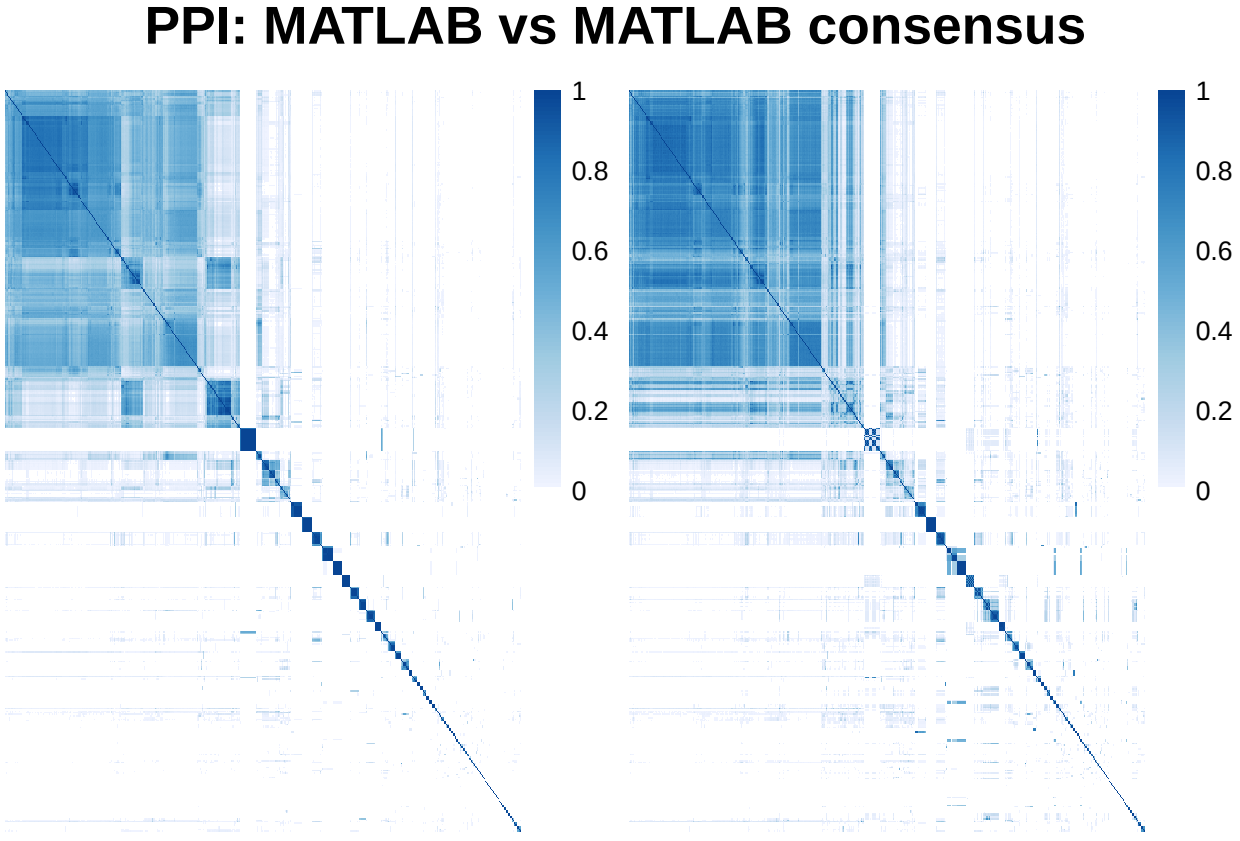
Harbison: MATLAB vs MATLAB consensus



PPI: MATLAB vs CMD Line







We can see that the consensus clustering works pretty well even with only 100 iterations. That's a nice result.