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:

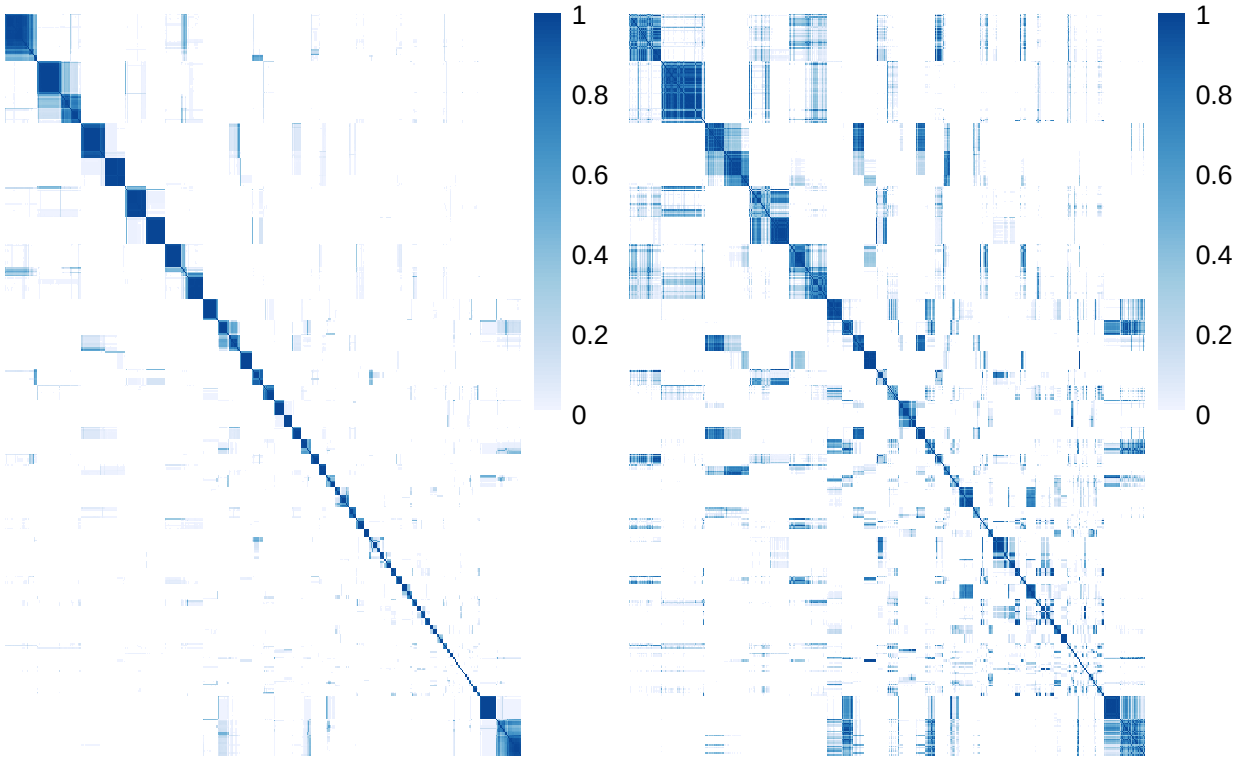
1. Command line MDI (358,500);
2. Command line consensus clustering (1,000 seeds for 500 iterations);
3. MATLAB MDI (8,080 iterations);
4. MATLAB consensus clustering (1,000 seeds for 100 iterations); and
5. MATLAB consensus clustering (1,000 seeds for 500 iterations).

The number of iterations is an odd number for the long chains due to constraints on the HPC.

In each case the MATLAB MDI is being considered our gold standard. This might not be fair as there's no strong reason to think this chain has converged, but it's similar to results to the original MDI paper, so for now it's a good heuristic for checking for sensible results.

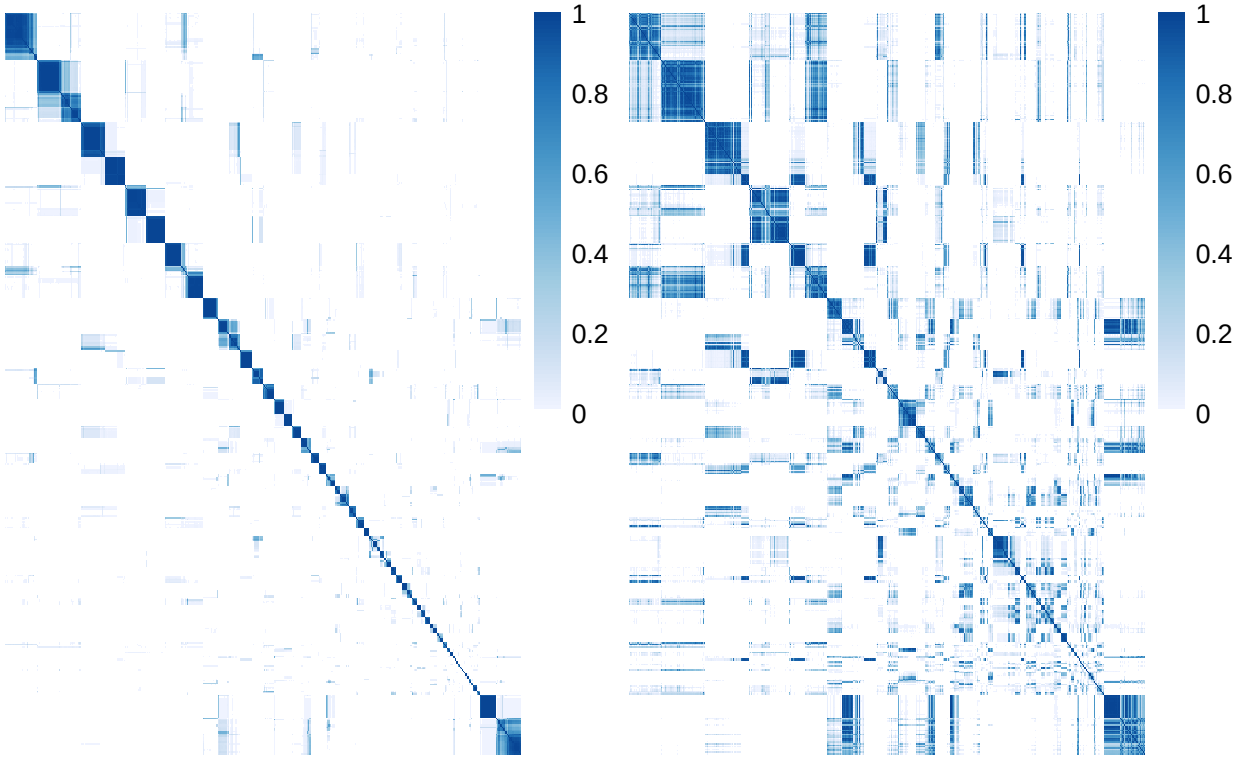
We have three datasets, the time course data (which seems most consistent across methods), which is run using the Gaussian Process setting, and then two sparse categorical datasets, the Chip-chip transcription factor data from the Harbison paper and the protein-protein interaction data (Harbison and 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

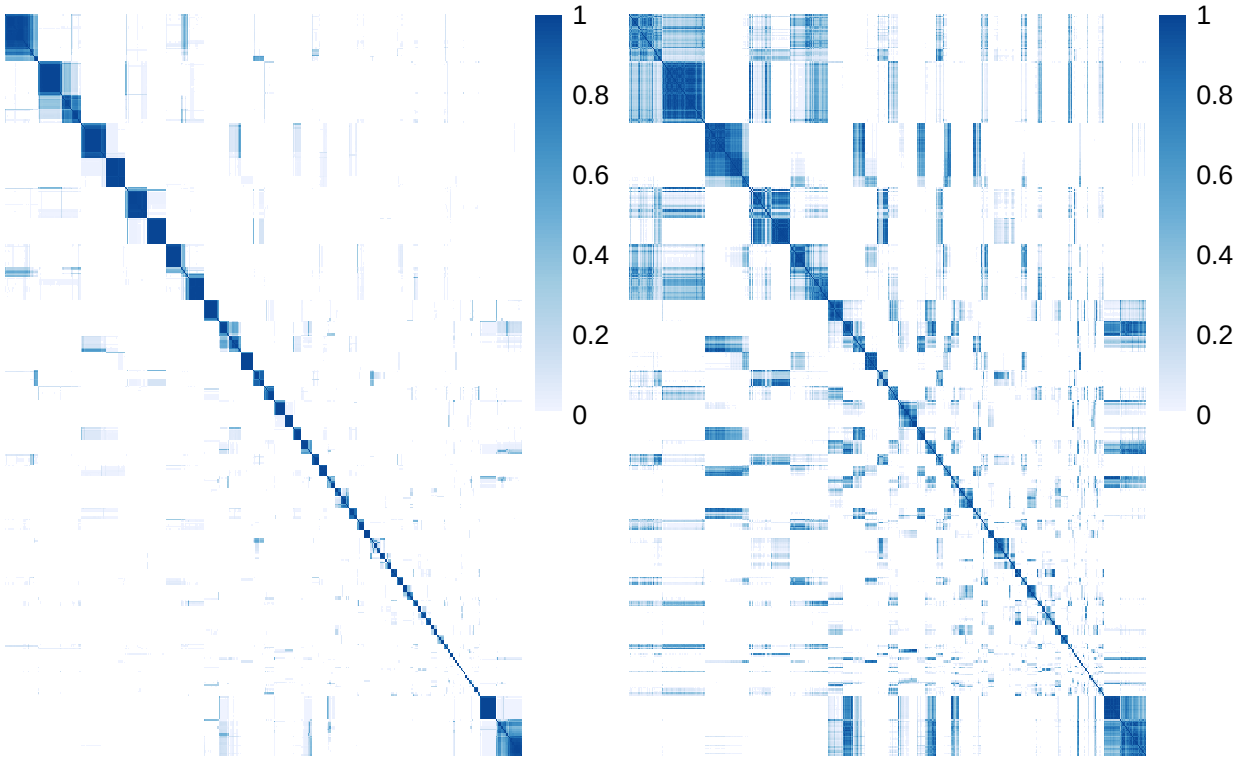


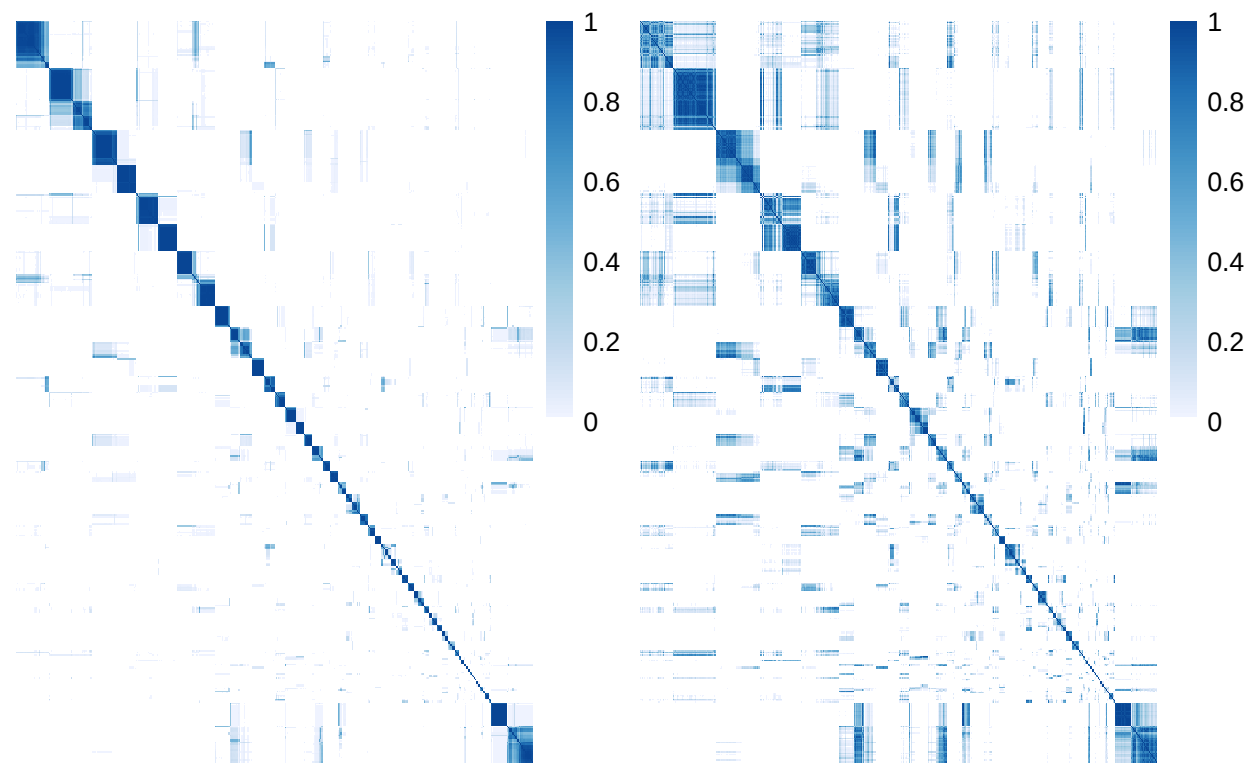
Comparing the MATLAB to both forms of consensus clustering:

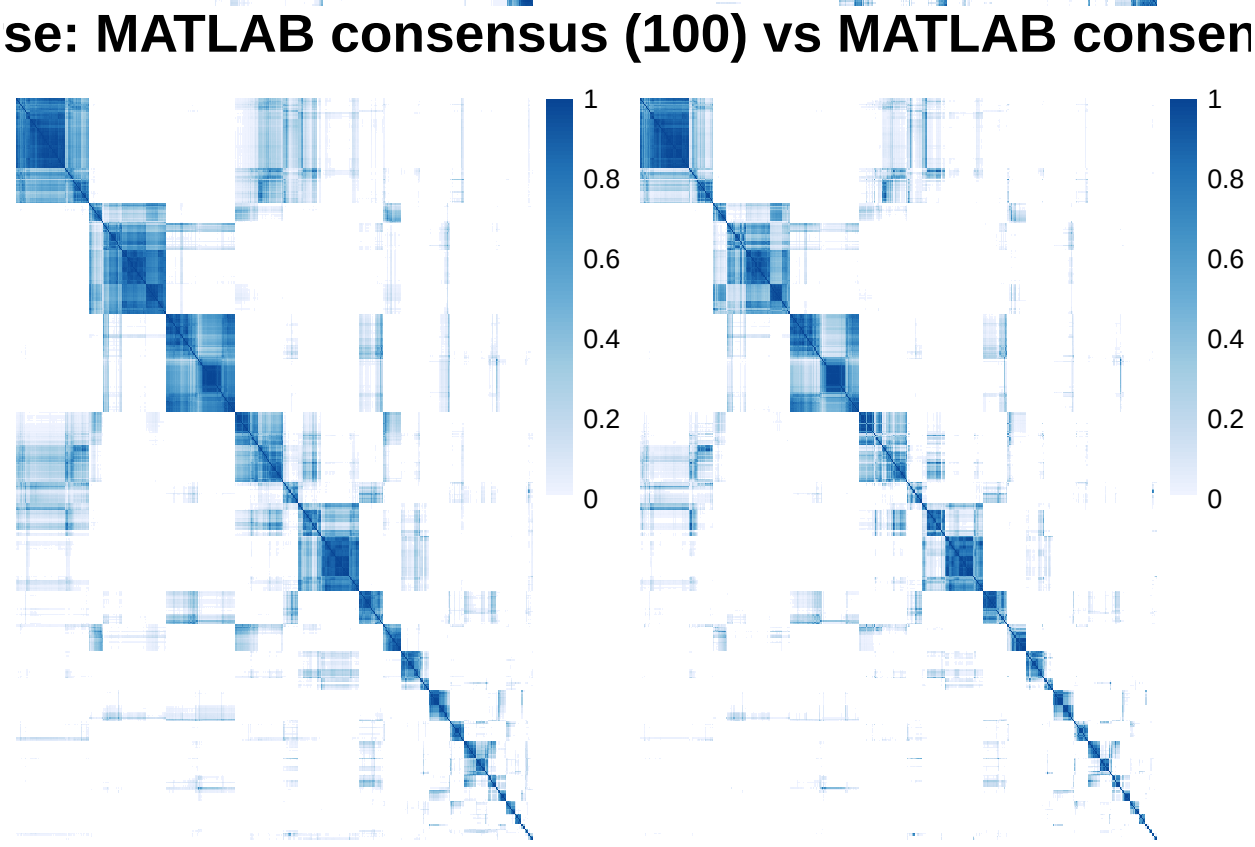
Timecourse: MATLAB vs CMD Line consensus



imecourse: MATLAB vs MATLAB consensus (100

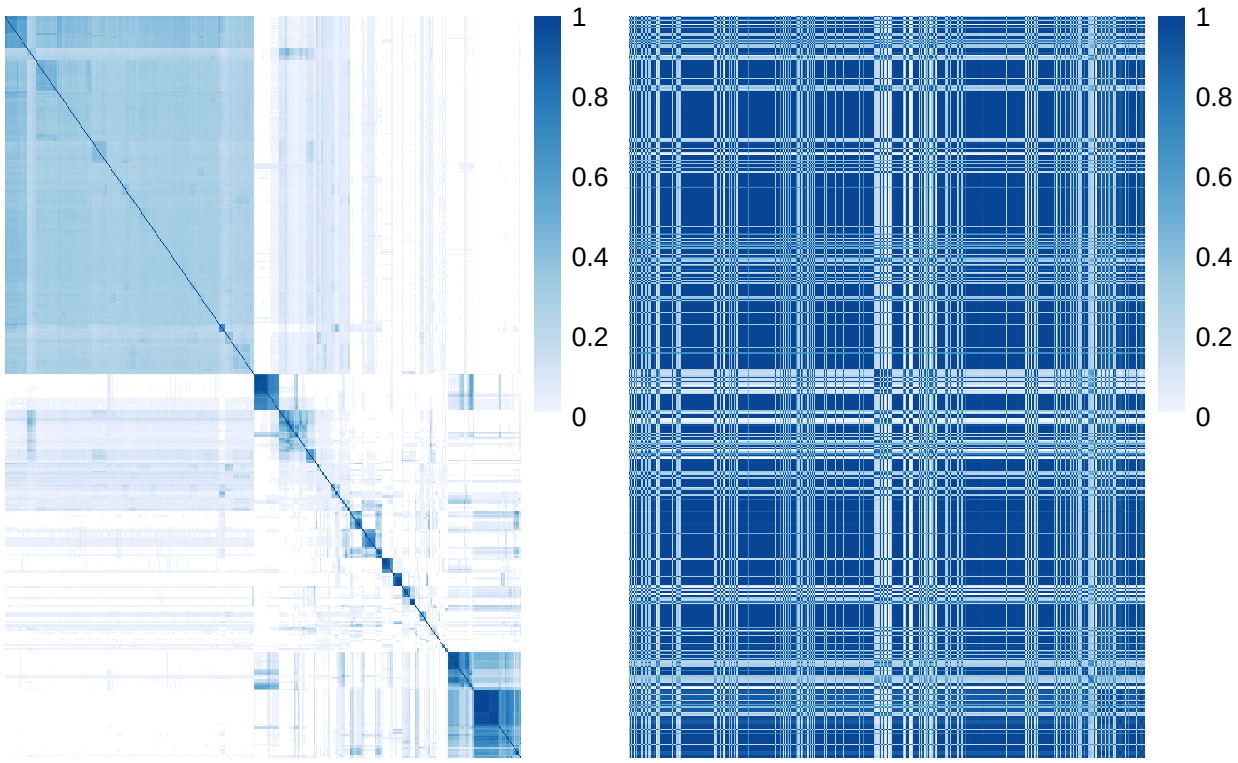


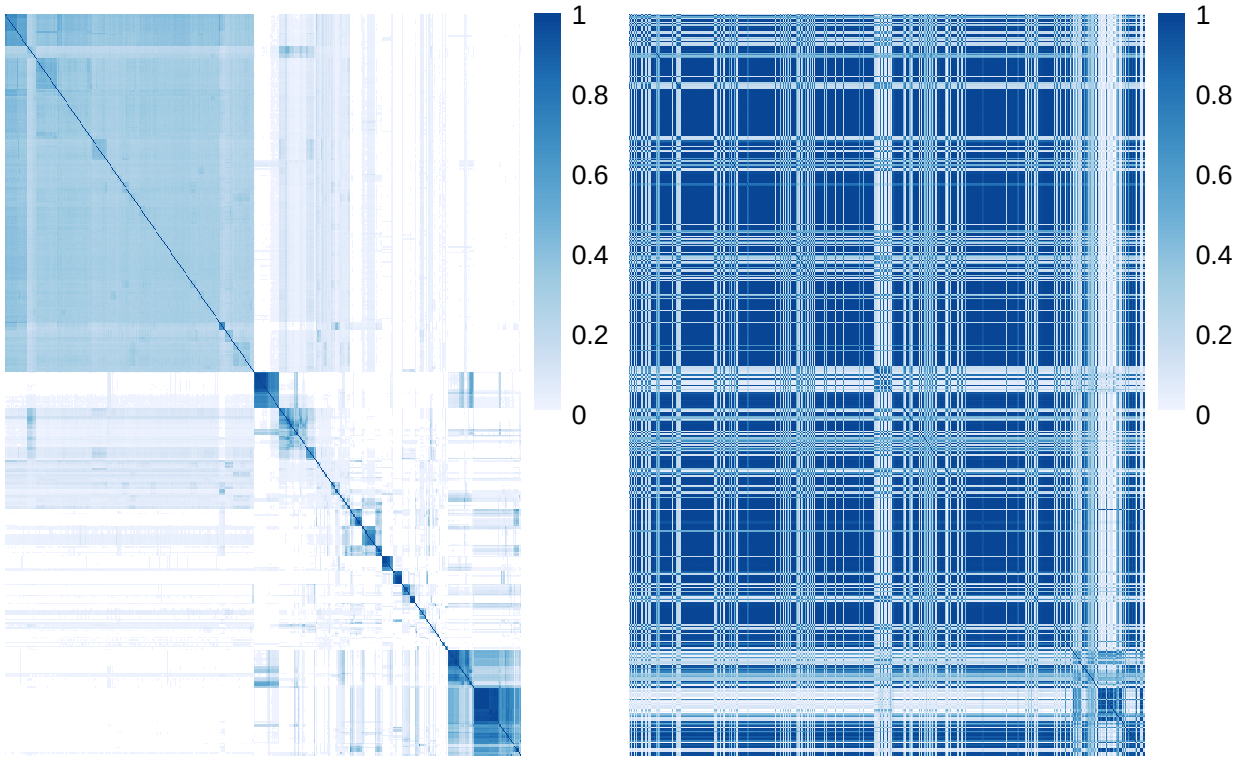




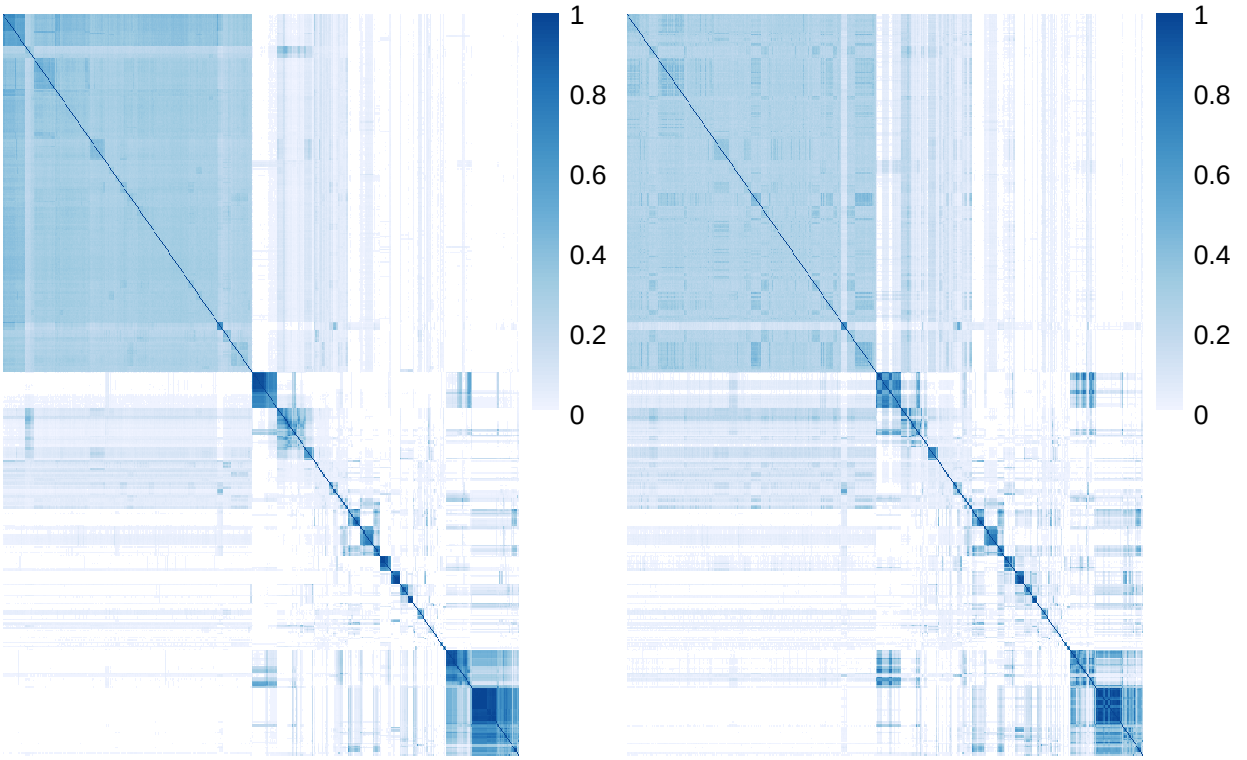
This is not too bad. There's slightly different stories but quite a lot of agreement too. The real difference is

within the categorical datasets:

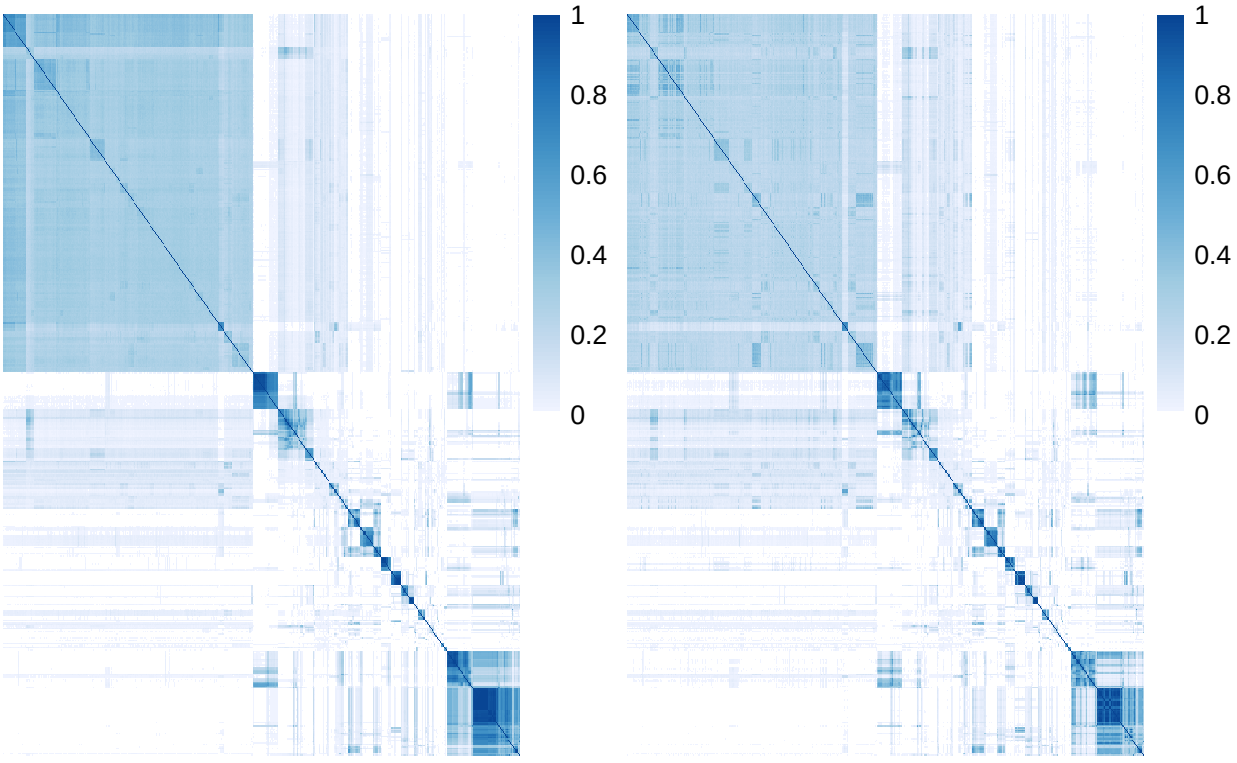




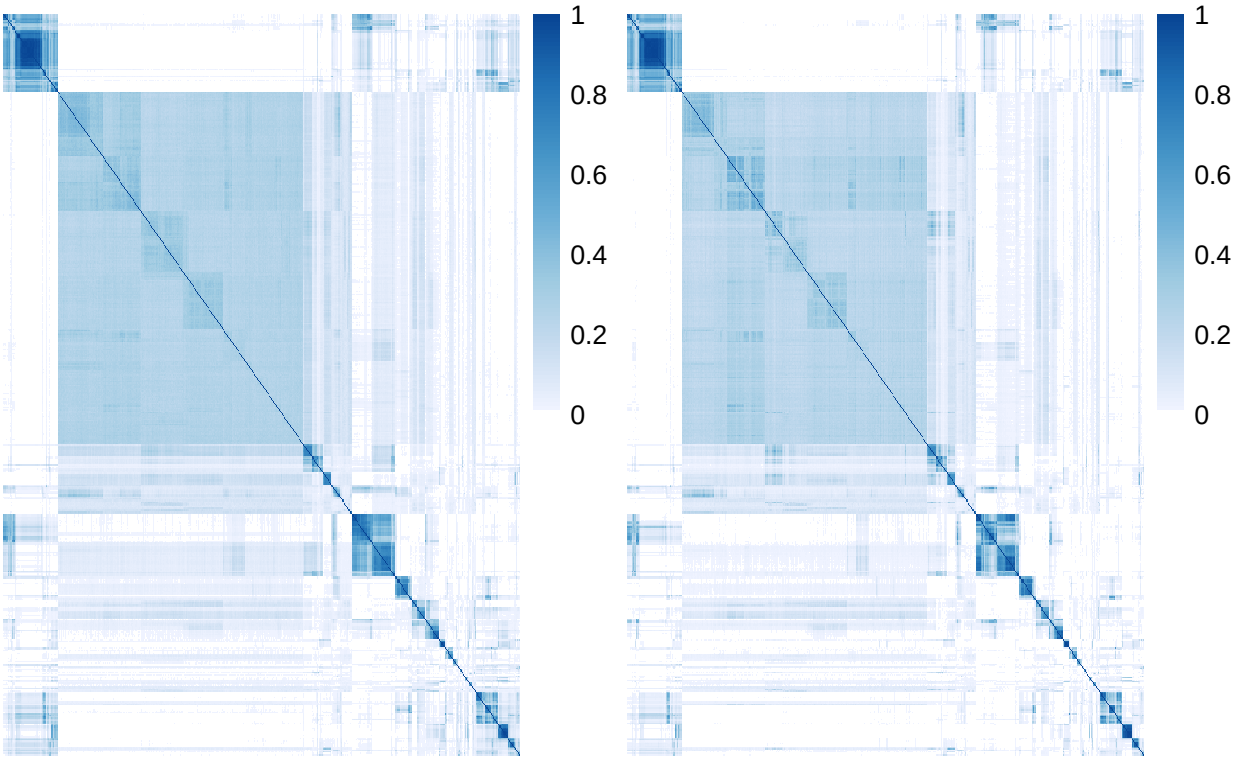
Harbison: MATLAB vs MATLAB consensus (100)



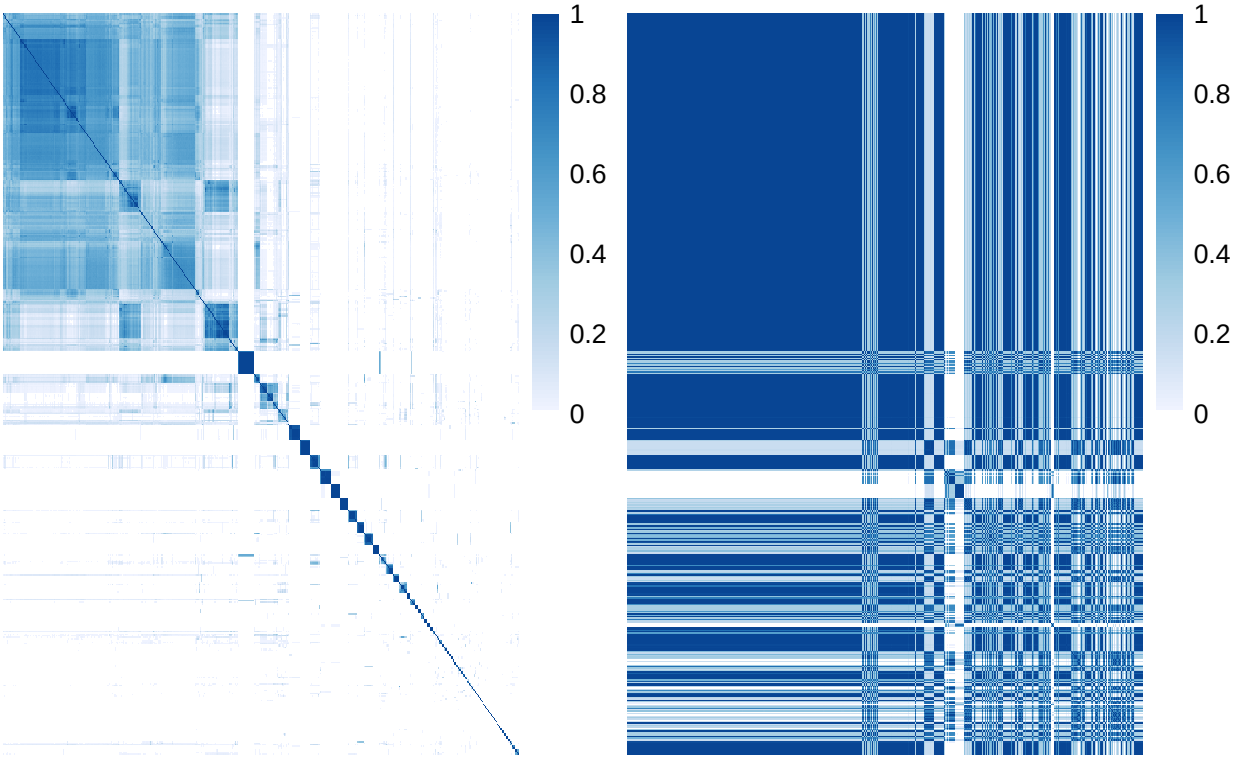
arbison: MATLAB MDI vs MATLAB consensus (50



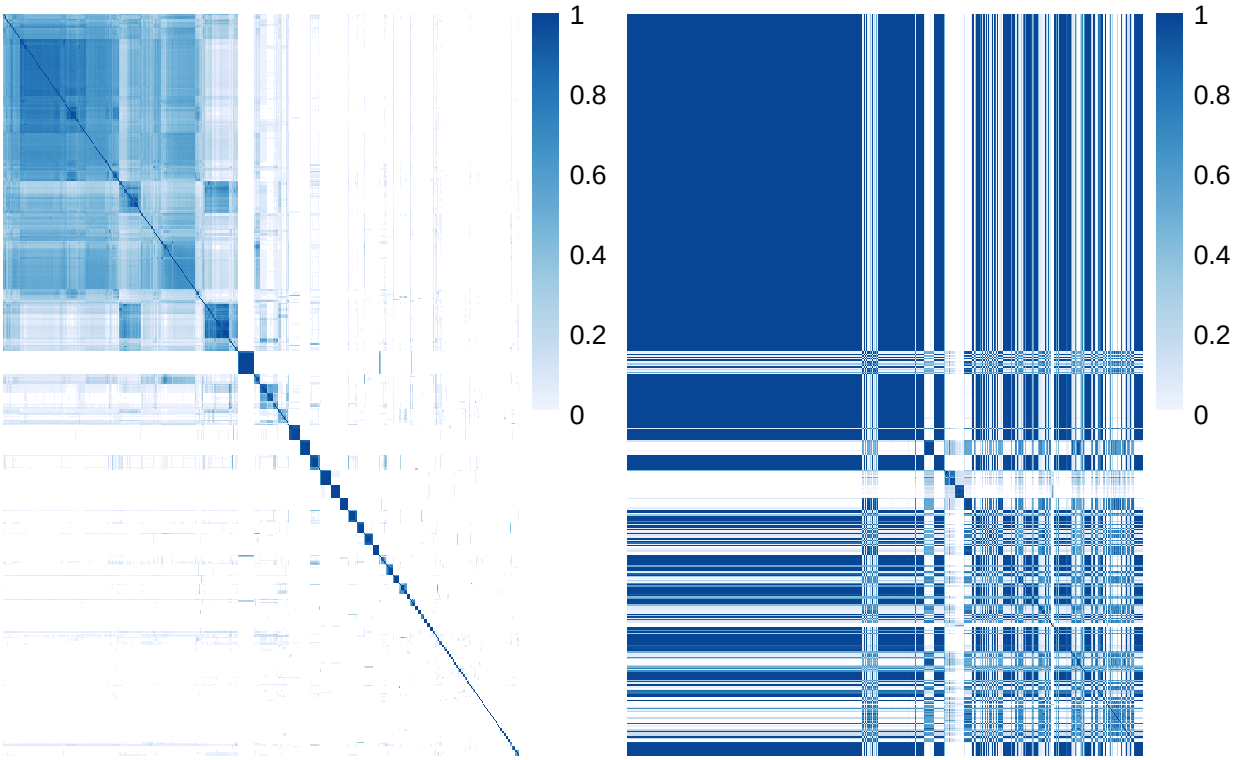
n: MATLAB consensus (100) vs MATLAB consens



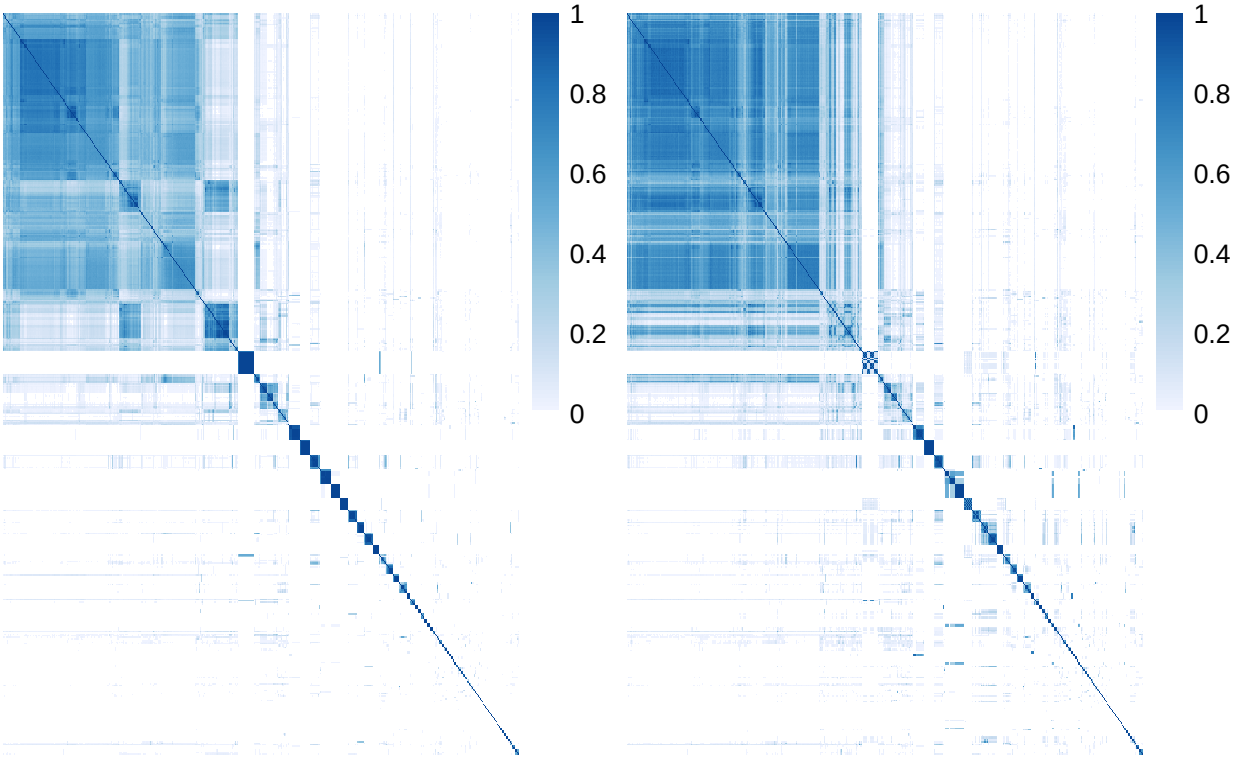
PPI: MATLAB vs CMD Line

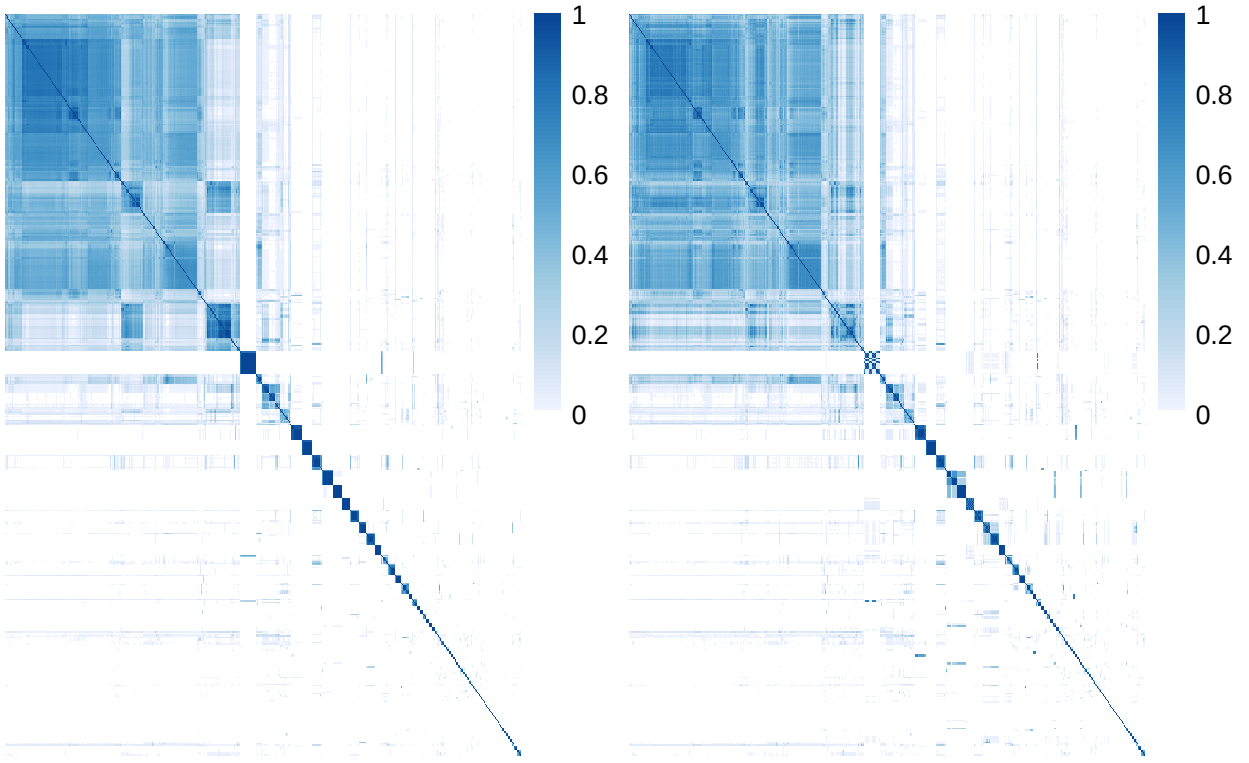


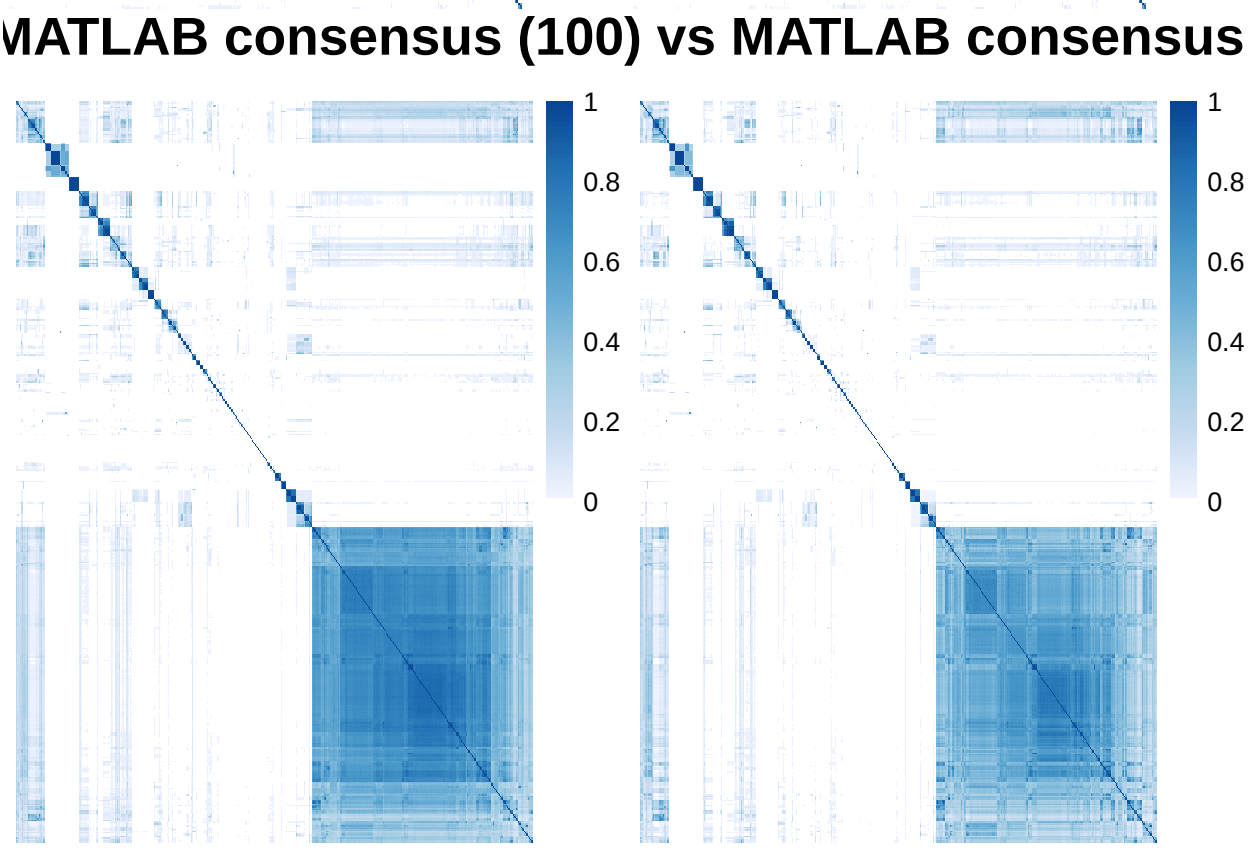
PPI: MATLAB vs CMD Line consensus



PPI: MATLAB vs MATLAB consensus (100)







We can see that the consensus clustering works pretty well even with only 100 iterations, but really 500 is a

fair bit better.

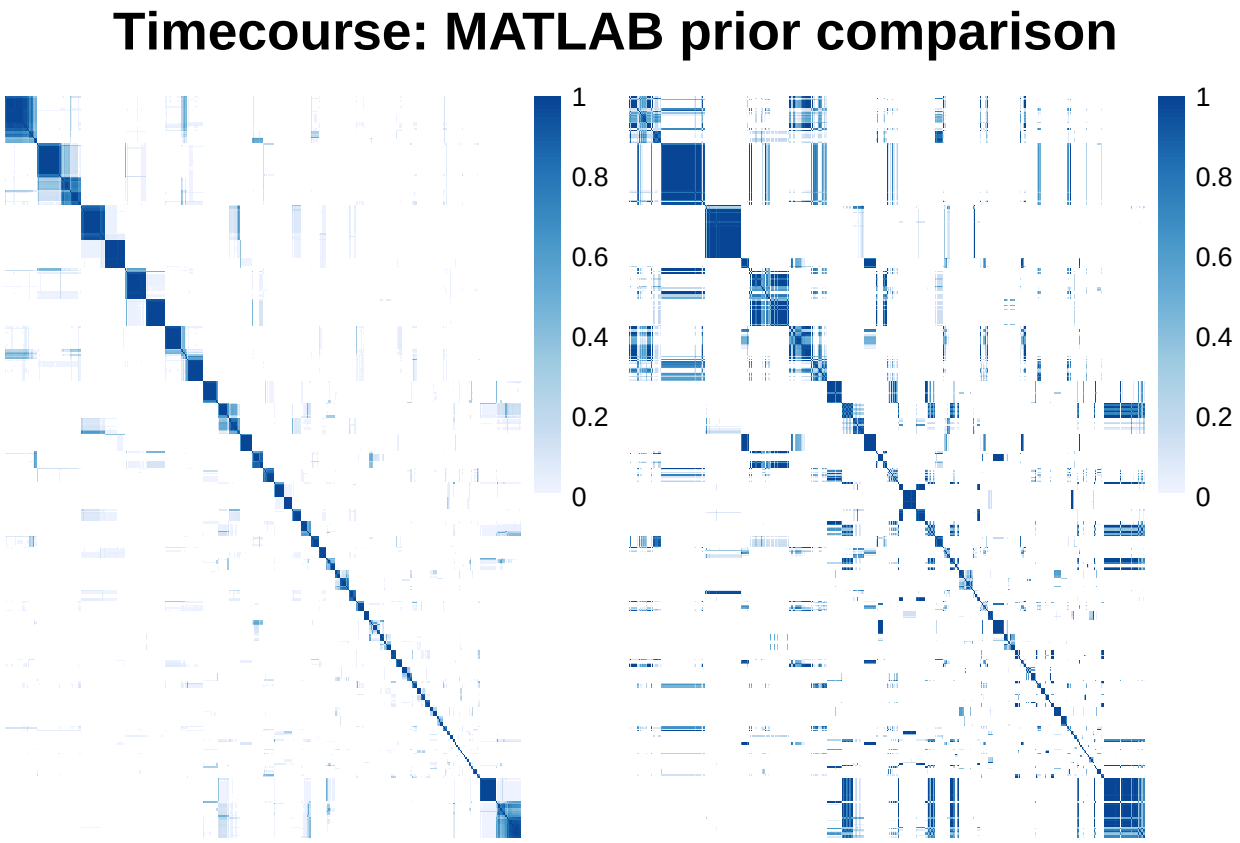
The priors were finally located in the Command line code. The categorical prior is set to 0.5. In the MATLAB code it's an empiricla prior based upon the proportion of each category across the entire datasets. In our example categorical datasets the priors are then:

$$\text{Command line prior: } \begin{pmatrix} 0.50 \\ 0.50 \end{pmatrix},$$

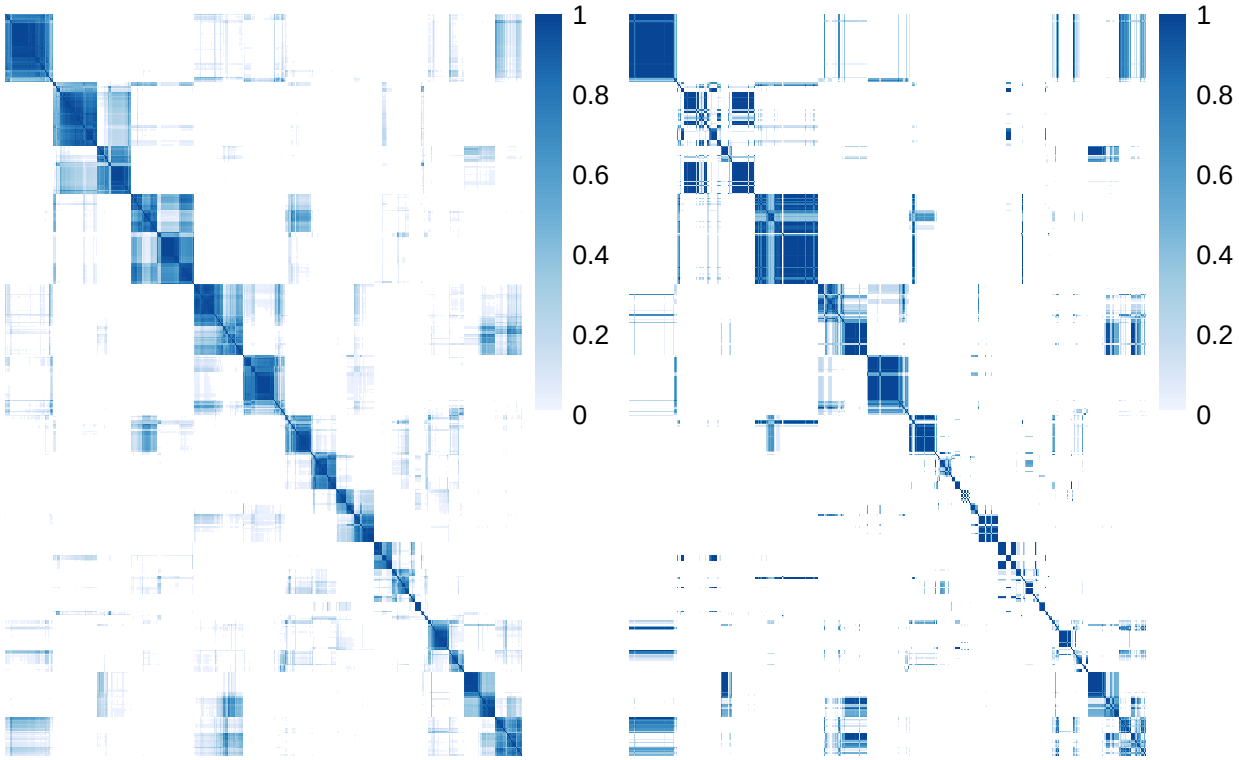
$$\text{MATLAB prior for the Transcription factor data: } \begin{pmatrix} 0.988 \\ 0.012 \end{pmatrix},$$

$$\text{MATLAB prior for the PPI data: } \begin{pmatrix} 0.983 \\ 0.017 \end{pmatrix}$$

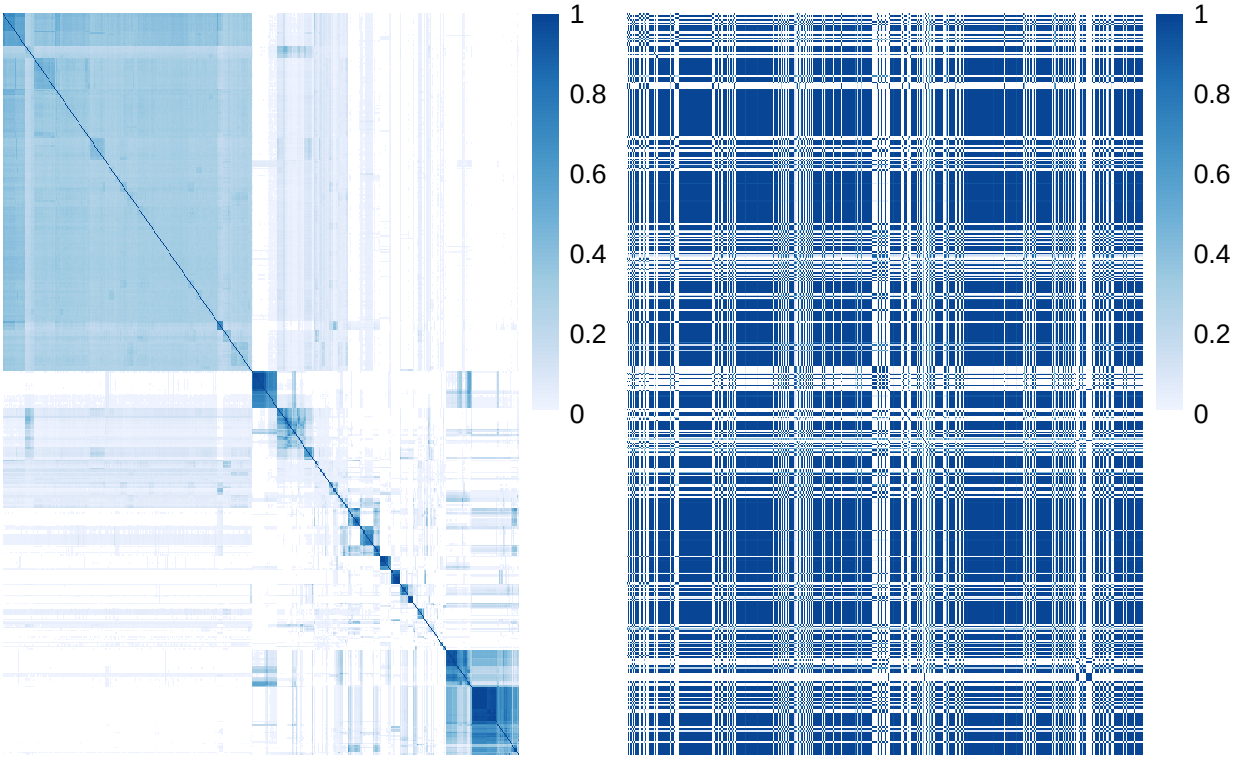
A clear difference. To check if this is the source of the issue the MATLAB code was re-run with a prior to match the command line code. The results can be seen below for each dataset:



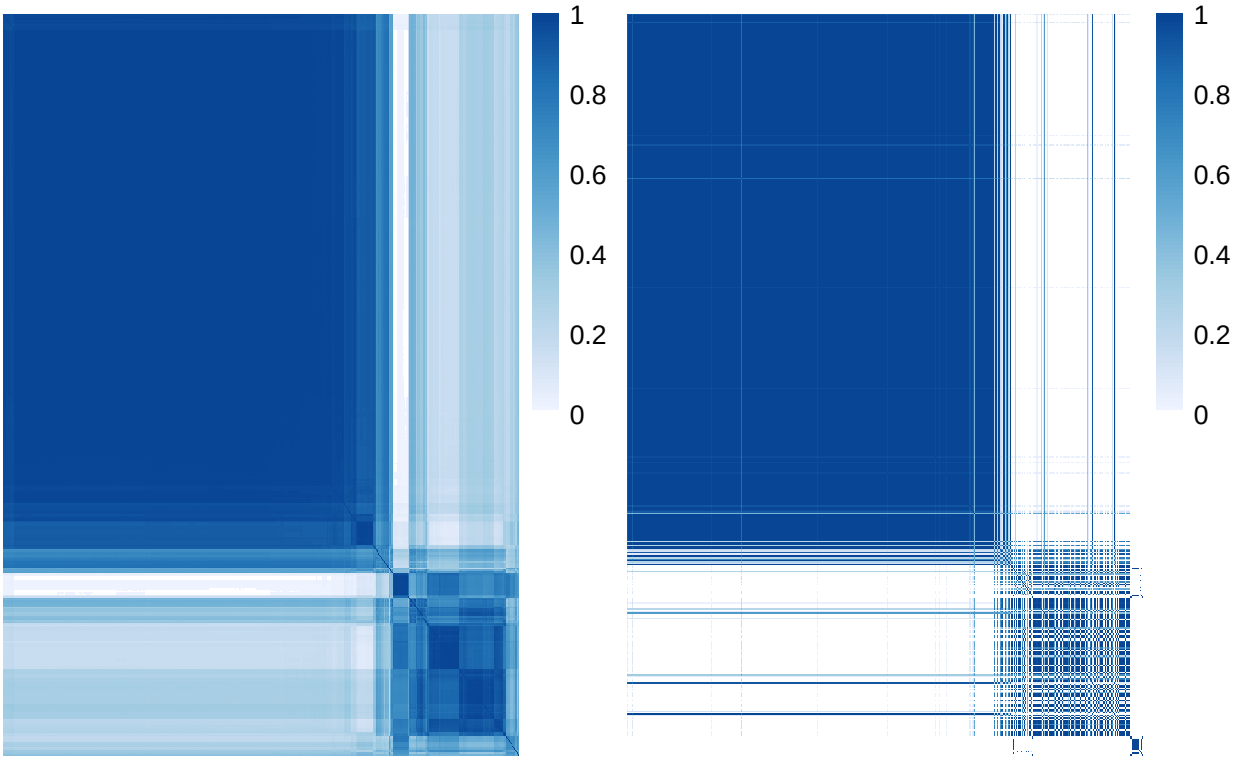
Timecourse: Cmd line vs MATLAB, common priors



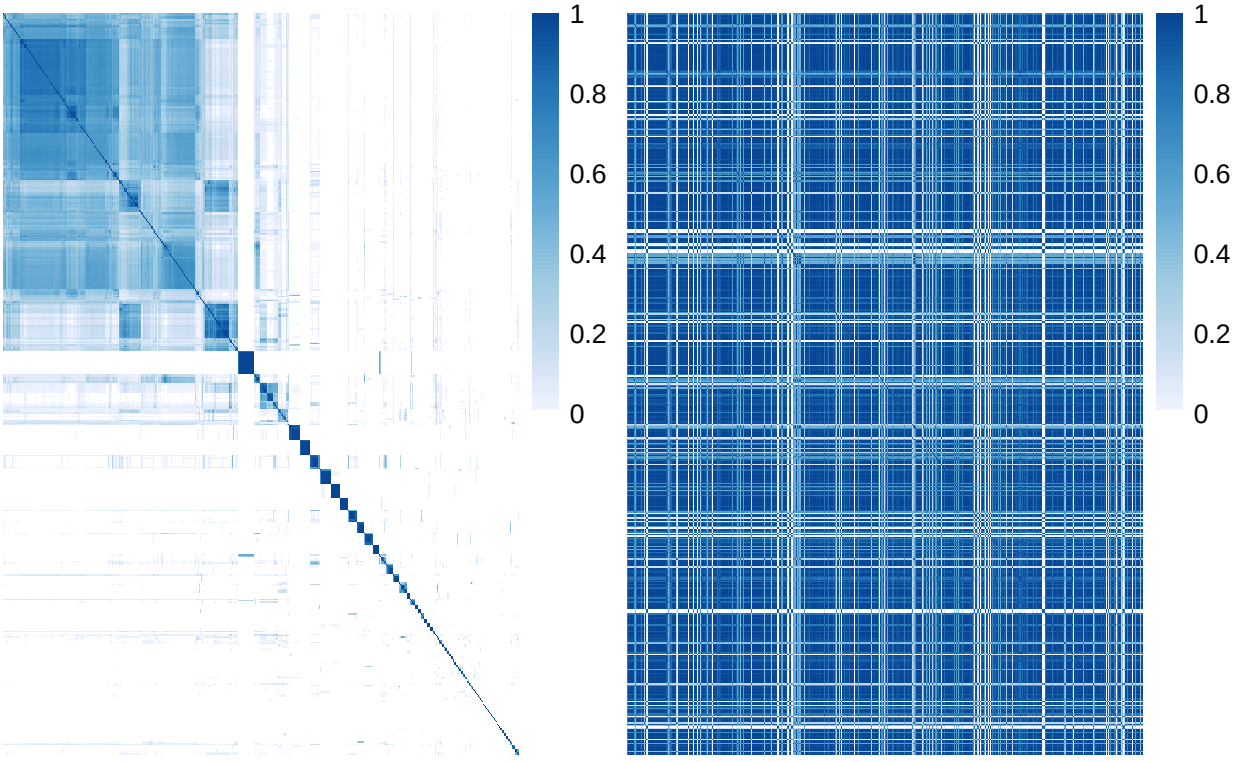
Harbison: MATLAB prior comparison



Harbison: Cmd line vs MATLAB, common priors



PPI: MATLAB prior comparison



PPI: Cmd line vs MATLAB, common priors