Consensus clustering for Bayesian mixture models: Supplementary materials

Stephen Coleman, Paul DW Kirk and Chris Wallace

October 5, 2020

Abstract

1 Yeast data

The "Yeast data" consists of three *S. cerevisiae* datasets with gene products associated with a common set of 551 genes present. The datasets are:

- microarray profiles of RNA expression from Granovskaia et al. (2010) a cell cycle dataset that comprises measurements taken at 41 time points (note: this is referred to as the **Timecourse** dataset in this paper),
- Chromatin immunoprecipitation followed by microarray hybridization (**ChIP-chip**) data from Harbison et al. (2004), and
- Protein-protein interaction (**PPI**) data from BioGrid (Stark et al., 2006).

The datasets were reduced to 551 items by considering only the genes identified by Granovskaia et al. (2010) as having periodic expression profiles with no missing data in the PPI and ChIP-chip data, following the same steps as the original MDI paper (Kirk et al., 2012). The datasets were modelled using a base measure of a Gaussian process in the Timecourse dataset and Multinomial distributions in the ChIP-chip and PPI datasets.

1.1 Consensus clustering analysis

An ensemble of depth $R = 501$ and size $S = 100$ was initially attempted. The consensus matrices for this ensemble was compared to those for the combinations of $R = (101, 501)$, $S = (50, 100)$ in the three datasets. This was deemed not sufficiently stable and thus a deeper and broader ensemble was considered.

1.2 Bayesian analysis

10 chains were run for 36 hours, resulting in 676,000 iterations per chain, thinned to every thousandth sample, resulting in 676 samples per chain. These chains were investigated for

- within-chain stationarity using the Geweke Z-statistics, and
- across-chain convergence using the Gelman-Rubin shrinkage coefficient ($\hat{R}$) and the Vats-Knudson extension (*stable* $\hat{R}$ Vats and Knudson, 2018)

References

Marina V Granovskaia, Lars J Jensen, Matthew E Ritchie, Joern Toedling, Ye Ning, Peer Bork, Wolfgang Huber, and Lars M Steinmetz. High-resolution transcription atlas of the mitotic cell cycle in budding yeast. *Genome biology*, 11(3):1–11, 2010.

Christopher T Harbison, D Benjamin Gordon, Tong Ihn Lee, Nicola J Rinaldi, Kenzie D Macisaac, Timothy W Danford, Nancy M Hannett, Jean-Bosco Tagne, David B Reynolds, Jane Yoo, et al. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104, 2004.

Paul Kirk, Jim E Griffin, Richard S Savage, Zoubin Ghahramani, and David L Wild. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, 28(24):3290–3297, 2012.

Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. Biogrid: a general repository for interaction datasets. *Nucleic acids research*, 34(suppl_1):D535–D539, 2006.

Dootika Vats and Christina Knudson. Revisiting the gelman-rubin diagnostic. *arXiv preprint arXiv:1812.09384*, 2018.