Consensus clustering for Bayesian mixture models: Supplementary materials

Stephen Coleman, Paul DW Kirk and Chris Wallace

October 12, 2020

Abstract

1 The models

In the simulations (see section 2) where individual datasets are modelled a *Bayesian mixture model* is used. The densities that model each cluster are *Gaussian*. The directed acyclic graph for this model is shown in figure 1. In the multiple dataset case (the *Yeast data*, see the analysis in section 3), *Multiple Dataset Integration* (Kirk et al., 2012) is used. The DAG for this model for three datasets is shown in figure 2.

2 Simulations

A number of scenarios are defined by the parameters that they use in generating individual simulations using algorithm 1. These parameters are shown in table 1.

Each scenario is seen as testing certain concepts or else specific characteristics of real data.

- *2D*: a low dimensional scenario within which `Mclust` is epxected to perform well and the long chains are expected to converge and explore the full support of the posterior distribution.
- *No structure*: included to reassure fears that Consensus clustering has a predilection to finding clusters where none exist (Şenbabaoğlu et al., 2014a,b).
- *Base case*: a highly informative setting used to benchmark a number of other scenarios that are variations of this setting.
- *Large standard deviation*: these two scenarios investigate the degree of distinction required between clusters for the methods to uncover their structure.

Input: Distance between means Δ_μ
A common standard deviation σ^2
A number of clusters K
The number of items to generate in total N
The number of features to generate in total P
An indicator vector of feature relevance $\phi = (\phi_1, \ldots, \phi_P)$
The expected proportion of items in each cluster $\pi = (\pi_1, \ldots, \pi_K)$
A method for sampling x times from the array y, with weights π: $Sample(y, x, \pi)$
A method for permuting a vector x: $Permute(x)$
A method for generating a value from a univariate Gaussian distribution with mean μ and standard deviation σ^2: $Gaussian(\mu, \sigma^2)$
Output: A dataset, X
The generating cluster labels $c = (c_1, \ldots, c_N)$
begin
  `/* initialise the empty data matrix */`
  $X \leftarrow 0_{N \times P}$;
  `/* create a matrix of` K `means */`
  $\mu \leftarrow (\Delta_\mu, \ldots, K\Delta_\mu)$;
  `/* generate the allocation vector */`
  $c \leftarrow Sample(1 : K, N, \pi)$;
  $\mathbf{M} \leftarrow \mathbf{0}_{N \times N}$;
  for $p = 1$ **to** P **do**
    `/* Test if the feature is relevant, if relevant generate data from a mixture of univariate Gaussians, otherwise draw all items from the same distribution */`
    if $\phi_p = 1$ **then**
      $\nu \leftarrow Permute(\mu)$;
      for $n = 1$ **to** N **do**
        $X(n, p) \leftarrow Gaussian(\nu_{c_n}, \sigma^2)$
      end
    end
    if $\phi_p = 0$ **then**
      for $n = 1$ **to** N **do**
        $X(n, p) \leftarrow Gaussian(0, \sigma^2)$
      end
    end
  end
  `/* Mean centre and scale the data */`
  $X \leftarrow Normalise(X)$
end

Algorithm 1: Data generation for a mixture of Gaussian with independent features. This algorithm is implemented in the `generateSimulationDataset` function from the `mdiHelpR` package available at `www.github.com/stcolema/mdiHelpR`.

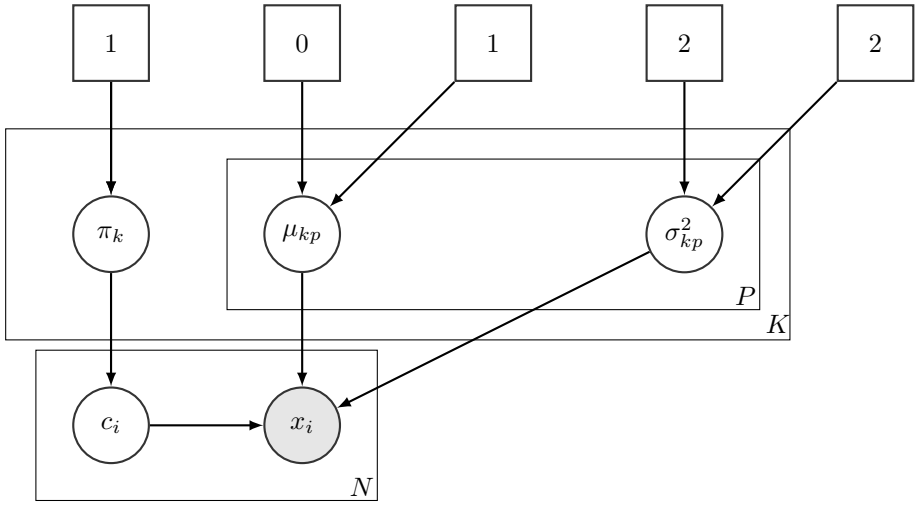

Figure 1: Directed acyclic graph for the mixture of Gaussians used.

- *Irrelevant features*: these scenarios investigate how robust the methods are to irrelevant features.
- *Varying proportions*: these scenarios investigate how well each method uncovers clusters when the clusters have significantly different membership counts.
- *Small N, large P*: an investigation of behaviour when the number of features is far greater than the number of items.

2.1 Bayesian analysis

For each simulation 10 chains of 1 million iterations were run, thinning to every thousandth sample. A burn-in of 10,000 iterations was then applied to the samples, leaving 990 samples per chain. These chains were investigated for

- within-chain stationarity using the Geweke convergence diagnostic (Geweke et al., 1991), and
- across-chain convergence using the potential scale reduction factor ($\hat{R}$, Gelman et al., 1992) and the Vats-Knudson extension (*stable* $\hat{R}$, Vats and Knudson, 2018).

The Geweke convergence diagnostic is a standard Z-score; it compares the sample mean of two sets of samples (in this case buckets of samples from the first half of the samples to the sample mean of the entire second half of samples). It is calculated under the assumption that the two parts of the chain are asymptotically independent and if this assumption holds than the scores are expected

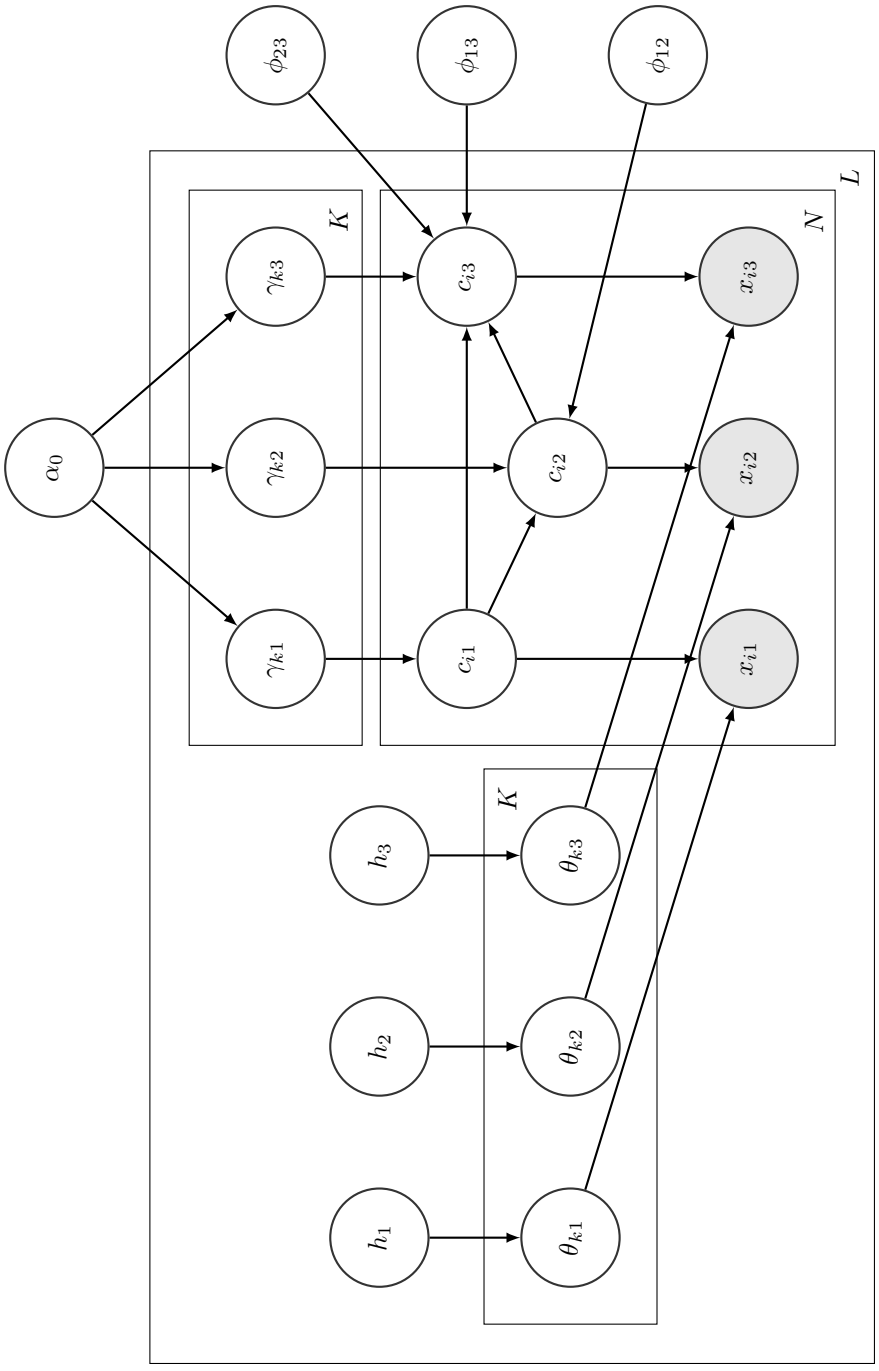

Figure 2: Directed acyclic graph for the Multiple Dataset Integration model for $L = 3$ datasets.

Scenario	N	P_s	P_n	K	Δ_μ	σ^2	π
2D	100	2	0	5	3.0	1	$(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$
No structure	100	0	2	1	0.0	1	1
Base Case	200	20	0	5	1.0	1	$(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$
Large standard deviation	200	20	0	5	1.0	3	$(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$
Large standard deviation	200	20	0	5	1.0	5	$(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$
Irrelevant features	200	20	10	5	1.0	1	$(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$
Irrelevant features	200	20	20	5	1.0	1	$(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$
Irrelevant features	200	20	100	5	1.0	1	$(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$
Varying proportions	200	20	0	5	1.0	1	$(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{16})$
Varying proportions	200	20	0	5	0.4	1	$(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{16})$
Small N, large P	50	500	0	5	1.0	1	$(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$
Small N, large P	50	500	0	5	0.2	1	$(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$

Table 1: Parameters defining the simulation scenarios as used in generating data and labels.

to be standard normally distributed presenting evidence for within chain stationarity. If a chain's Geweke convergence diagnostic passed a Shapiro-Wilks test for normality (Shapiro and Wilk, 1965) (based upon a threshold of 0.05), it was considered to have achieved stationarity and to be included in the model performance analysis.

$\hat{R}$ is expected to approach 1.0 if the set of chains are converged. Low $\hat{R}$ is not sufficient in itself to claim chain convergence, but values above 1.1 are clear evidence for a lack of convergence (Gelman et al., 2013). Vats and Knudson (2018) show that this threshold is significantly too high (1.01 being a better choice) and propose extensions to $\hat{R}$ that enable a more formal rule for a threshold. It is their method as implemented in the R package `stableGR` (Knudson and Vats, 2020) that is the final check of convergence.

We focus upon stationarity of the continuous variables. This is as convergence of the allocation labels is difficult due to the label-switching problem. In our simulations, the only recorded continuous variable is the concentration parameter of the Dirichlet distribution for the component weights.

The samples from the "stationary" chains were pooled and a Posterior similarity matrix created. There are three possibilities to consider this decision under.

- The chains are converged and agree upon the distribution sampled.
- The chains are not in agreement upon the partition sampled, becoming trapped in different modes. However, a mode does dominate being the mode present in a majority of chains.
- The chains are not in agreement and no one mode dominates among chains.

In the first case pooling has no effect upon the predicted clustering compared to using any one chain. In the second case it feels natural that one would use

the mode that dominates. Pooling the samples effectively does this for the predictive performance of the method as the mode with the greatest number of samples across the chains dominates, however the uncertainty for this mode is increased. In the third case the analysis is non-trivial and further thought, chains and samples would be required. Thankfully the generating process used means that only a small number of modes ever emerge in the PSMs and this case does not arise (based upon the simulations we investigated).

2.2 Consensus clustering analysis

A range of ensembles are investigated. All combinations of chain depth, $R = (1, 10, 100, 1000, 10000)$, and the number of chains, $S = (1, 10, 30, 50, 100)$ were used, a total of 25 different ensembles. A consensus matrix was constructed from the samples generated by each ensemble by finding the proportion of samples within which any pair of items are coclustered.

2.3 `Mclust`

`Mclust` was called using the default settings and a range of inputs for the choice of K. This was $K = (2, \ldots, \min(\frac{N}{2}, 50))$. The choice of $K = \min(\frac{N}{2}, 50)$ was made to mirror the default value of $K_{max} = 50$ used for the overfitted mixture models, with the limit of $\frac{N}{2}$ to avoid fitting 50 clusters in the *Small N, large P* scenario where $N = 50 = K_{max}$. The model choice is performed using the Bayesian Information Criterion (Schwarz et al., 1978, as implemented in `Mclust`).

2.4 Model performance

The different models (Bayesian (pooled), `Mclust` and the 25 Consensus clustering ensembles) were compared under their ability to predict the generating clustering and their uncertainty about this quantity.

3 Yeast

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

- microarray profiles of RNA expression from Granovskaia et al. (2010). This a cell cycle dataset that comprises measurements taken at 41 time points (the **Timecourse** dataset).
- Chromatin immunoprecipitation followed by microarray hybridization (**ChIP-chip**) data from Harbison et al. (2004). This dataset has 117 features.
- Protein-protein interaction (**PPI**) data from BioGrid (Stark et al., 2006). This dataset has 603 features.

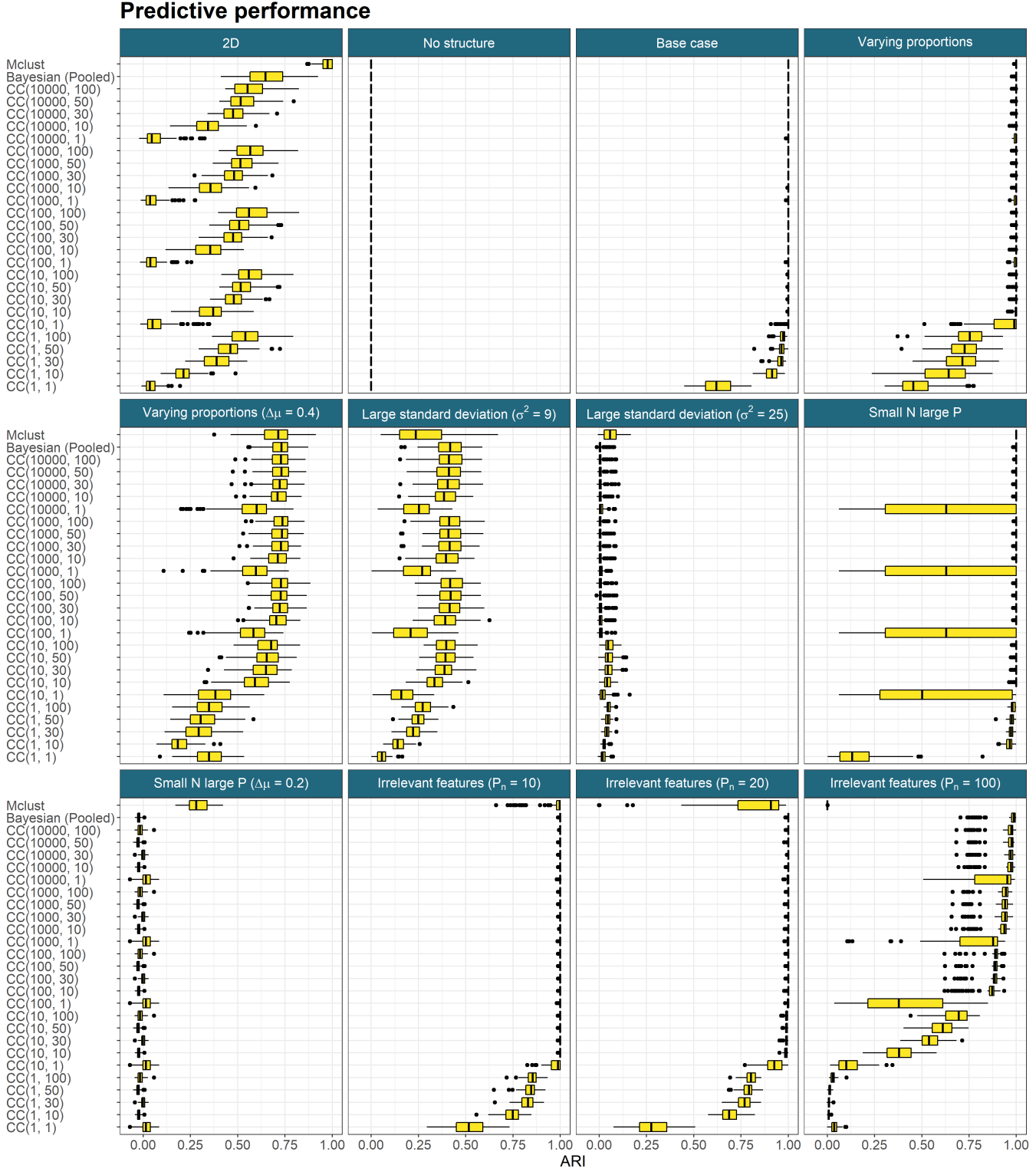

Figure 3: Predictive performance across all simulations. $CC(R, S)$ denotes consensus clustering using the R^{th} sample from S different chains.

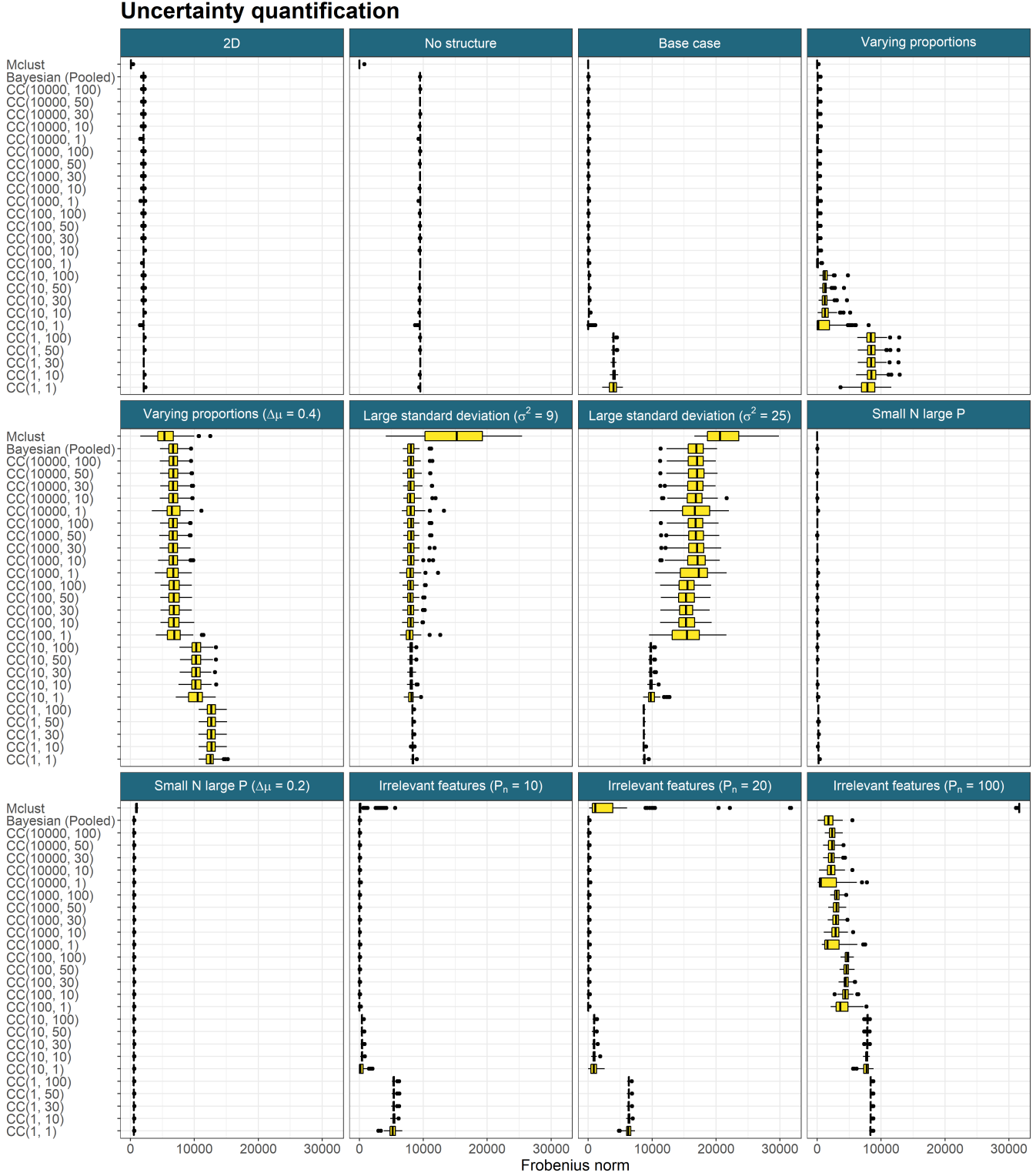

Figure 4: Frobenius norm across simulations. $CC(R, S)$ denotes consensus clustering using the R^{th} sample from S different chains. Lower values are better.

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.

3.1 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. Similar to section 2.1 these chains were investigated for

- within-chain stationarity using the Geweke convergence diagnostic (Geweke et al., 1991), and
- across-chain convergence using $\hat{R}$ (Gelman et al., 1992) and the Vats-Knudson extension (*stable* $\hat{R}$, Vats and Knudson, 2018).

Again we focus upon stationarity of the continuous variables. In MDI, the continuous variables consist of the concentration parameters of the Dirichlet distribution for the dataset-specific component weights and the ϕ_{ij} parameter associated with the correlation between the i^{th} and j^{th} datasets.

We plot the Geweke-statistic for each chain in figure 5 and the series of the ϕ parameters alone in figure 6, excluding the most poorly behaved chain (chain 9). Very few of the chains appear to be truly stationary, but some behave far worse than others. Based upon this we exclude chains 1, 2, 4, 6 and 9, restricting the analysis to the 5 better, if not ideally, behaved chains. Further evidence that even these chains are not converged can be see in figure 7, where the values of $\hat{R}$ do not drop below 1.25 for the ϕ parameters. Stable $\hat{R}$ is also too high, with several million more samples recommended before convergence is expected.

Investigating the Posterior similarity matrices (PSMs) we can see that the Timecourse data appears to have only the mildest of disagreement between the PSMs from different chains. The lack of convergence between chains emerges in the ChIP-chip data and, to a far greater degree, in the PPI data.

3.2 Consensus clustering analysis

We investigate an ensemble of depth $R = 1001$ and width $S = 10000$. The consensus matrices for this ensemble was compared to those for the combinations of $R = (1, 101, 501, 1001, 5001, 10001)$, $S = (1, 100, 500, 1,000)$ in the three datasets. We use a heuristic to decide if the ensemble is sufficiently deep and wide to stop growing. For a given depth r and width s, if there is no visible difference between the consensus matrices from the ensembles using $R = (ar, r)$, $S = (s, bs)$ (in our analysis we used $a = b = 0.5$, but the smaller the choice of a, b the more extreme the stopping criterion), then we consider the ensemble to have

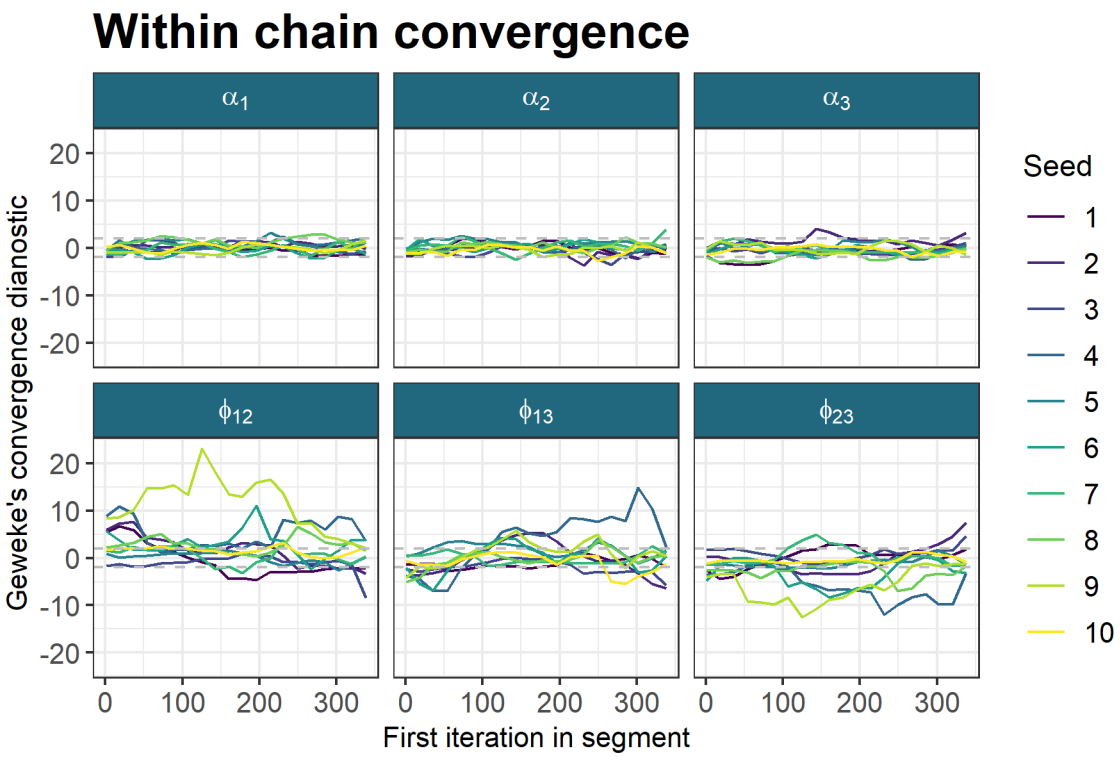

Figure 5: Chain 9 can be seen to have the most extreme behaviour in the distribution of the Geweke diagnostic for ϕ_{12}, ϕ_{13} and ϕ_{23}. We remove this chain from the analysis. We also see that is in these same variables that the chains reveal poor behaviour and focus on these.

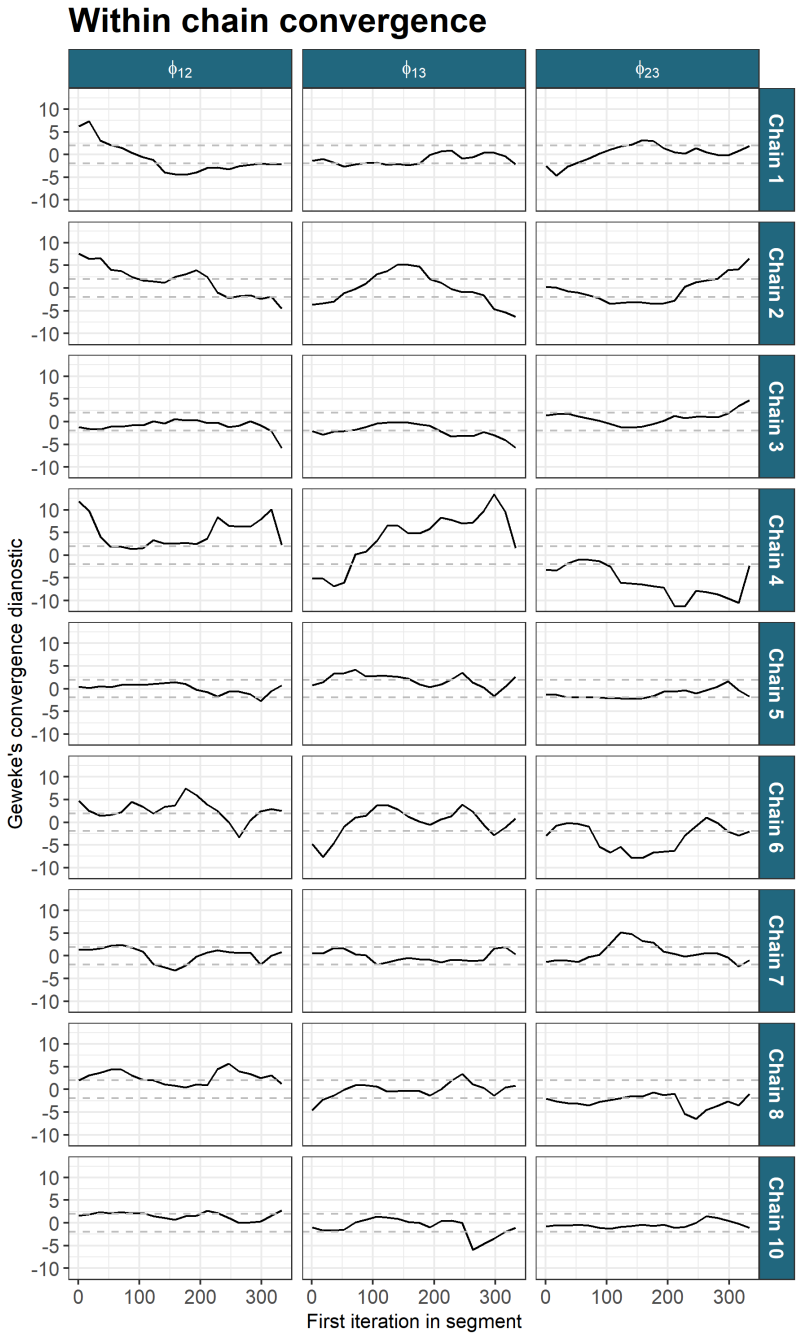

Figure 6: None of the chains appear to be standard normal in their distribution. Chain 4 behaves very strangely and is also dropped from the analysis. Of the remaining chains there is less clear distinctions, but chains 1, 2, and 6 appear most extreme and thus are dropped.

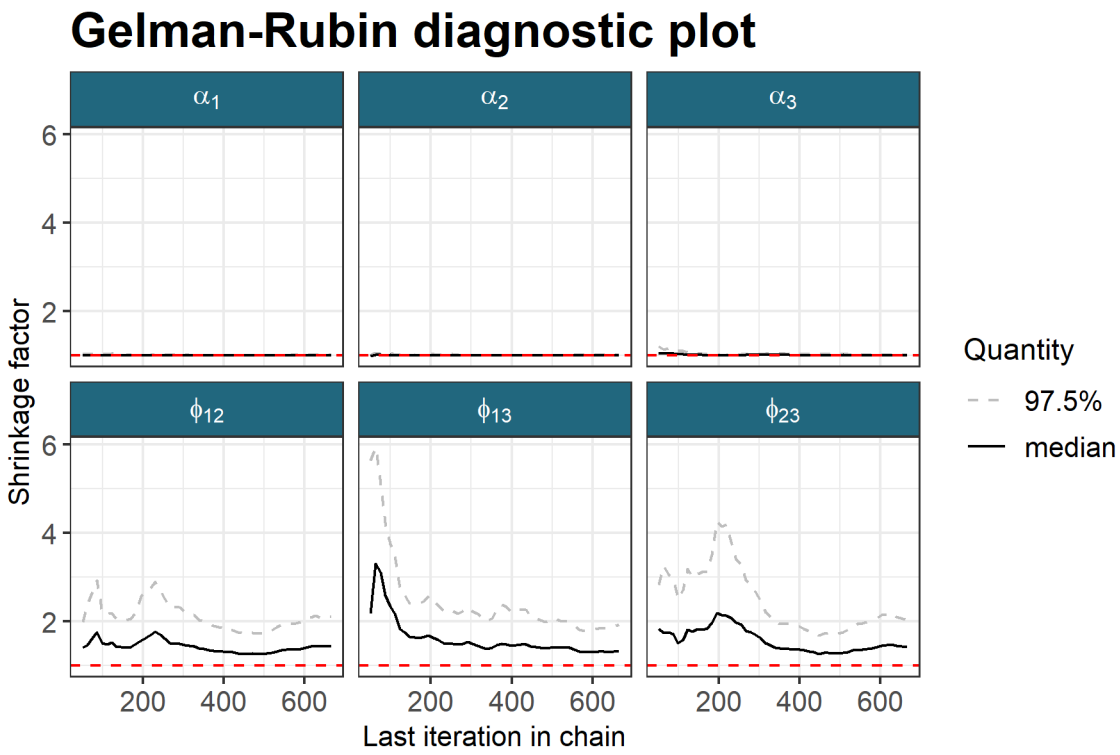

Figure 7: The chains still appear to be unconverged with $\hat{R}$ remaining above 1.25 for the ϕ_{12}, ϕ_{13} and ϕ_{23} parameters. Stable $\hat{R}$ is also too high with values of 1.049, 1.052 and 1.057.

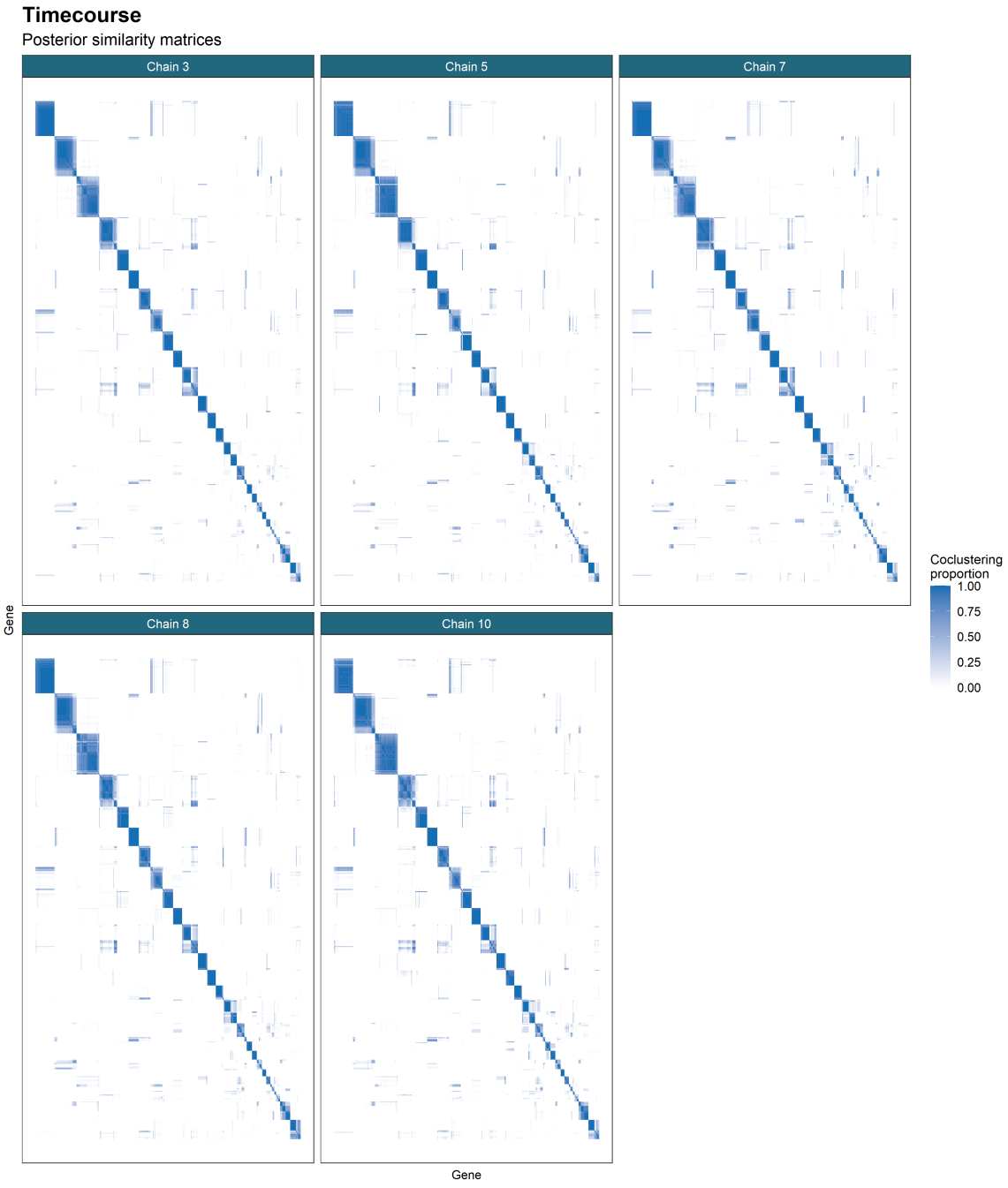

Figure 8: No marked difference.

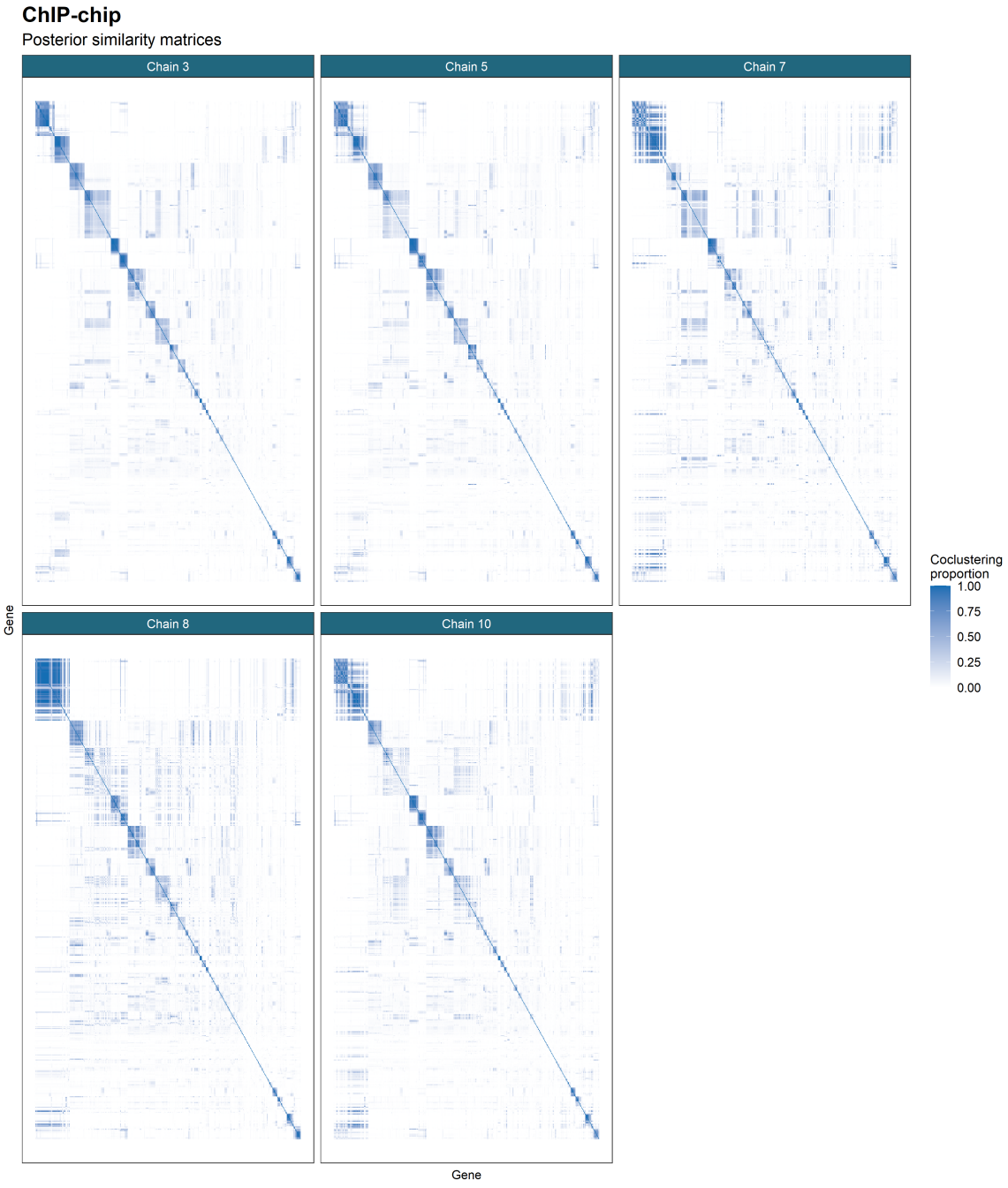

Figure 9: Some difference.

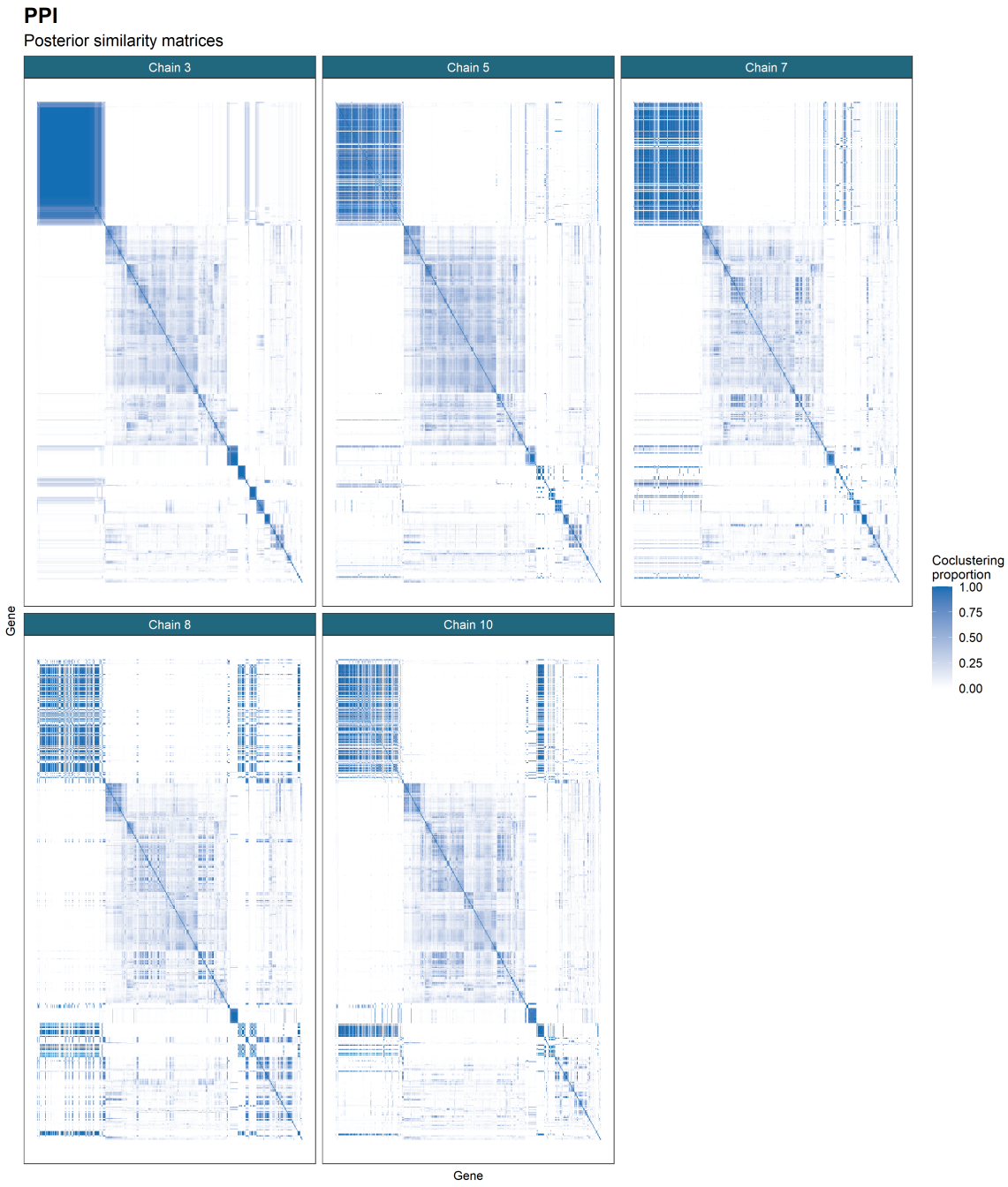

Figure 10: These PSMs have very large disagreements between eachother. There is some common agreement in the square in the centre of each plot. However, the other sections (which consist of the most confident allocations) appear to completely fail to overlap. These sections appear to be approximately random in the partition defined.

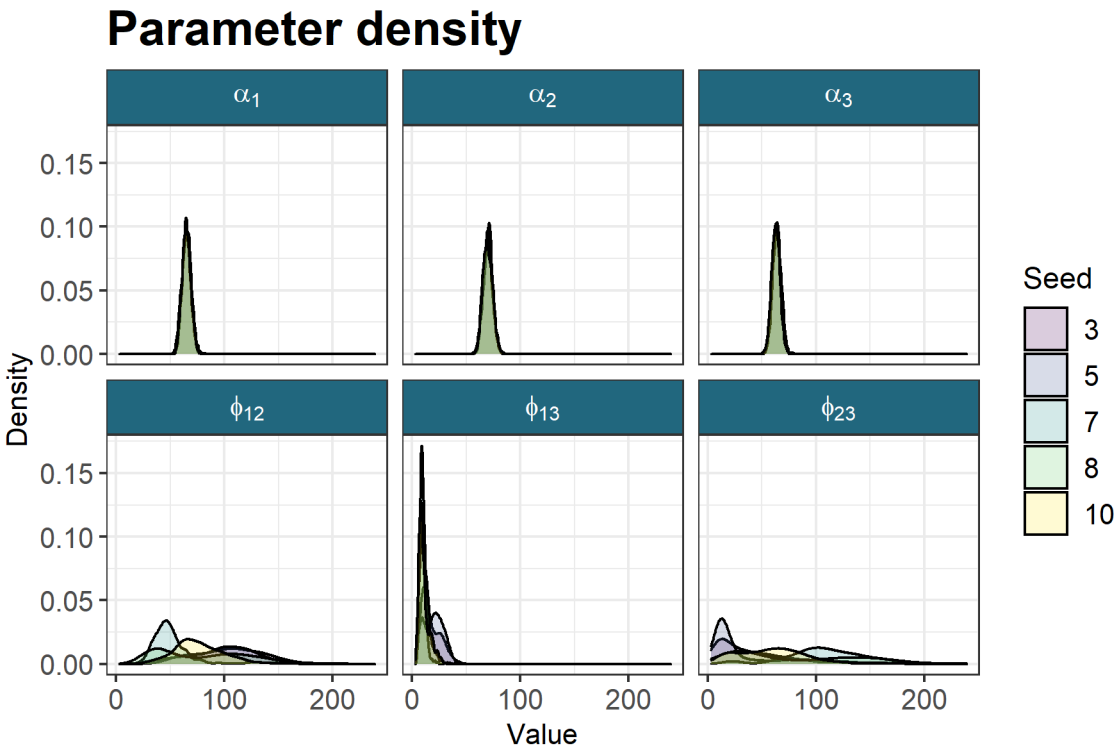

Figure 11: The densities of the continuous variables across the 5 chains kept for analysis. The mean sampled values are $\alpha_1 = 64.84$, $\alpha_2 = 69.85$, $\alpha_2 = 63.22$, $\phi_{12} = 81.76$, $\phi_{13} = 13.87$, and $\phi_{23} = 65.03$. It can be seen that different modes are being sampled for the ϕ parameters in each chain.

stabilised. This is inspired by the belief that a clustering method should produce stable results across similar datasets (Von Luxburg and Ben-David, 2005; Meinshausen and Bühlmann, 2010). We believe that if the method is still producing a partition that is visibly changing for additional chains and depth, than the random initialisation is influencing the result sufficiently that it is unlikely to be stable for similar datasets or reproducible for a random choice of seeds. An example of this logic can be seen in figures 13 and 14 (and to a lesser degree in figure 12). Here the decision to stop growing the ensemble is made as there is no apparent gain in increasing chain depth from $R = 5001$ to $R = 10001$, but it can be seen that a chain depth of $R = 1001$ is insufficient as there is a marked difference in the consensus matrices for the PPI dataset particularly between $R = 1001$ and $R = 5001$. The number of chains appears required appears to have stabilised quickly, as there is no obvious change in increasing S from 100.

If we compare the distribution of sampled values for the ϕ parameters for the Bayesian chains that we keep based upon their convergence diagnostics, the final ensemble used ($R = 10001$, $S = 1000$) and the pooled samples from the 5 long chains, then we see that the ensemble consisting of the long chains (which might be believed to sampling different parts of the posterior distribution) is closer in its appearance to the distributions sampled by the Consensus clustering than to any single chain.

3.3 GO term over-representation

To validate our analysis we test if the predicted clusters have a higher concentration of specific Gene Ontology (GO) terms than would be expected by chance, conditioning on the background set of the 551 yeast genes in the data. The `Bioconductor` packages `clusterProfiler` (Yu et al., 2012), `biomaRt` (Durinck et al., 2009) and the annotation package `org.Sc.sgd.db` (Carlson et al., 2014) were used. Clusters were predicted from the Posterior similarity matrices of the chains kept from section 3.1.1 and the consensus matrix of the largest ensemble run (i.e. $CC(10001, 1000)$). The gene labelled YIL167W was not found in the annotation database and was dropped from the analysis leaving a background universe of 550 genes. A hypergeometric test was used to check if the number of genes associated with specific GO terms within a cluster was greater than expected by random given the 550 possible genes. The false discovery rate of this test was controlled using the Benjamini-Hochberg correction (Benjamini and Hochberg, 1995) and significance threshold of 0.05 was used. The over-represented GO terms were then plotted to compare methods. The three different ontologies of "Molecular function" (**MF**), "Biological process" (**BP**) and "Cellular component" (**CC**) were all investigated.

References

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal*

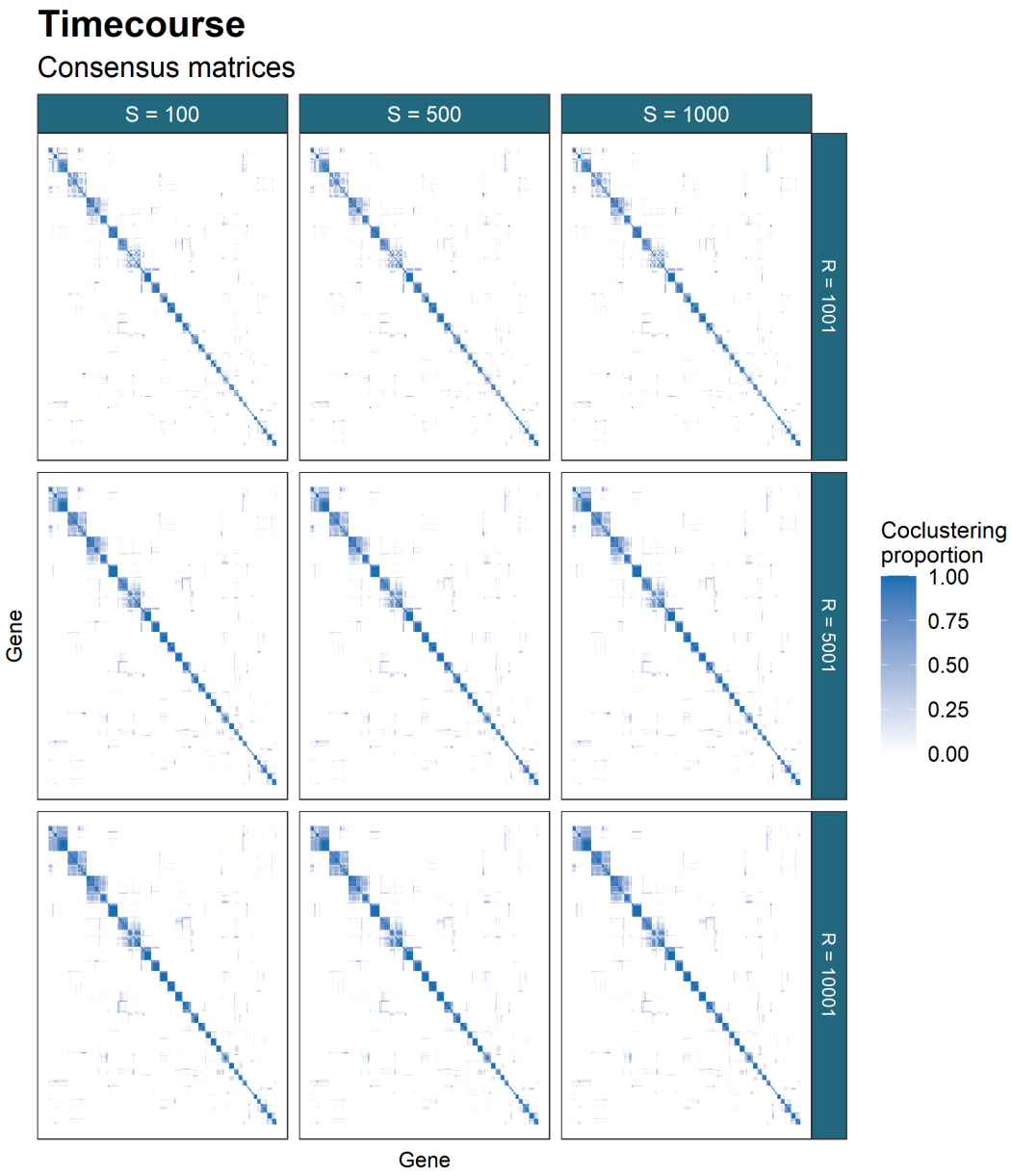

Figure 12: Consensus matrices for different ensembles of MDI for the Timecourse data. This dataset has stable clustering across the different choices of number of chains, S, and chain depth, R, with some components merging as the chain depth increases.

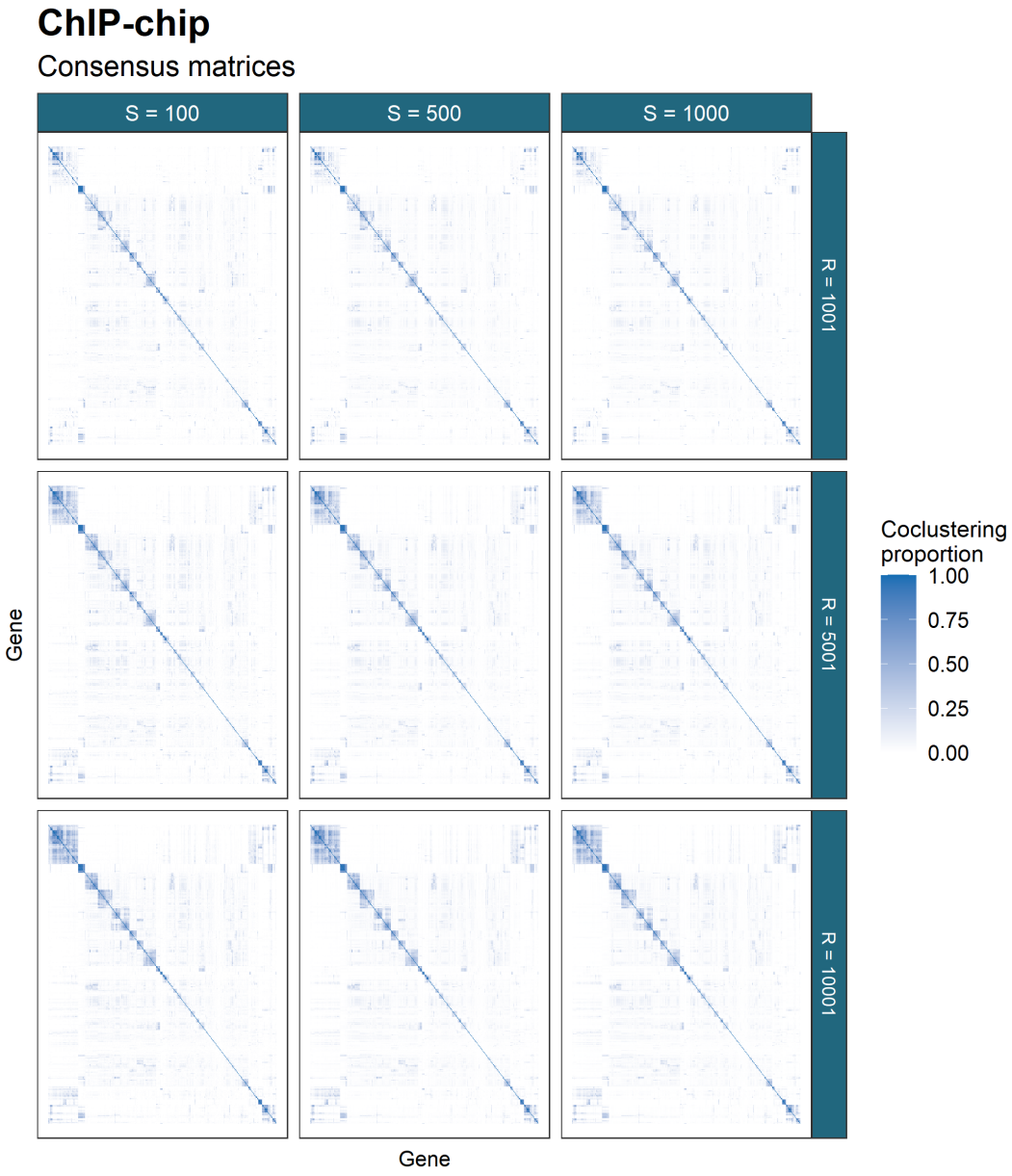

Figure 13: The ChIP-chip dataset is more sparse than the Timecourse data. In keeping with the results from the simulations for mixture models, deeper chains are required for better performance. It is only between $R = 5,001$ and $R = 10,001$ that no change in the clustering can be observed and the result is believed to be stable. In this dataset the number of chains used, S, appears relatively unimportant, with similar results for $S = 100, 500, 1000$.

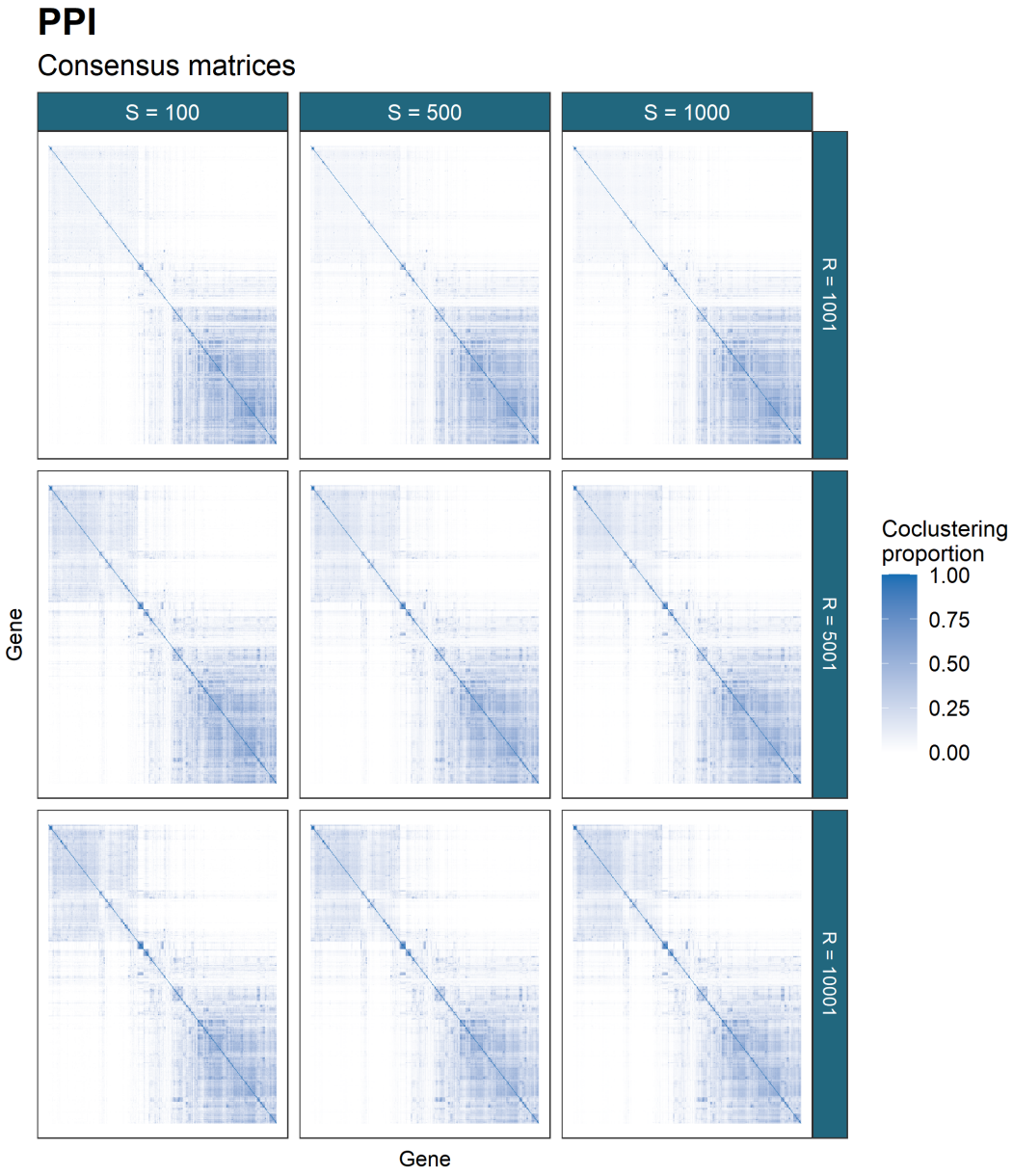

Figure 14: The PPI dataset has awkward characteristics for modelling. A wide, sparse dataset it is again chain depth that is the most important parameter for the ensemble. Similar to the results in figure 13, the matrices only stabilise from $R = 5001$ to $R = 10001$.

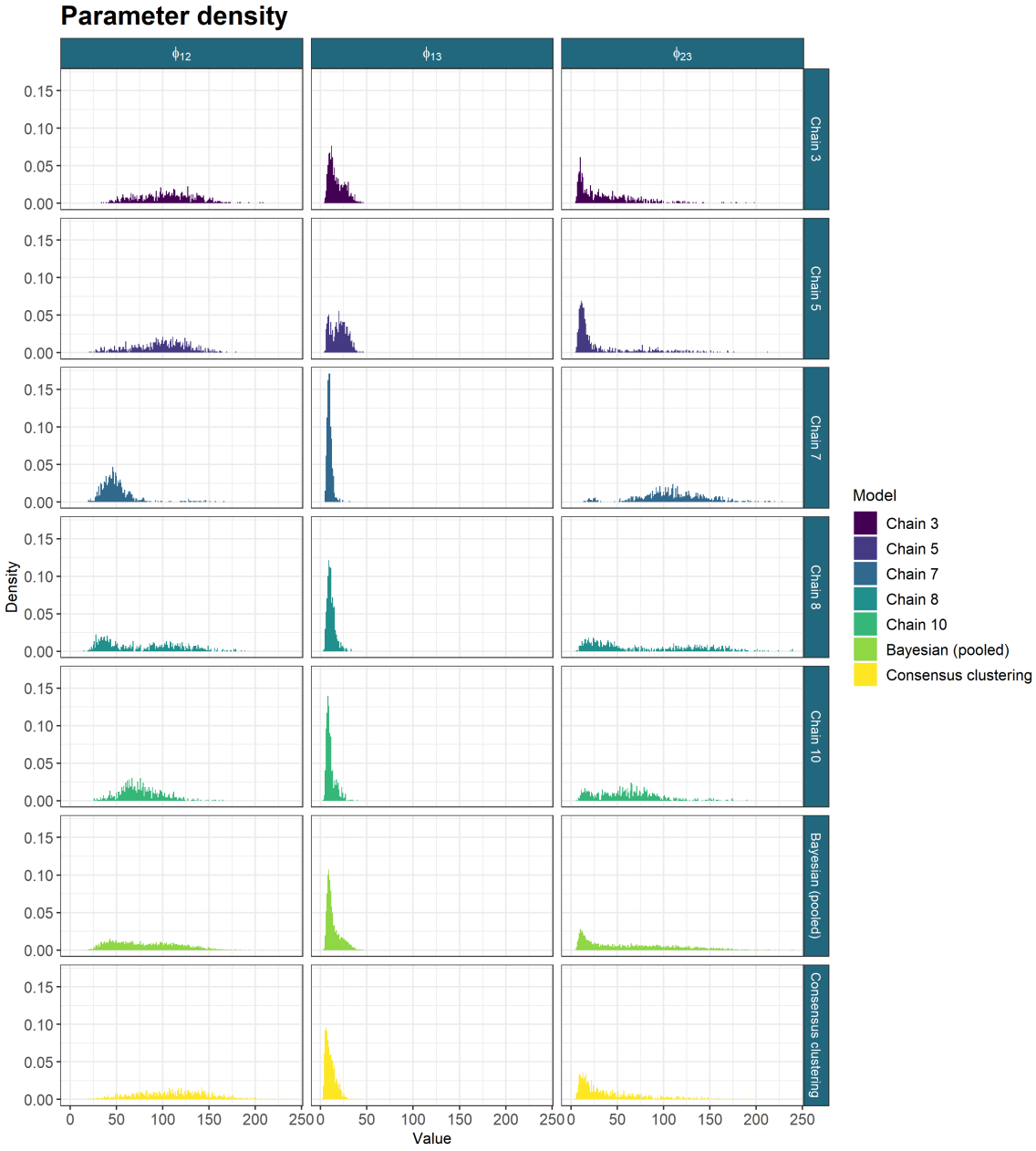

Figure 15: The sampled values for the ϕ parameters from the long chains, their pooled samples and the consensus using 1000 chains of depth 10,001. The long chains display a variety of behaviours. Across chains there is no clear consensus on the nature of the posterior distribution. The samples from any single chain are not particularly close to the behaviour of the pooled samples across all three parameters. It is the Consensus clustering that most approaches this pooled behaviour.

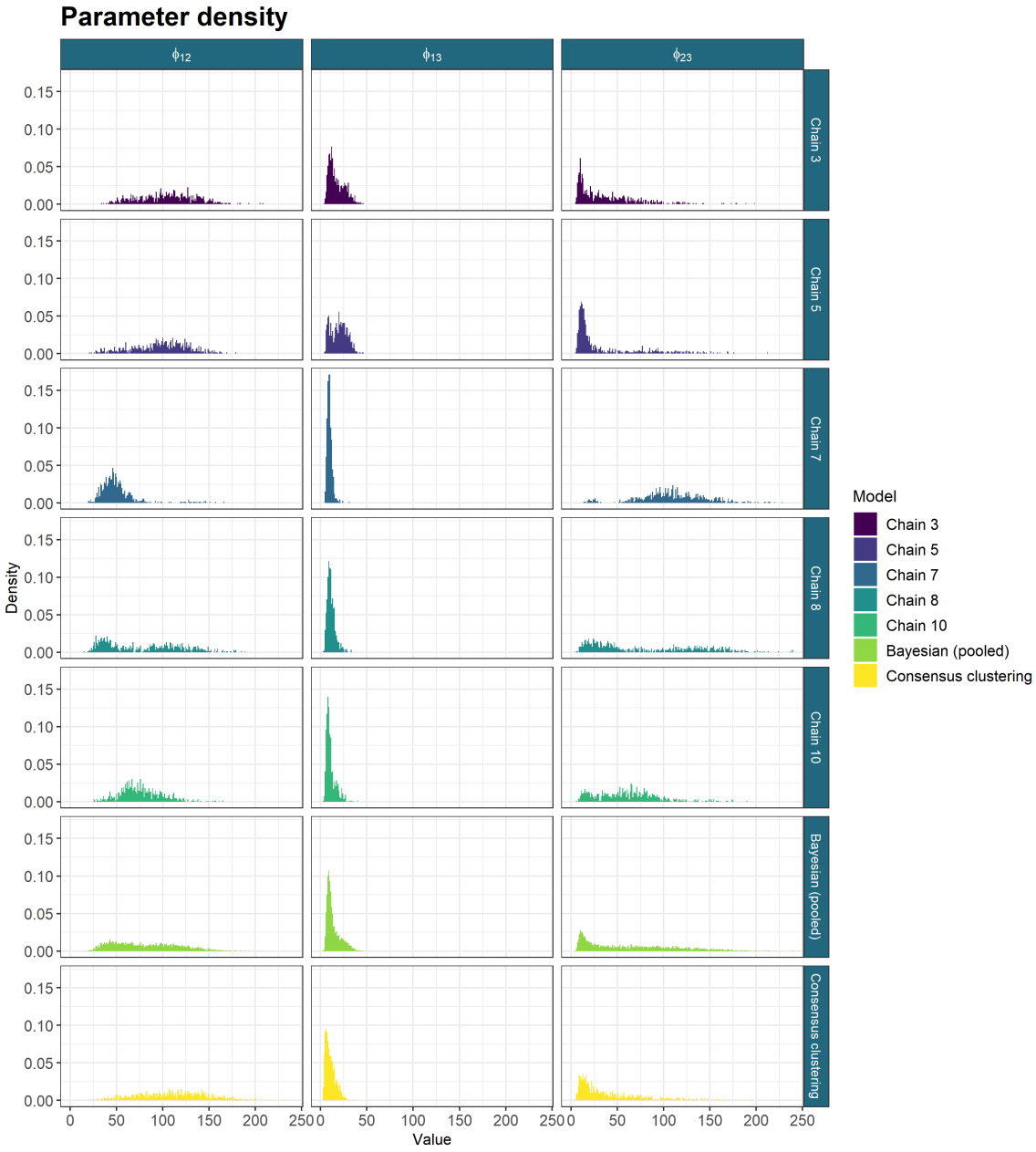
Parameter density
ϕ_{12}
ϕ_{13}
ϕ_{23}
Chain 3
Chain 5
Chain 7
Chain 8
Chain 10
Bayesian (pooled)
Consensus clustering
Density
0.00
0.05
0.10
0.15
0
50
100
150
200
250
Value
Model
Chain 3
Chain 5
Chain 7
Chain 8
Chain 10
Bayesian (pooled)
Consensus clustering

Figure 16: .

statistical society: series B (Methodological), 57(1):289–300, 1995.

M Carlson, S Falcon, H Pages, and N Li. Org. sc. sgd. db: Genome wide annotation for yeast. *R package version*, 2(1), 2014.

Steffen Durinck, Paul T Spellman, Ewan Birney, and Wolfgang Huber. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nature protocols*, 4(8):1184, 2009.

Yasin Şenbabaoğlu, George Michailidis, and Jun Z Li. A reassessment of consensus clustering for class discovery. *bioRxiv*, page 002642, 2014a.

Yasin Şenbabaoğlu, George Michailidis, and Jun Z Li. Critical limitations of consensus clustering in class discovery. *Scientific reports*, 4(1):1–13, 2014b.

Andrew Gelman, Donald B Rubin, et al. Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472, 1992.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

John Geweke et al. *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*, volume 196. Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN, 1991.

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.

Christina Knudson and Dootika Vats. *stableGR: A Stable Gelman-Rubin Diagnostic for Markov Chain Monte Carlo*, 2020. URL `https://CRAN.R-project.org/package=stableGR`. R package version 1.0.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

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.

Ulrike Von Luxburg and Shai Ben-David. Towards a statistical theory of clustering. In *Pascal workshop on statistics and optimization of clustering*, pages 20–26. Citeseer, 2005.

Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He. clusterprofiler: an r package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology*, 16(5):284–287, 2012. doi: 10.1089/omi.2011.0118.