MDI SMC: Progress Report

Nathan Cunningham

August 31, 2016

General algorithm

Algorithm 1 Gibbs sampler

```
Initialise $\Gamma$ matrix of prior allocation weights and $\Phi$ matrix of dataset concordance values
for i = 1, ..., number of iterations do
    Conditional on $\Gamma_{i-1}$ and $\Phi_{i-1}$ update the cluster labels, $c_i$, using alg. 2
    Conditional on $c_i$ update $\Gamma_i$ and $\Phi_i$
end for
```

Algorithm 2 Particle filter to update cluster allocations

```
for i = 1, ..., n do                                                    ▷ Loop over observations
    for m = 1, ..., M do                                                ▷ Loop over particles
        for j = 1, ..., d do                                            ▷ Loop over datasets
            Sample $c_{i,j}^{(m)}$                                      ▷ Propose a cluster for each datum
            $q(c_{i,j}^{(m)} = k) \propto k^*(y_{i,j}|c_{i,j}^{(m)} = k)\gamma_{i,k,j}$
            $\xi^{(m)} = \xi^{(m)} \times \gamma_{i,k,j}(1 + \phi_i)k^*(y_{i,k}|c_{i,j}^{(m)} = k)$
        end for
    end for
    Resample particles according to $\xi^{(m)}$
end for
Update cluster labels using allocation in particle with largest $\xi^{(m)}$
```

Where

$$k^*(y_{i,k}|c_{i,k}^{(m)} = k) = (\mathbf{y_{i,k}} - \mu_\mathbf{k})\mathbf{\Sigma^{-1}}(\mathbf{y_{i,k}} - \mu_\mathbf{k})^\top \tag{1}$$

$$\Phi \text{ is a measure of cluster label correspondence across datasets} \tag{2}$$

$$\gamma_{i,k,j} \text{ is a prior weight for assigning observation i, in dataset k to cluster j} \tag{3}$$

1 Algorithmic issues

- Particle resampling step is currently disabled as it can crash the algorithm—possibly related to `NA` logweights being assigned.
- Not currently taking advantage of parallel computing which means the algorithm is slower than it could be. It scales linearly with the number of particles.
- As the MDI algorithm originally was not set up for multiple particles, a single particle is selected at the end of the particle filter (based on $\xi^{(m)}$) and that is fed into the next step of the Gibbs sampler. A possible solution would be to treat the particles as separate datasets, so all the particles are returned and the Γ and Φ values are updated based on all of the particles.

2 Multinomial data

Extended the algorithm to work on multinomial data.

For each gene an observation is made on each of Q features, each of which take a value r from the set $\{1, \ldots, R\}$. Denote the cluster-specific probability of getting value r for feature q by $\theta_{rq}^{(m)}$, such that $\Sigma_{r=1}^{R} \theta_{rq}^{(m)} = 1$ for each particle, m. Adopt a $Dirichlet\{\beta_{1q}^{(m)}, \ldots, \beta_{Rq}^{(m)}\}$ prior for $\theta_{1q}^{(m)}, \ldots, \theta_{Rq}^{(m)}$. The Dirichlet prior hyperparameter, $\beta_{rq}^{(m)}$ is taken as 0.5. Denote the resulting prior probabilities as $b_{rq}^{(m)}$

Assuming independence between features, the probability of assigning the i^{th} observation to a cluster, j, in particle m is calculated as:

$$p(c_i^{(m)} = j | x_{1:i}^{(m)}, c_{1:(n-1)}^{(m)}, b_{1q}^{(m)}, \ldots, b_{Rq}^{(m)}, a^{(m)}) = \prod_{q=1}^{Q} \frac{(n_{j,x_i}^{(m)} \times a^{(m)}) + (b_{rq}^{(m)} \times (1 - a^{(m)}))}{(n_{j,\cdot}^{(m)} \times a^{(m)}) + (1 \times (1 - a^{(m)}))} \tag{4}$$

Where $n_{j,x_i}^{(m)}$ is the number of observations in cluster j with the same value as $x_i^{(m)}$, $n_{j,\cdot}^{(m)}$ is the total number of observations in cluster j and $a^{(m)}$ is a particle-specific centring measure. This is a weighted average of the empirical probability of the value of x_i in the cluster j, and the prior probability.

Output of the algorithm clustering multinomial and gaussian data is presented in Figures 3 and 3.

3 Markov chain diagnostics

Convergence of the markov chain can be examined by viewing plots of the allocation agreement. Figure 3 shows the pairwise agreement of genes in clusters ("fusion"). Although these are relatively stable there appears to be some evidence of an upward trend, so it may be that the algorithm has not yet converged. The algorithm was run for 1000 iterations.

Figure 3 shows the distribution of pairwise allocation agreements. The x axis indicates the number of fused genes across the two datasets, while the height represents the frequency this occurs in the Markov chain. The number of genes fused appears to be uniformly distributed, when it should be hoped that in most cases a large number of genes are fused.

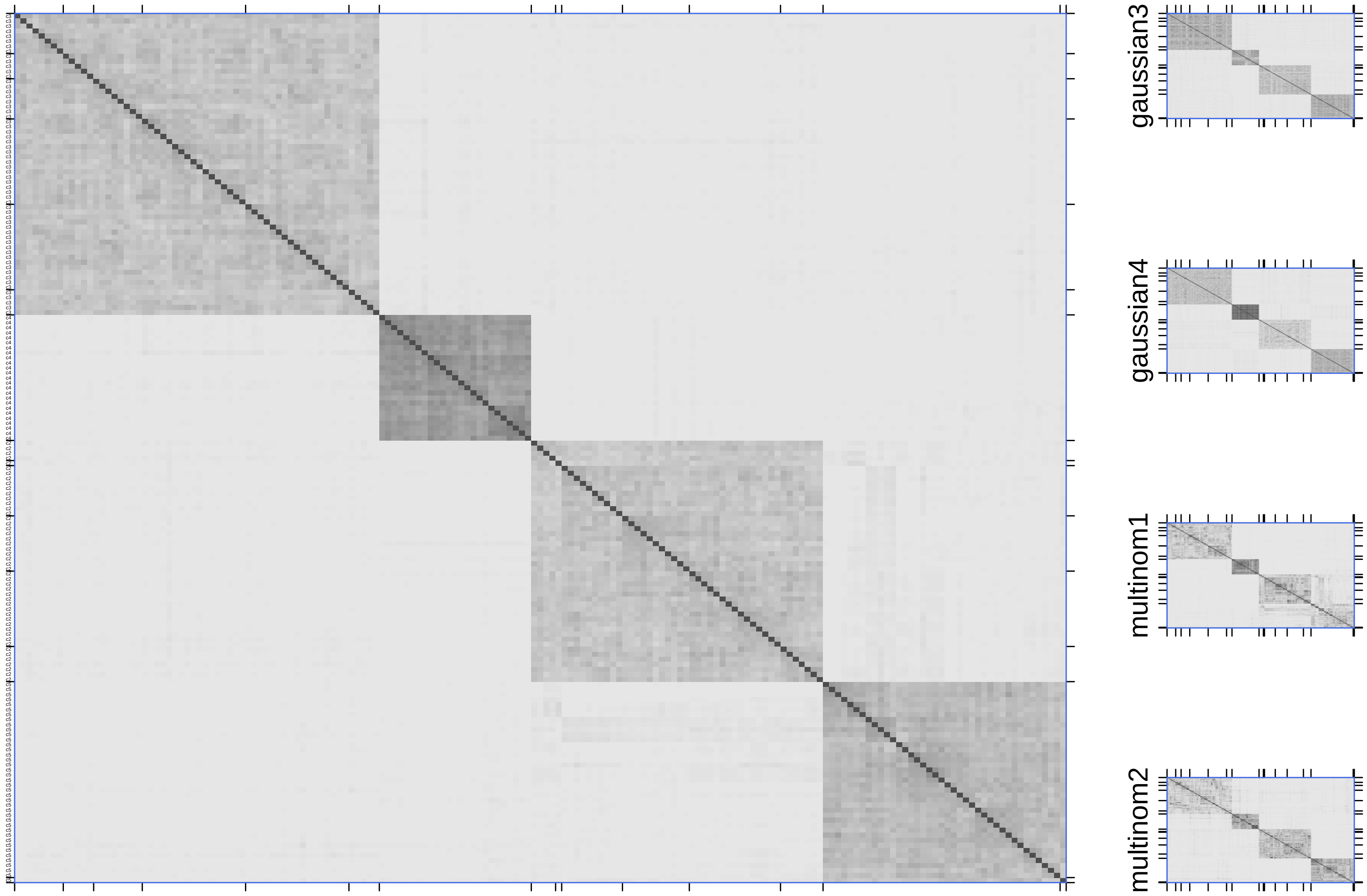

Figure 1: Running MDI using four datasets: two gaussian and two multinomial, with a shared clustering order. The algorithm is attempting to partition the data into 15 clusters, but appears to be recovering the four true clusters.

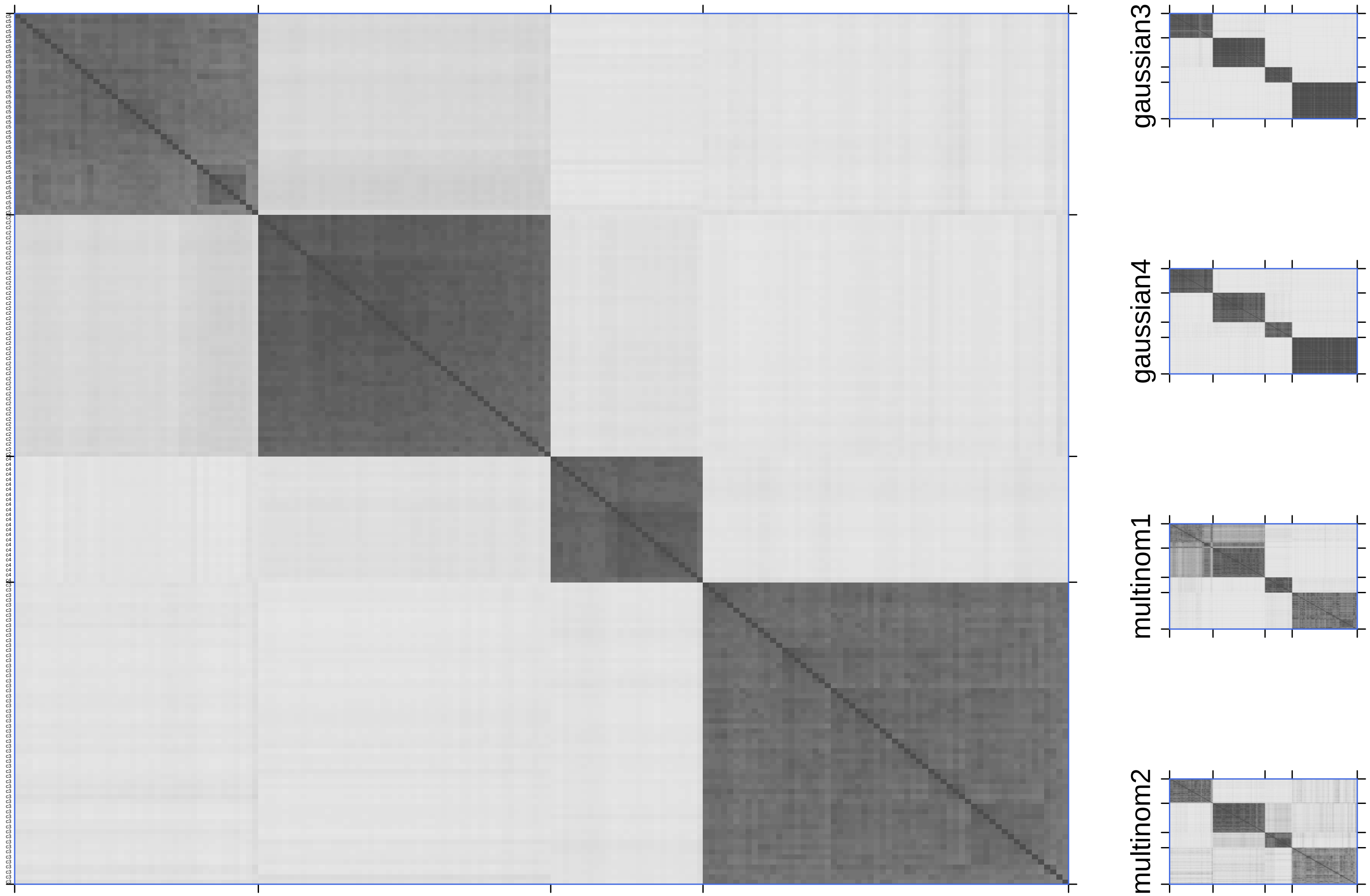

Figure 2: Running MDI using four datasets: two gaussian and two multinomial, with a shared clustering order. The algorithm is attempting to partition the data into four clusters and is recovering the true clustering.

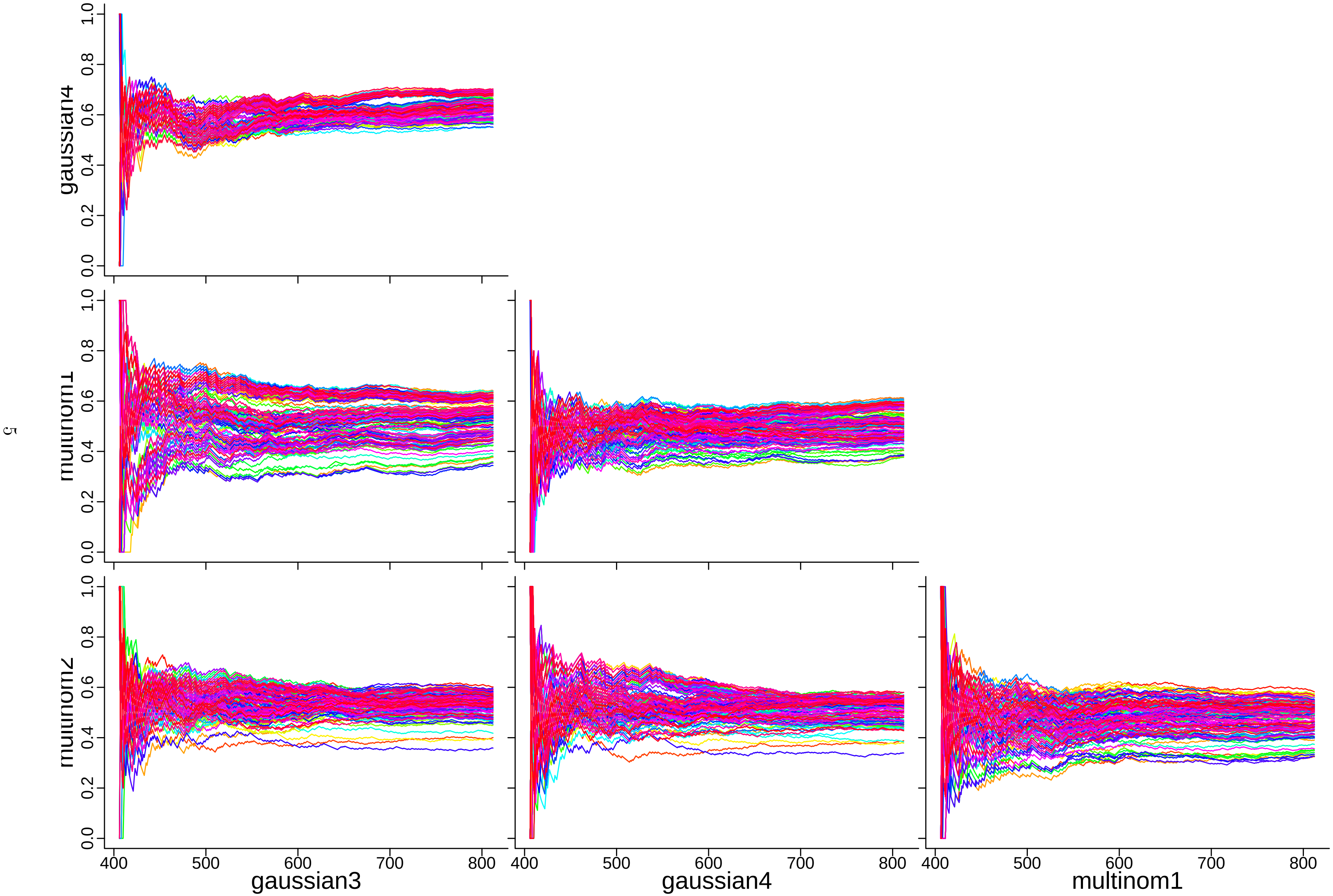

Figure 3: Allocation agreement for MDI run using 4 datasets

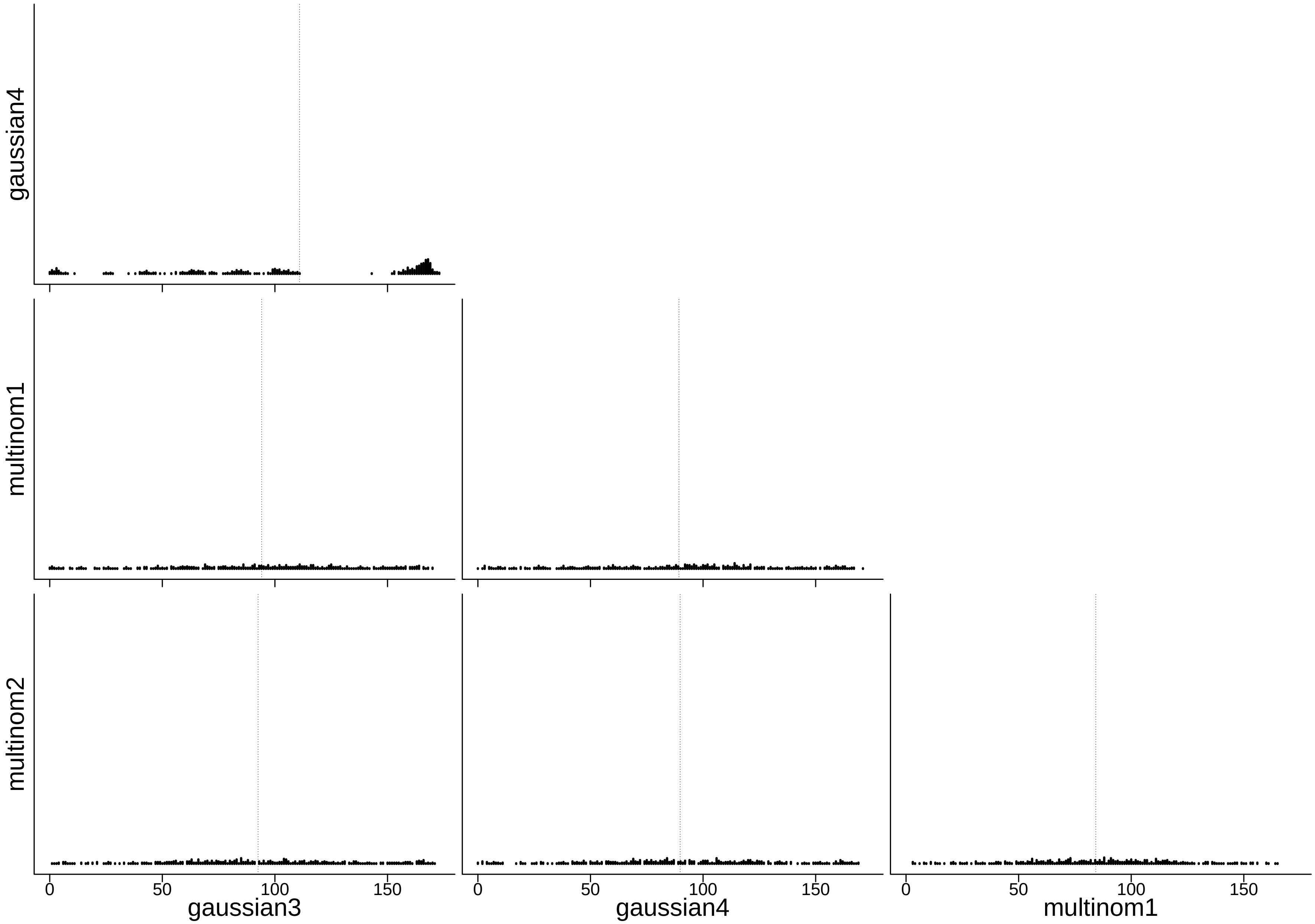

Figure 4: Allocation agreement histogram for MDI run using 4 datasets