MDI SMC: Progress Report

Nathan Cunningham

September 7, 2016

1 What have I done?

- Implemented the multinomial resampling step, carried out after each 'per-gene' iteration. This doesn't appear to have corrected the sensitivity of the algorithm to the ordering of the data. The Griffin [1] paper mentions randomly permuting the data also, "the data were randomly permuted and the SMC algorithms was run with 5000 particles".
- The particle Gibbs steps is a work in progress still. The approach I'm taking is to hold one of the particles as fixed, then resample the allocation holding this reference trajectory fixed.
- I examined the output from the original MDI on the same data as I'd tested the SMC algorithm on. The output appears to converge quickly to complete fusion across clusters. Figures presented below.
- The weight updates, ξ, have not been changed, as I think they are being updated correctly in the code, but I might have written the math inaccurately. Algorithm updated below.

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
            $q(c_{i,j}^{(m)} = k) \propto k^*(y_{i,j}|c_{i,j}^{(m)} = k) \times \gamma_{i,k,j}$
        end for
        $\xi^{(m)} = \xi^{(m)} \times \prod_{k=1}^{d-1} \prod_{l=k+1}^{d} (1 + \phi_{kl} \mathbb{1}(c_{ik} = c_{il}) \prod_{j=1}^{d} \gamma_{i,k,j} k^*(y_{i,k}|c_{i,j}^{(m)} = k)$
    end for
    Resample particles according to $\xi^{(m)}$
end for
Select single cluster label allocation according to $\xi^{(m)}$
```

Where

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

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

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

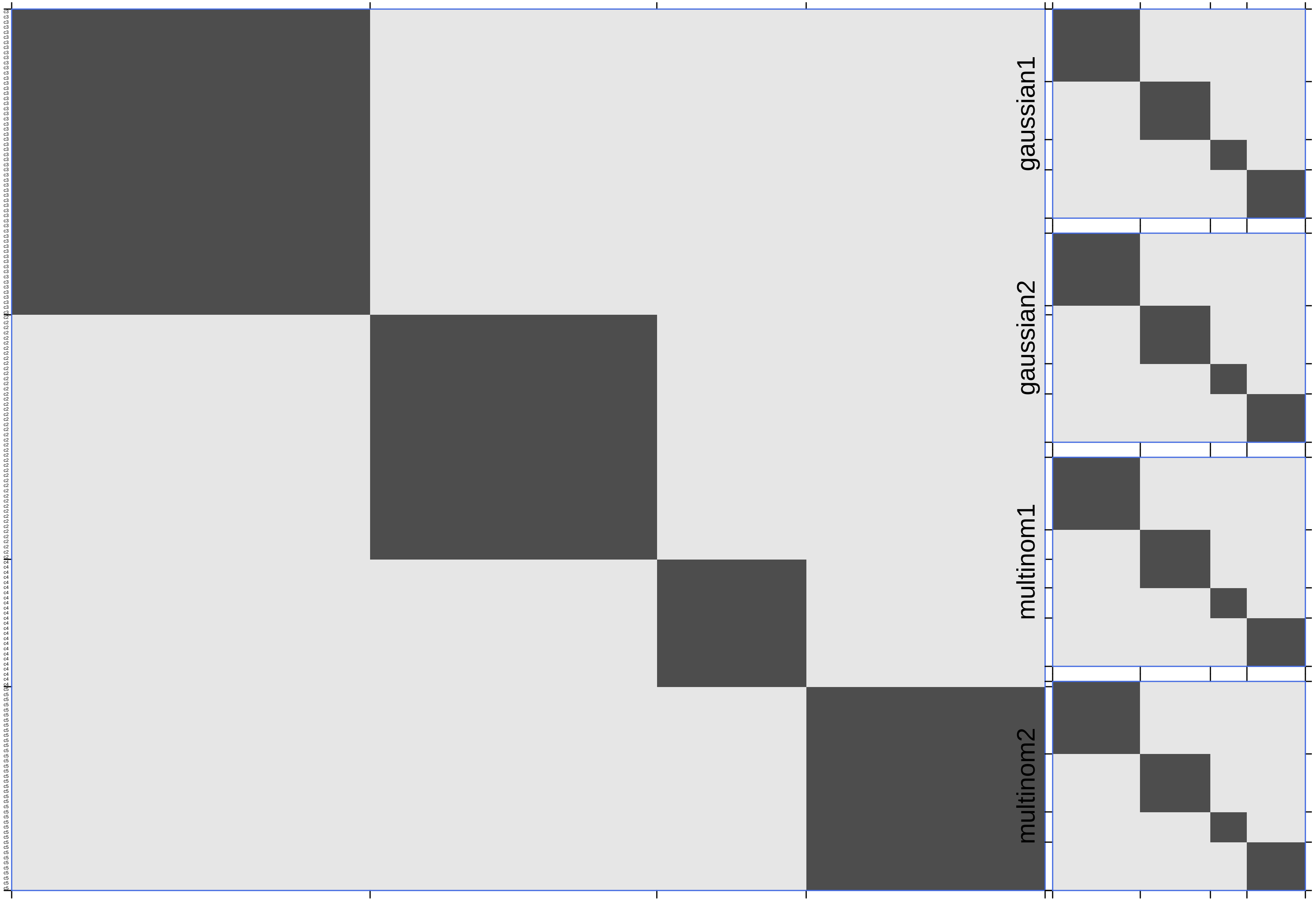

Figure 1: Running the original MDI using four datasets: two gaussian and two multinomial, with a shared clustering order.

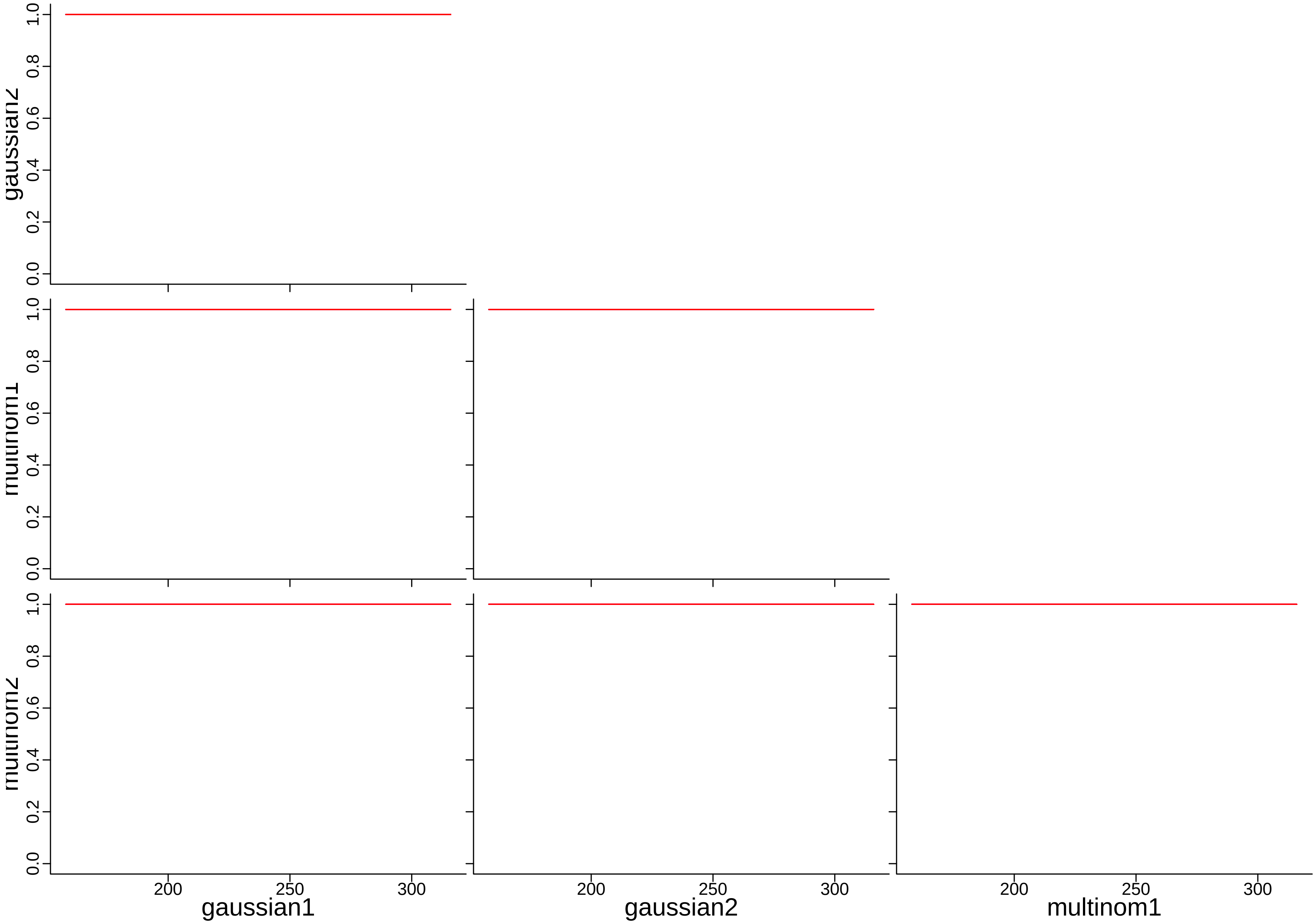

Figure 2: Allocation agreement the original MDI using four datasets: two gaussian and two multinomial, with a shared clustering order. Genes are allocated to the same clusters across datasets.

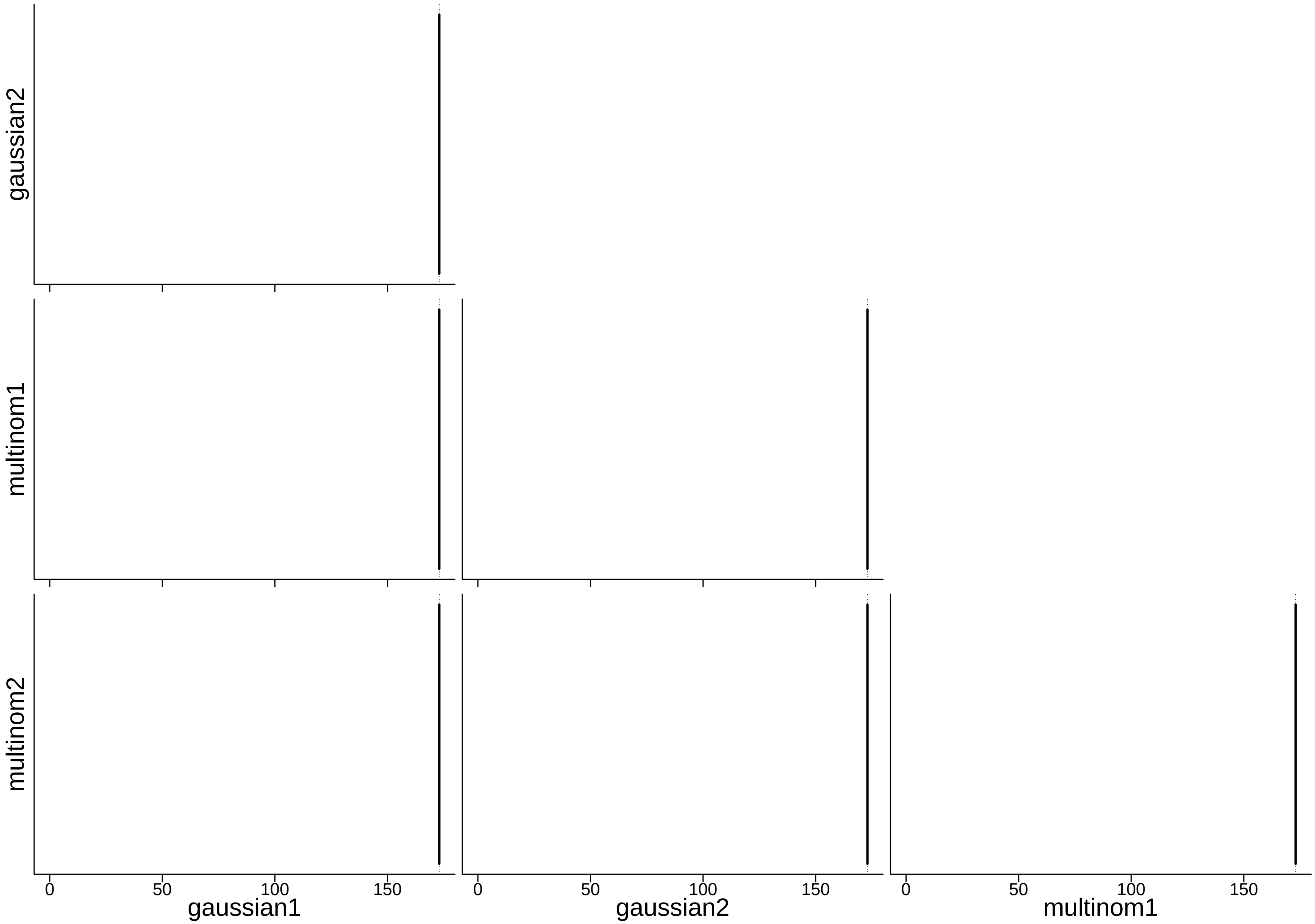

Figure 3: Allocation agreement the original MDI using four datasets: two gaussian and two multinomial, with a shared clustering order. Genes are allocated to the same clusters across datasets.

References

[1] JE Griffin. Sequential monte carlo methods for mixtures with normalized random measures with independent increments priors. *Statistics and Computing*, pages 1–15, 2014.