Population structure inference using phenotype data

Uses

✳ Absence of genotype data
✳ Contrasting with genetic population structure
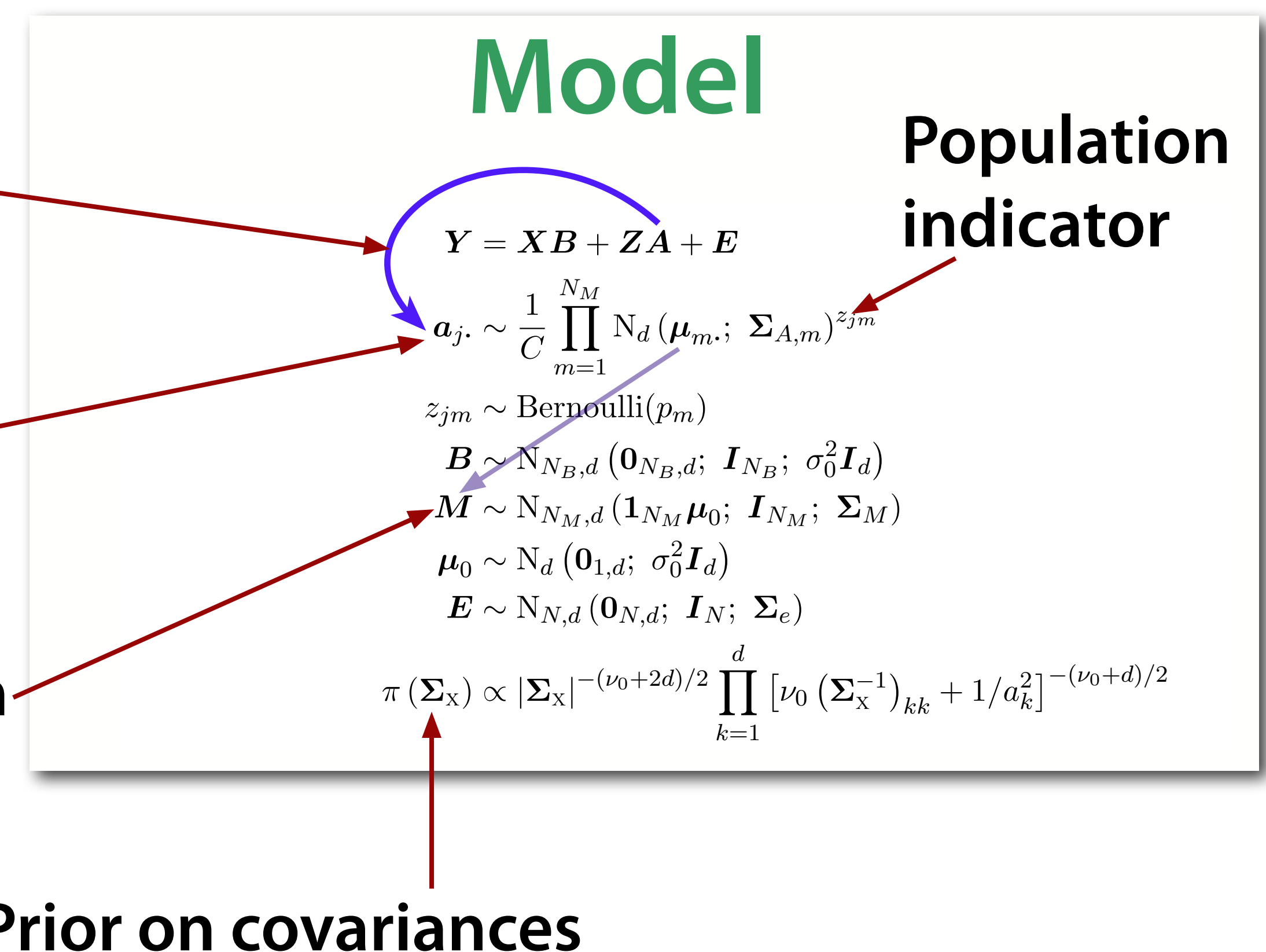
Implementation

✳ Hamiltonian Monte Carlo (NUTS)
✳ R package: MuGaMix
    github.com/tonymugen/MuGaMix

Model

Replicated individuals (e.g., inbred lines)

Population indicator

Multiplicative Gaussian mixture model

Population mean

$$Y = XB + ZA + E$$

$$a_{j.} \sim \frac{1}{C} \prod_{m=1}^{N_M} N_d\left(\mu_m; \; \Sigma_{A,m}\right)^{z_{jm}}$$

$$z_{jm} \sim \text{Bernoulli}(p_m)$$

$$B \sim N_{N_B,d}\left(0_{N_B,d}; \; I_{N_B}; \; \sigma_0^2 I_d\right)$$

$$M \sim N_{N_M,d}\left(1_{N_M}\mu_0; \; I_{N_M}; \; \Sigma_M\right)$$

$$\mu_0 \sim N_d\left(0_{1,d}; \; \sigma_0^2 I_d\right)$$

$$E \sim N_{N,d}\left(0_{N,d}; \; I_N; \; \Sigma_e\right)$$

$$\pi\left(\Sigma_x\right) \propto |\Sigma_x|^{-(\nu_0+2d)/2} \prod_{k=1}^{d}\left[\nu_0\left(\Sigma_x^{-1}\right)_{kk} + 1/a_k^2\right]^{-(\nu_0+d)/2}$$
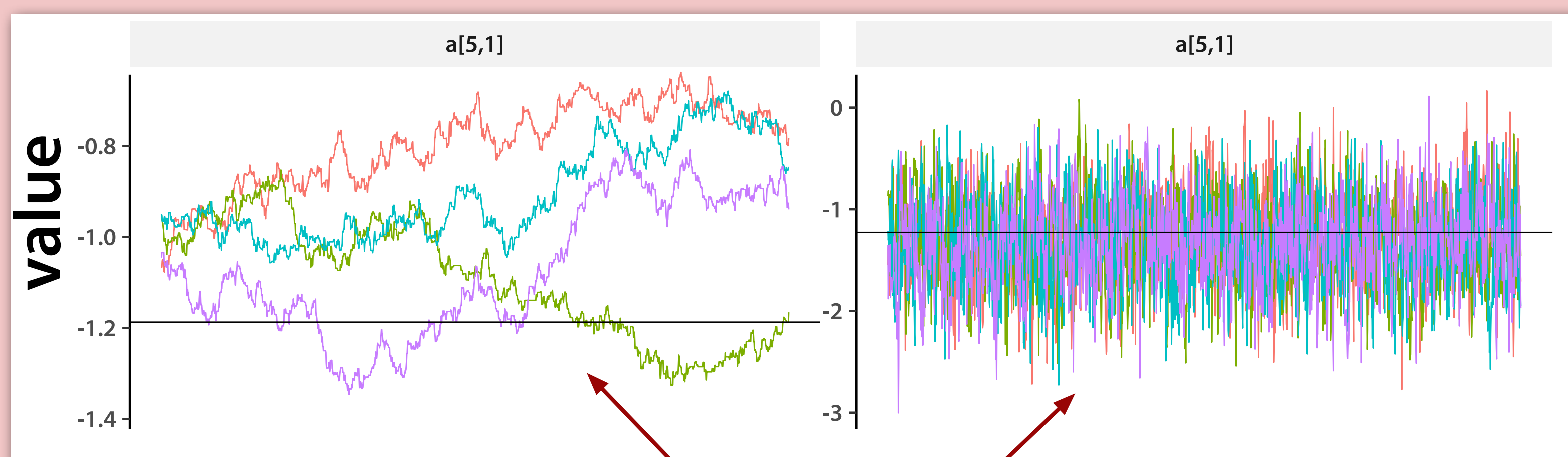
Prior on covariances

Advantage of Hamiltonian Monte Carlo:

Metropolis   NUTS

Dramatic improvement in convergence and mixing
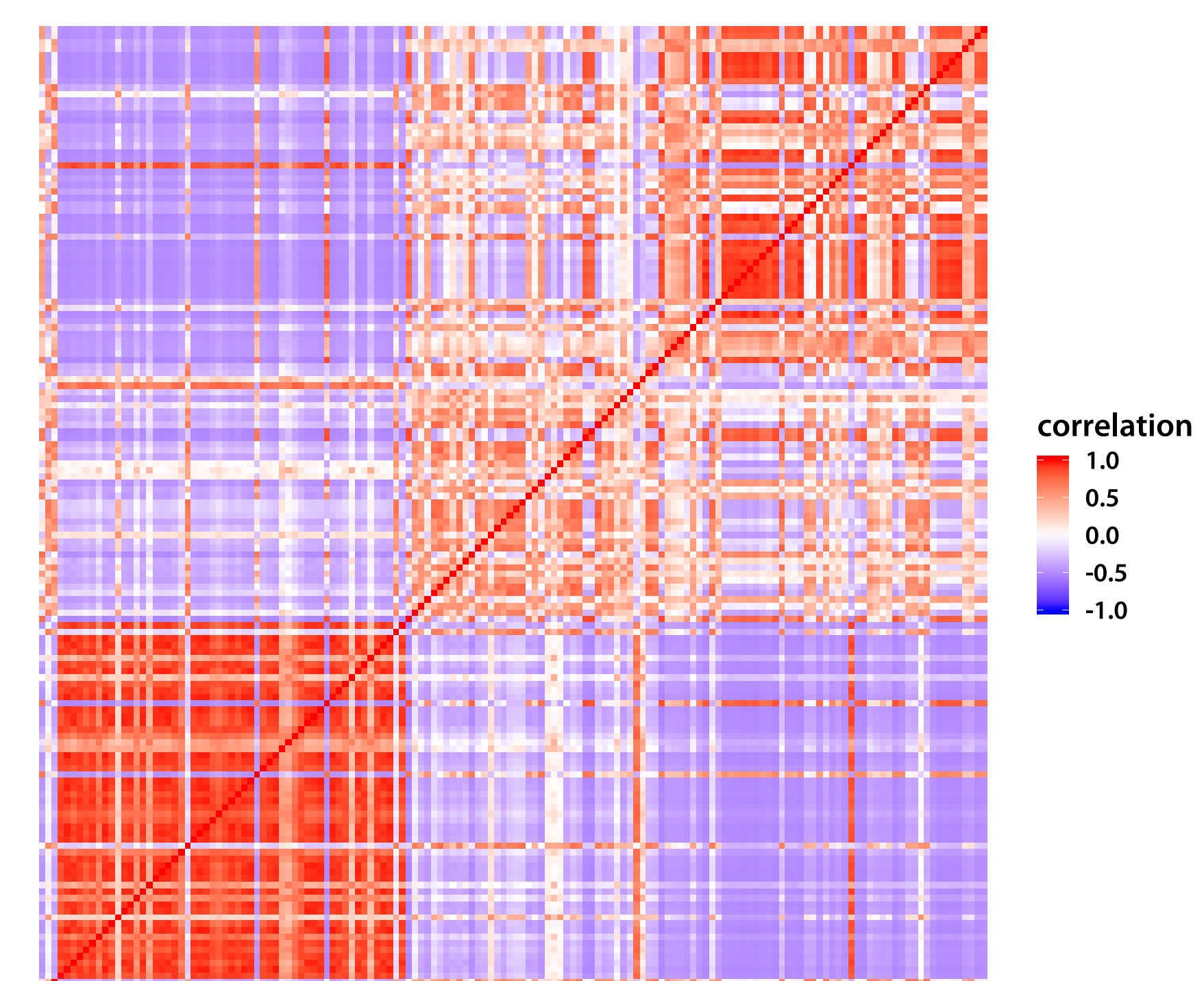
Challenge:
Population label switching

Population mean MCMC chains

chainID
— c1
— c2
— c3
— c4

Simulated data results
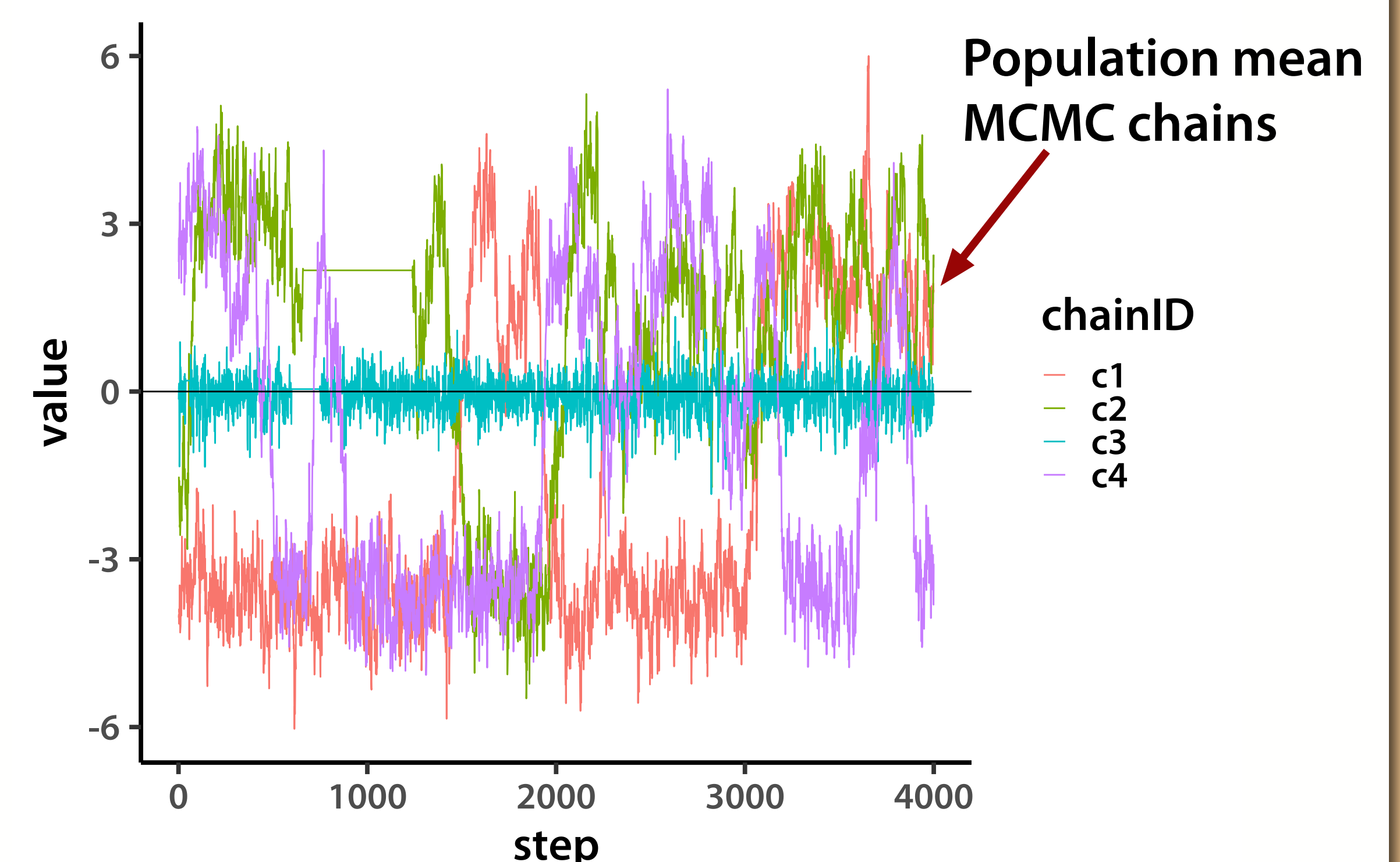
population
1
2
3
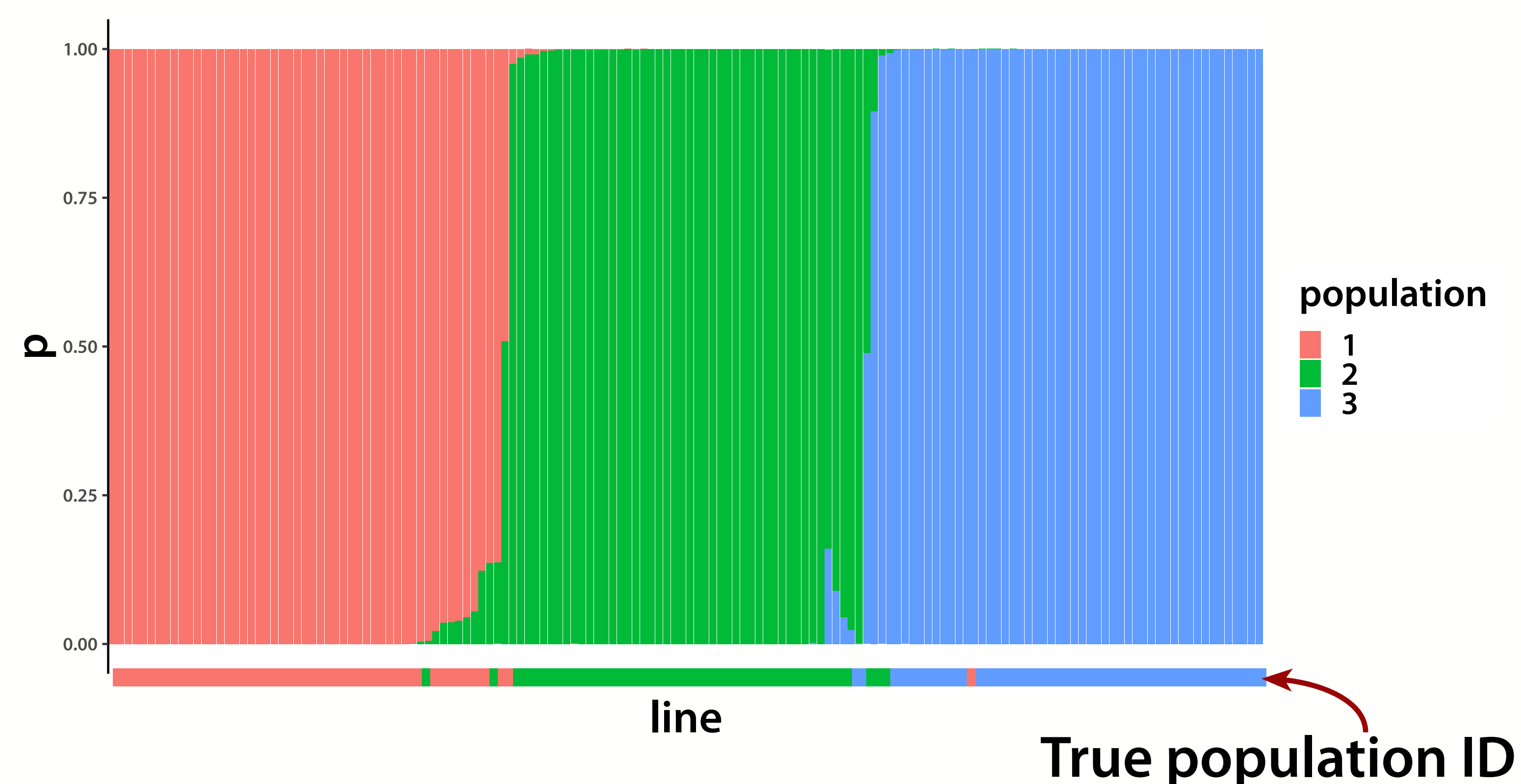
True population ID

Solutions

1. Sort populations using PC1(unstable)
2. Use concerted switching of population IDs in the Markov chain:

correlation
1.0
0.5
0.0
-0.5
-1.0

3. Left-ordered matrices (to be implemented)