Bayesian computation for high-dimensional Gaussian graphical models

Déborah Sulem (Università della Svizzera Italiana)

Università della Svizzera italiana

7th November 2024, AHIDI 2024 Workshop, Verona.

Acknowledgements

Jack Jewson
(Monash University)

David Rossell
(Pompeu Fabra University)

Outline

1. High-dimensional Gaussian graphical modelling

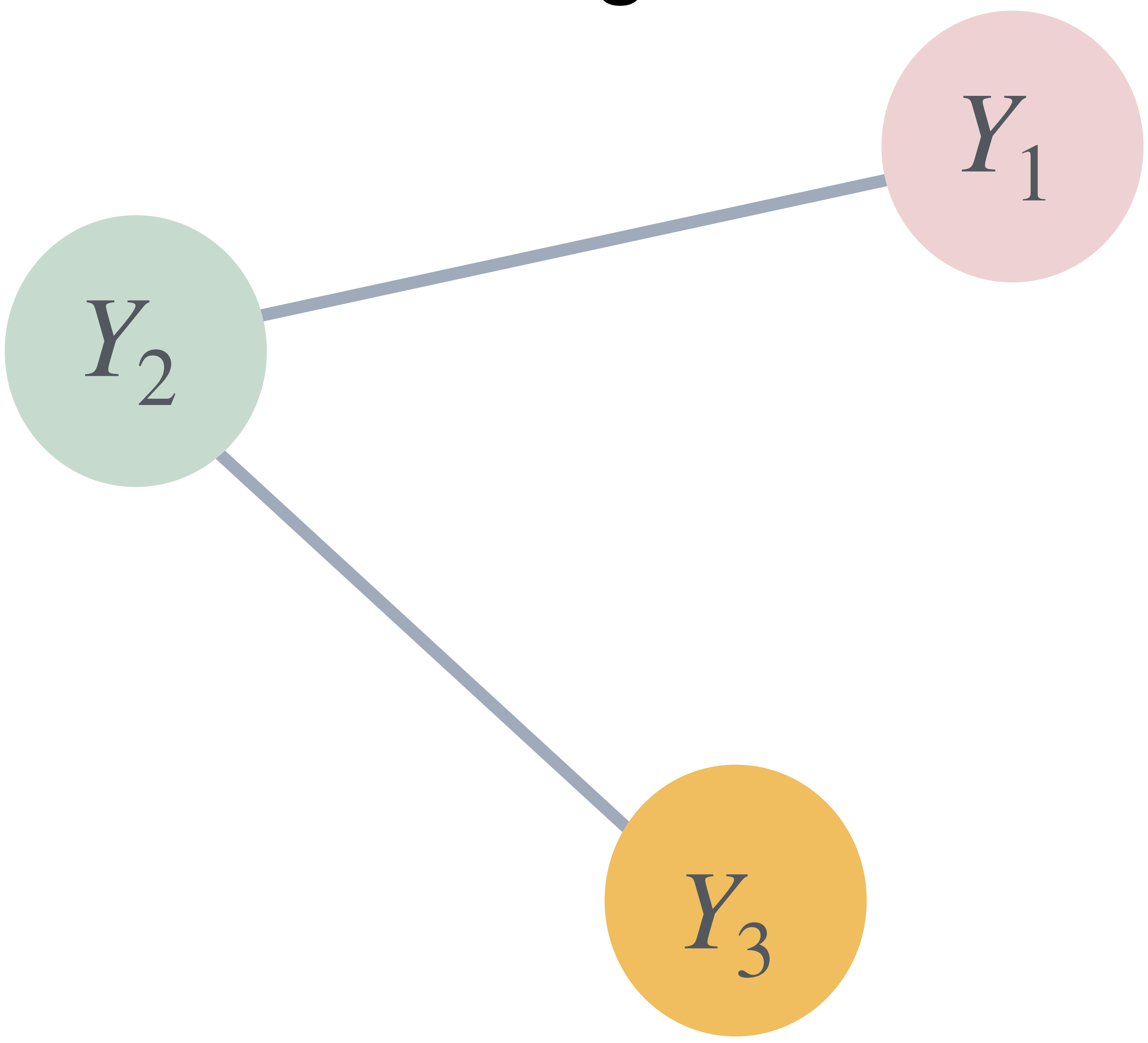
2. Sparse Bayesian inference

3. Local/Global Metropolis-within-Gibbs algorithms

4. Numerical results

Estimating partial dependencies

- Observations of a set of p variables $Y = (Y_1, \ldots, Y_p)$
- Statistical goal: estimate their partial dependencies $\mathscr{L}(Y_i, Y_j \mid \{Y_k, k \neq i, j\})$
- Graphical modelling



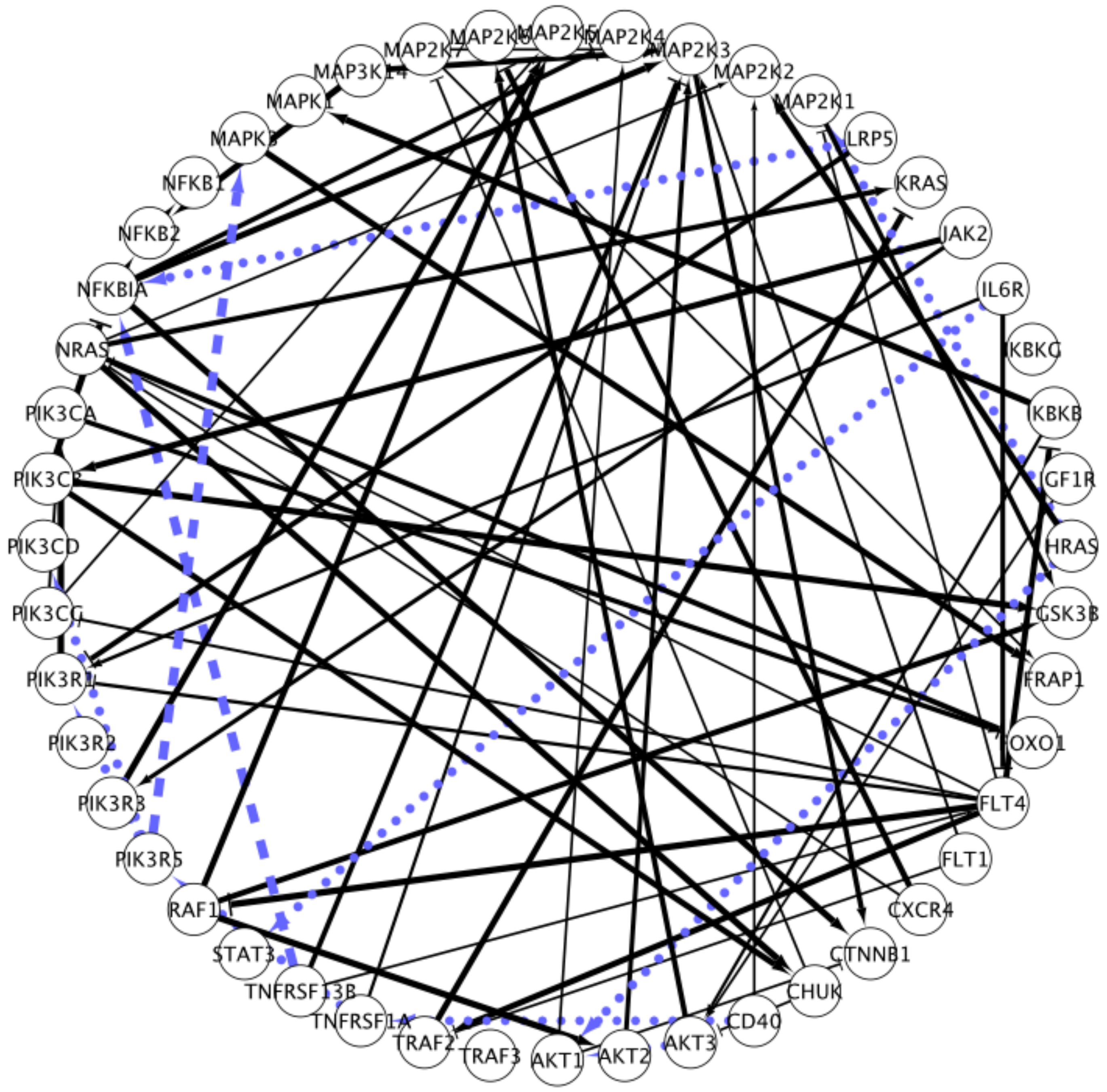
$$Y_1 - Y_2 \iff Y_1 \not\perp Y_2 \mid Y_3$$

$$Y_1 \not- Y_3 \iff Y_1 \perp Y_3 \mid Y_2$$

Applications with large p

- Gene-wise association studies

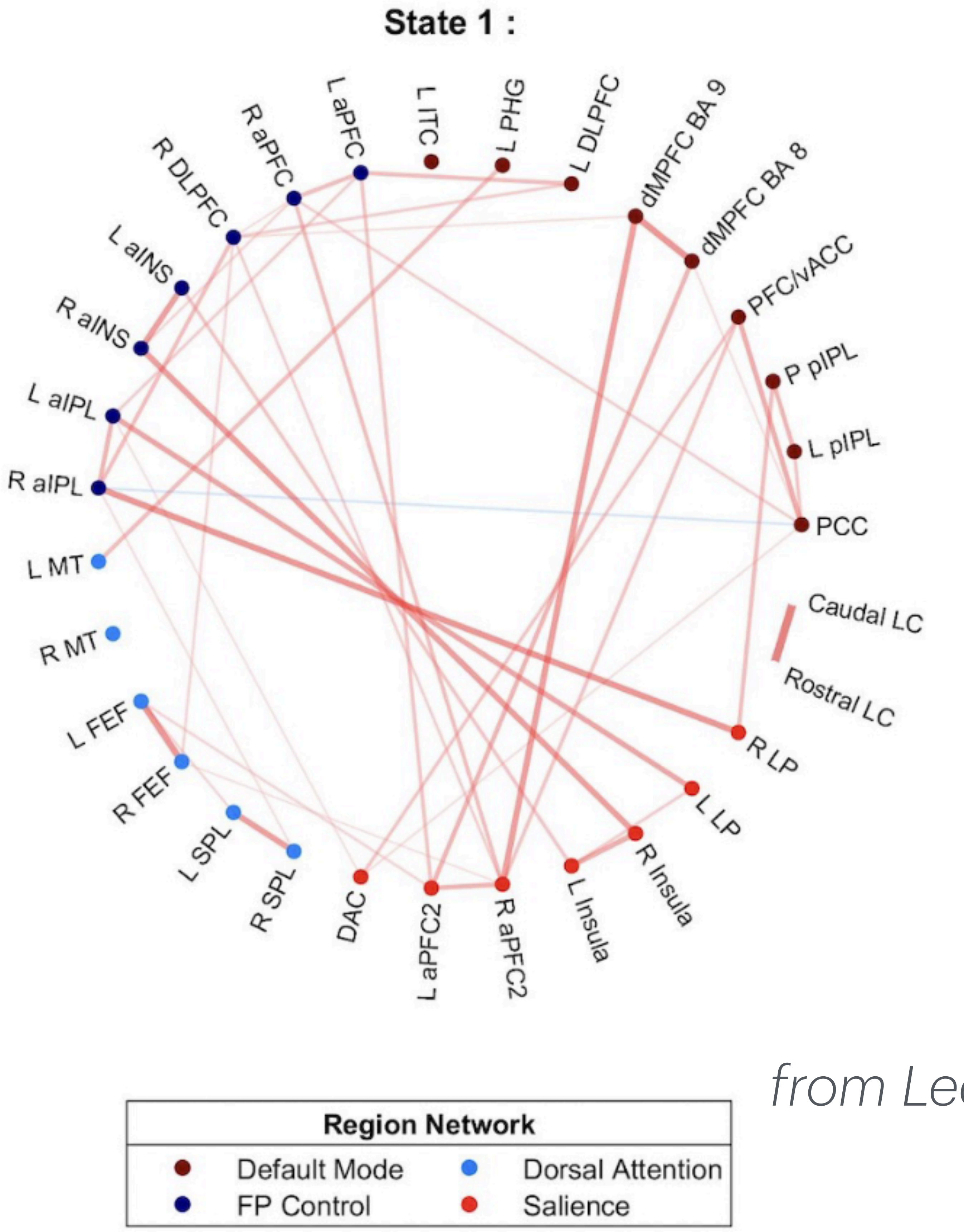
Y_j : level of expression of gene j



from Ni et al., 2022

- Neuroimaging data

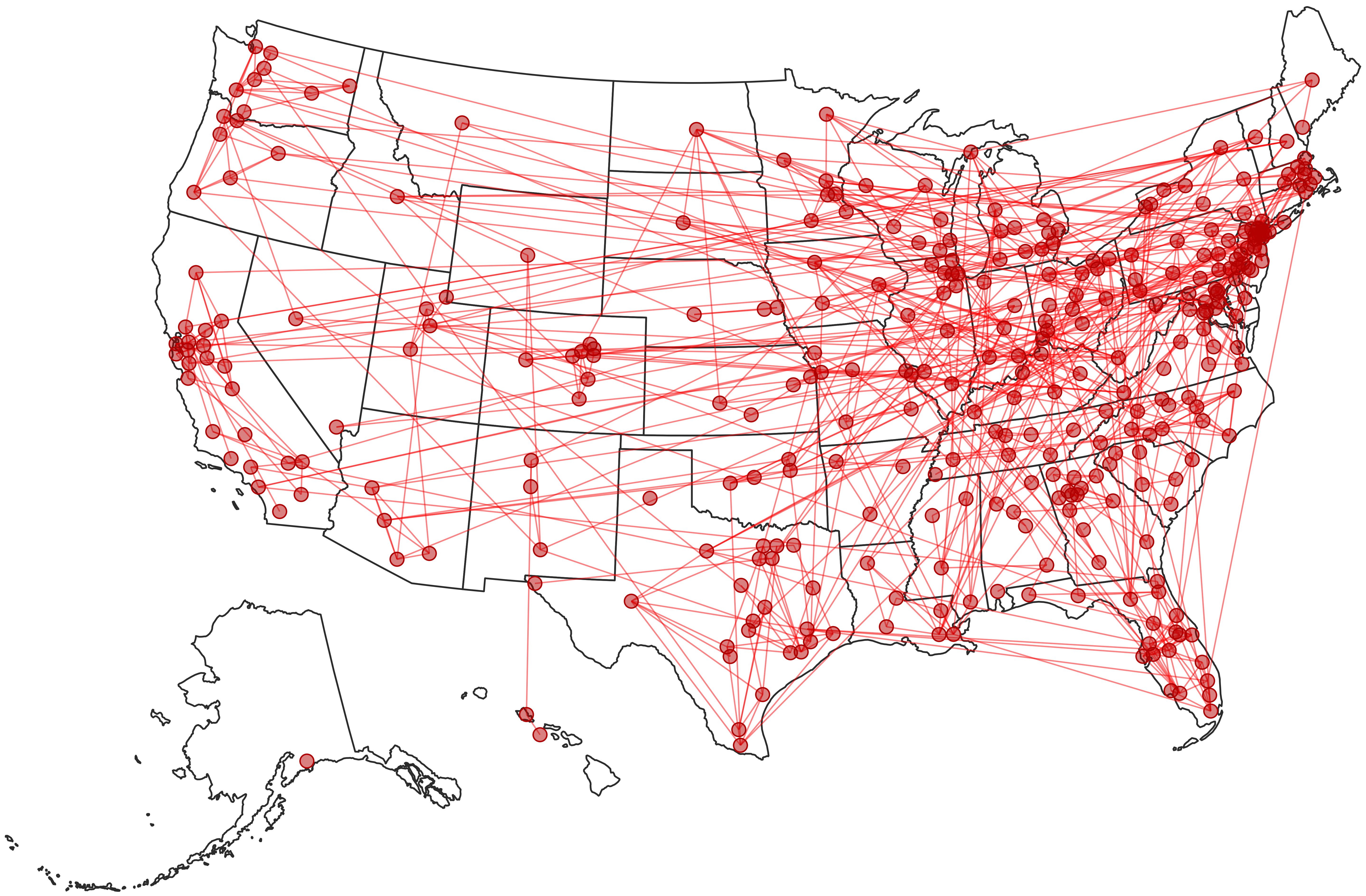
Y_j : level of activation of region of interest j



from Lee et al., 2023

COVID-19 log infection rates in the US

Weekly rates from Jan 2020 to Nov 2021 in 332 counties



Gaussian graphical modelling (GGM)

- Assume $Y = (Y_1, \ldots, Y_p) \sim \textbf{MVN}(\mathbf{0}_p, \Sigma)$

  with $\Sigma = (\sigma_{ij})_{1 \leq i,j \leq p}$ the covariance matrix $(p \times p)$

- With $\Omega := \Sigma^{-1} = (\Omega_{ij})_{1 \leq i,j \leq p}$ the precision (inverse-covariance) matrix, for $i \neq j$,

$$\Omega_{ij} = 0 \iff Y_i \perp Y_j \,|\, \{Y_k, k \neq i,j\}$$

➡ The precision matrix encodes the graphical model!

- <u>Statistical goal:</u> from n i.i.d observations of Y, $y_1, \ldots, y_n \in \mathbb{R}^p$, estimate Ω and the corresponding graphical model $Z = (z_{ij})_{i,j}$ with $z_{ij} = 1$if $\Omega_{ij} \neq 0$ and $z_{ij} = 0$ otherwise.

Estimation in high-dimensional GGM

- If p is large, it is reasonable to look for a sparse graphical model.
- Penalised Maximum Likelihood Estimator [Friedman et al., 2008; Yuan and Lin, 2007]

Graphical LASSO

$$\hat{\Omega} = \arg\max_{\Omega \succ 0} \underbrace{\log|\Omega| - tr(\Omega\hat{\Sigma})}_{\text{log-likelihood}} - \underbrace{\lambda|\Omega|_1}_{\text{penalisation}}$$

with $\hat{\Sigma} = \frac{1}{n}\sum_i y_i y_i^T = \frac{1}{n}\mathbf{Y}^T\mathbf{Y}$ the sample covariance matrix.

Neighbourhood selection via Lasso [Meinshausen and Buhlmann, 2006]

$$\hat{\theta}^j = \arg\min_{\theta \in \mathbb{R}^{p-1}} \|\mathbf{Y}_{.j} - \mathbf{Y}_{.-j}\theta\|^2 + \lambda\|\theta\|_1, \qquad j = 1,\ldots,p\,.$$

  - Set $z_{ij} = 1$ if $\hat{\theta}^j_i \neq 0$ and/or $\hat{\theta}^i_j \neq 0$

- (G)LASSO estimators tend to shrink large coefficients.
- No uncertainty quantification on the graphical model.

Sparse Bayesian inference

Sparse conjugate prior

- G-Wishart distribution:

$$\Pi(\Omega) = \Pi(\Omega \,|\, Z)\Pi(Z) \quad \text{with e.g., } \Pi(Z) = \text{Ber}(\theta)^{\frac{p(p-1)}{2}} \text{ with } \theta \in (0,1)$$

and $$\Pi(\Omega \mid Z) = I_Z(b, D)\,|\,\Omega\,|^{\frac{b-2}{2}} \exp\{\frac{1}{2} Tr(\Omega D)\}$$

with $b > 2,\ D \succ 0$ and $I_Z(b, D)$ is an intractable normalising constant.

➡Requires some computational engineering and/or approximation to compute the posterior distribution via MCMC [Mohammadi et al. 2021, 2023].

➡Because of this, each MCMC step only modifies one edge.

Spike-and-slab prior for GGM

- Shrinks each coefficient via a "product" of univariate mixture [Wang, 2015]

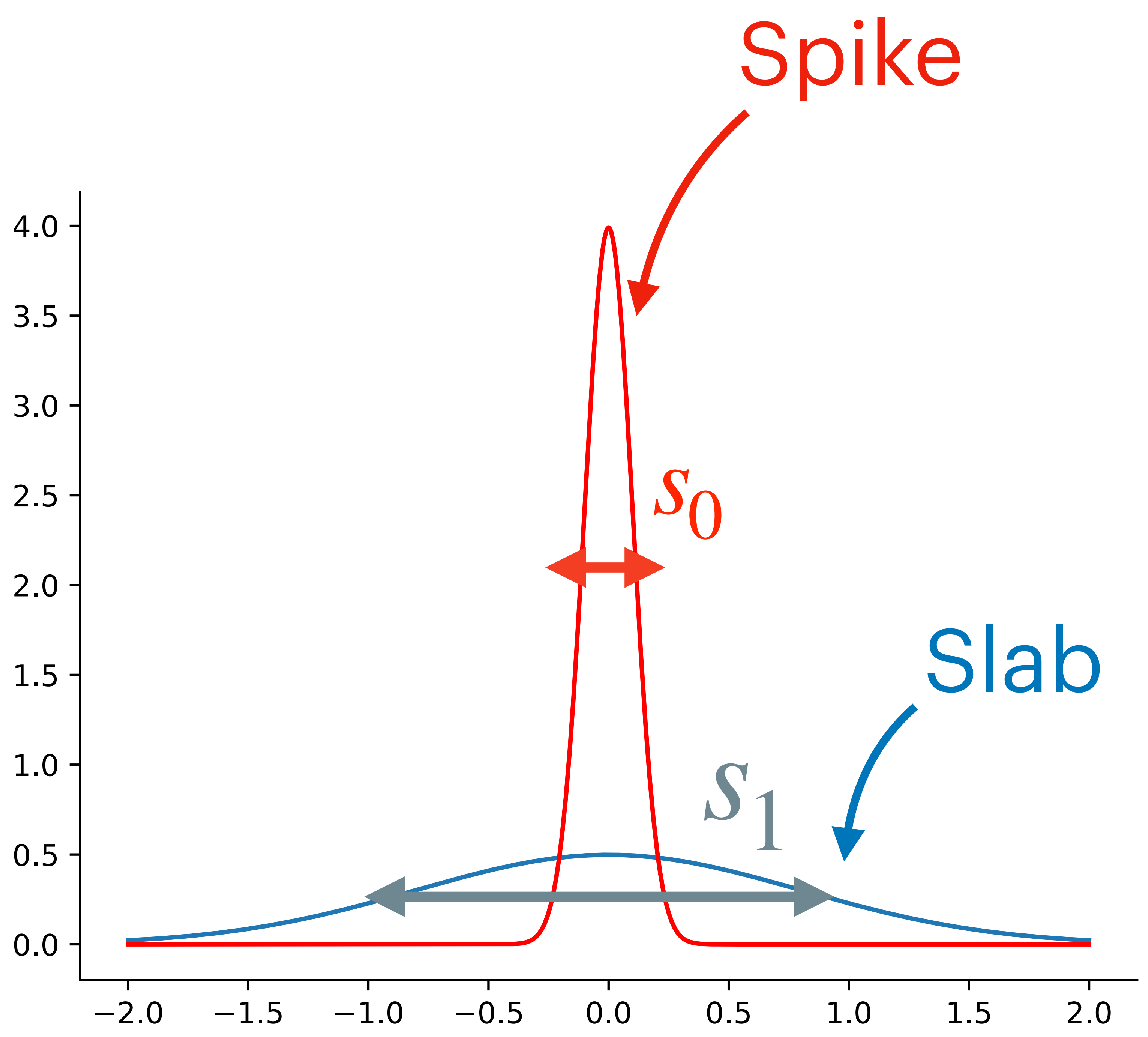
$$\Pi_{SAS}(\Omega) \propto \prod_i \mathrm{Exp}(\Omega_{ii}; \lambda) \prod_{i<j} \pi_{SAS}(\Omega_{ij}) \mathbf{1}_{\Omega \succ 0}$$

where $$\boxed{\pi_{SAS}(\Omega_{ij}) = (1-\theta)\ \underbrace{\mathrm{N}(\omega_{ij};\ 0, s_0^2)}_{\text{spike}} + \theta\ \underbrace{\mathrm{N}(\omega_{ij};\ 0, s_1^2)}_{\text{slab}}}$$

  - $\theta \in (0,1)$: slab's weight
  - $s_0, s_1 > 0$: spike's and slab's standard deviations

➡The spike models small (non-significant) coefficients while the slab allows large coefficients.

➡The precision matrix is not sparse!



Discrete spike-and-slab

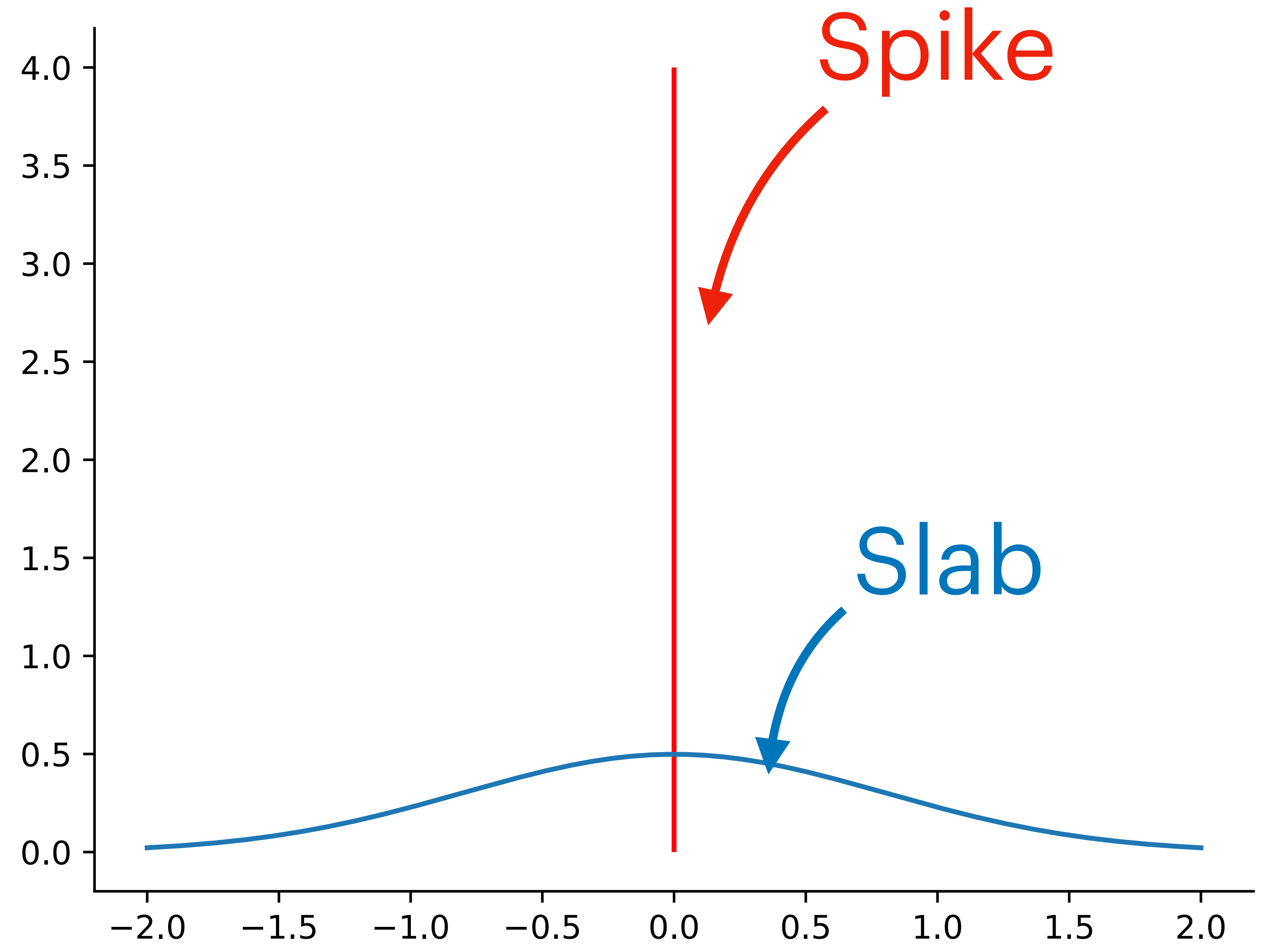
- We replace the spike's normal density by a Dirac measure at 0:

$$\Pi_{DSAS}(\Omega) \propto \prod_i \mathrm{Exp}(\Omega_{ii}; \lambda) \prod_{i<j} \pi_{DSAS}(\Omega_{ij}) \mathbf{1}_{\Omega>0}$$

where $$\boxed{\pi_{DSAS}(\Omega_{ij}) = (1-\theta)\ \underbrace{\delta_0(\Omega_{ij})}_{\text{spike}} + \theta\ \underbrace{\mathrm{N}(\omega_{ij};\ 0, s_1^2)}_{\text{slab}}}$$

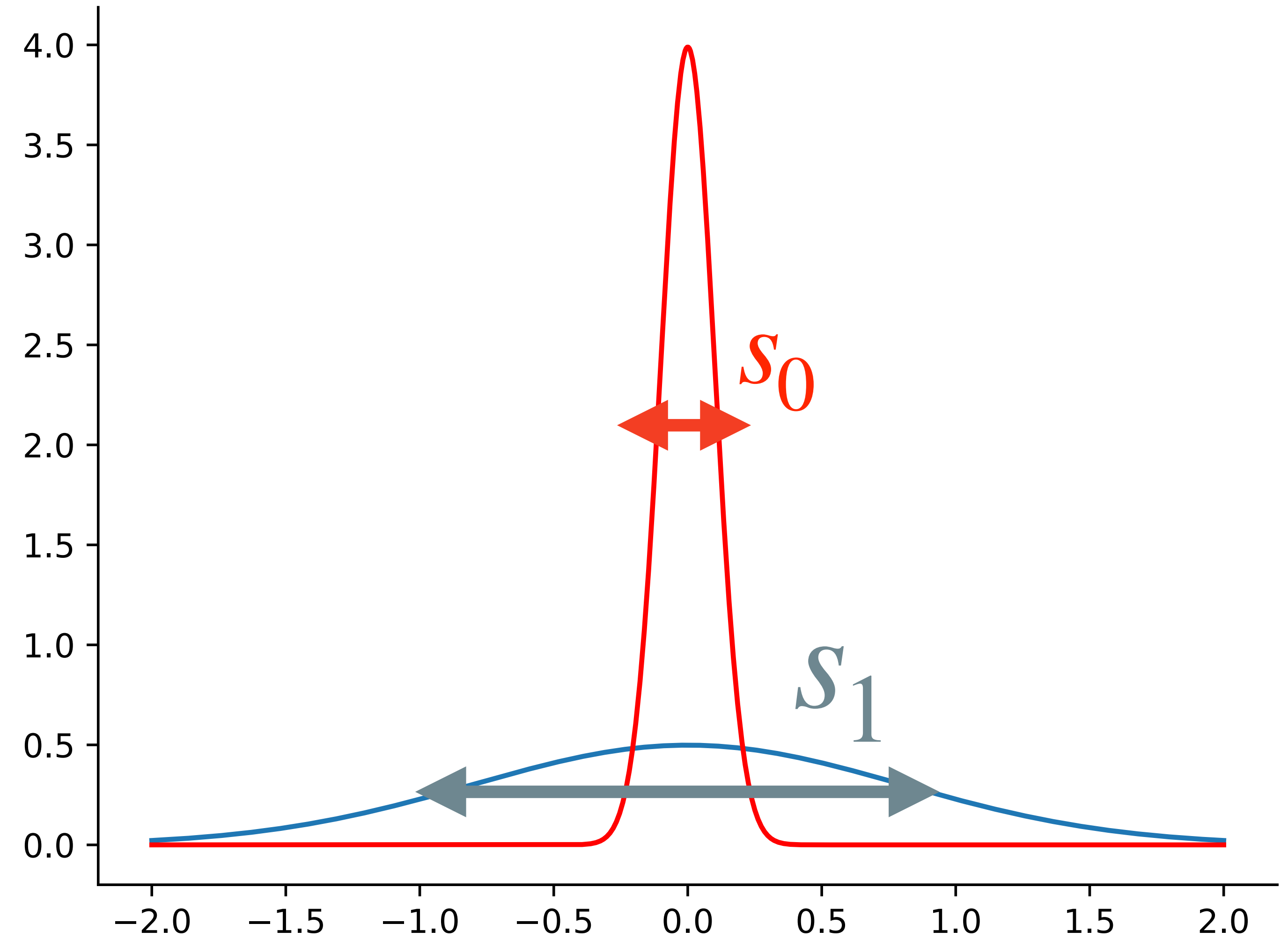
  - $\theta \in (0,1)$: slab's weight
  - $s_1 > 0$: slab's standard deviation

➡The spike now allows exact 0 while the slab allows large coefficients.

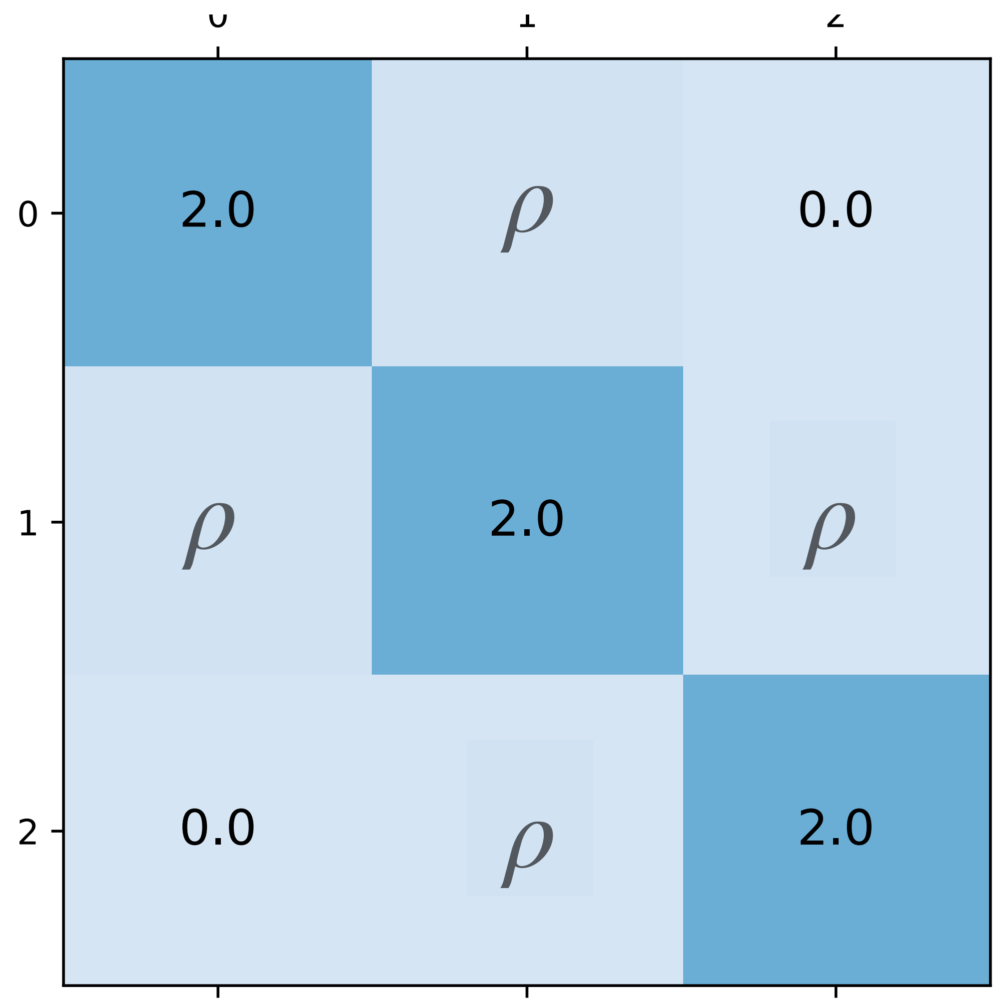


Continuous vs Discrete SAS

- Coefficients $\lesssim s_0$ are often estimated at 0



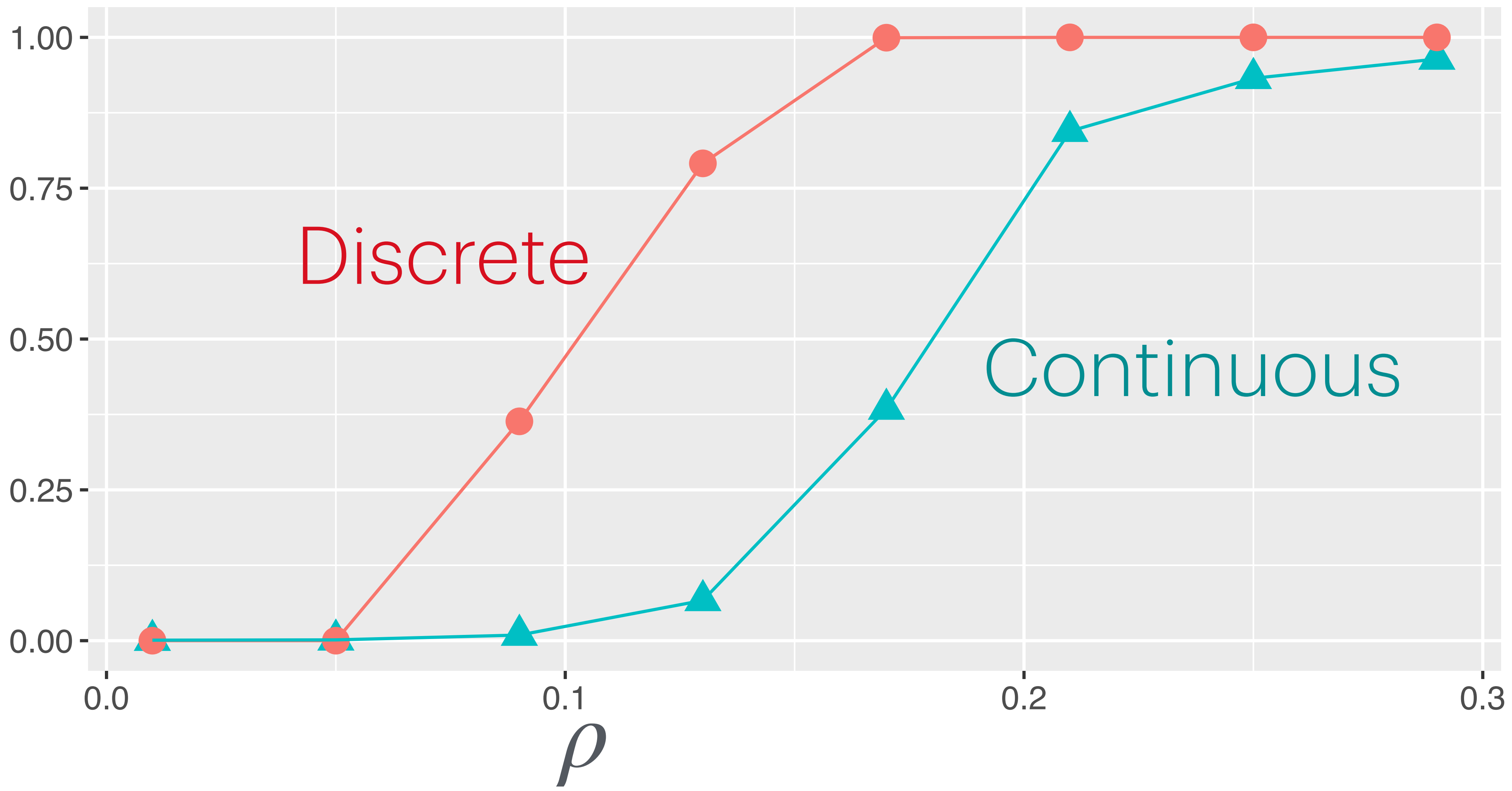
$$\Omega^0 = \begin{pmatrix} 2.0 & \rho & 0.0 \\ \rho & 2.0 & \rho \\ 0.0 & \rho & 2.0 \end{pmatrix}$$



$n = 5000,\ s_0 = 0.05$

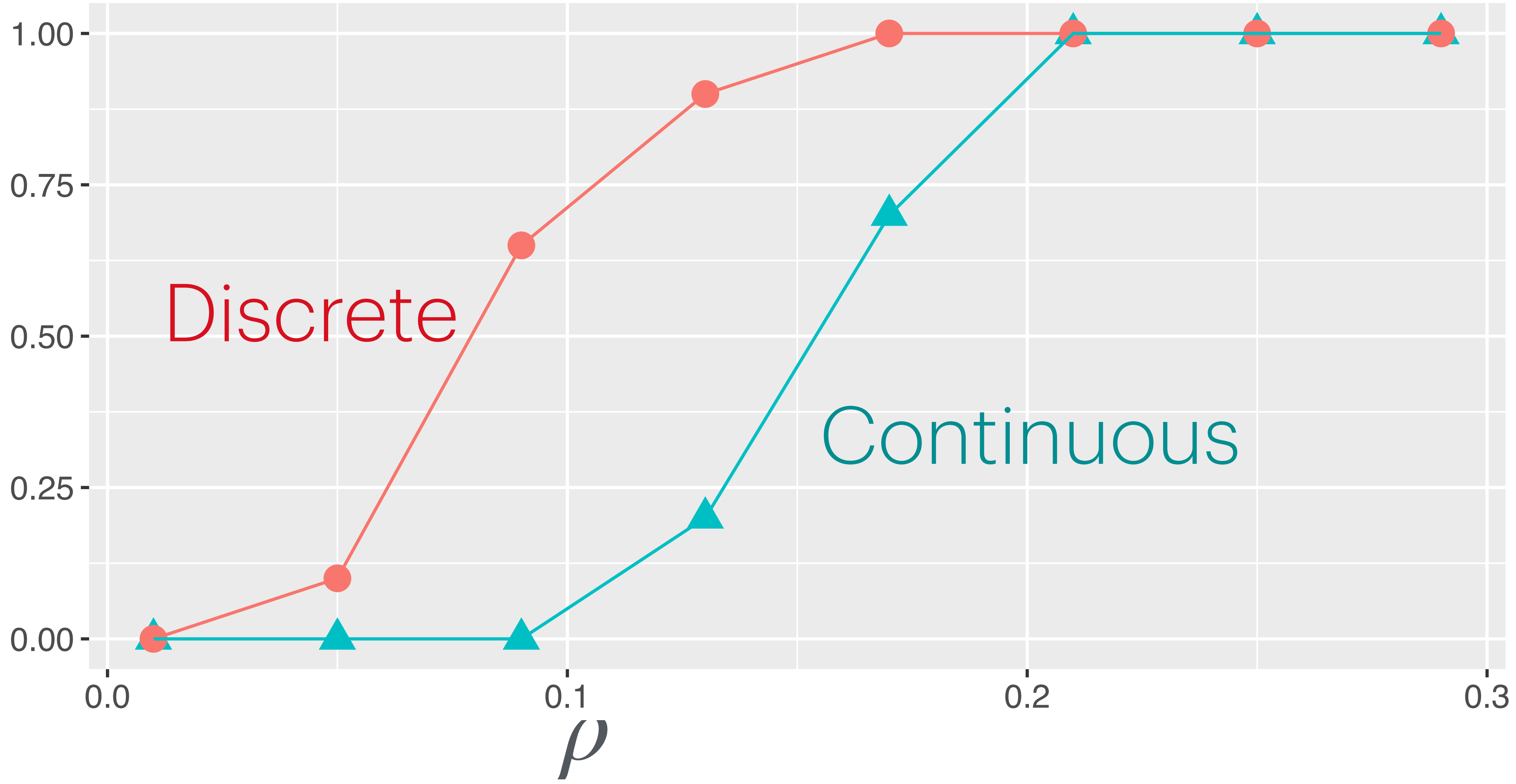
Posterior proba. on true model

$\Pi(Z^0 \mid \mathbf{Y})$



Power

$$\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

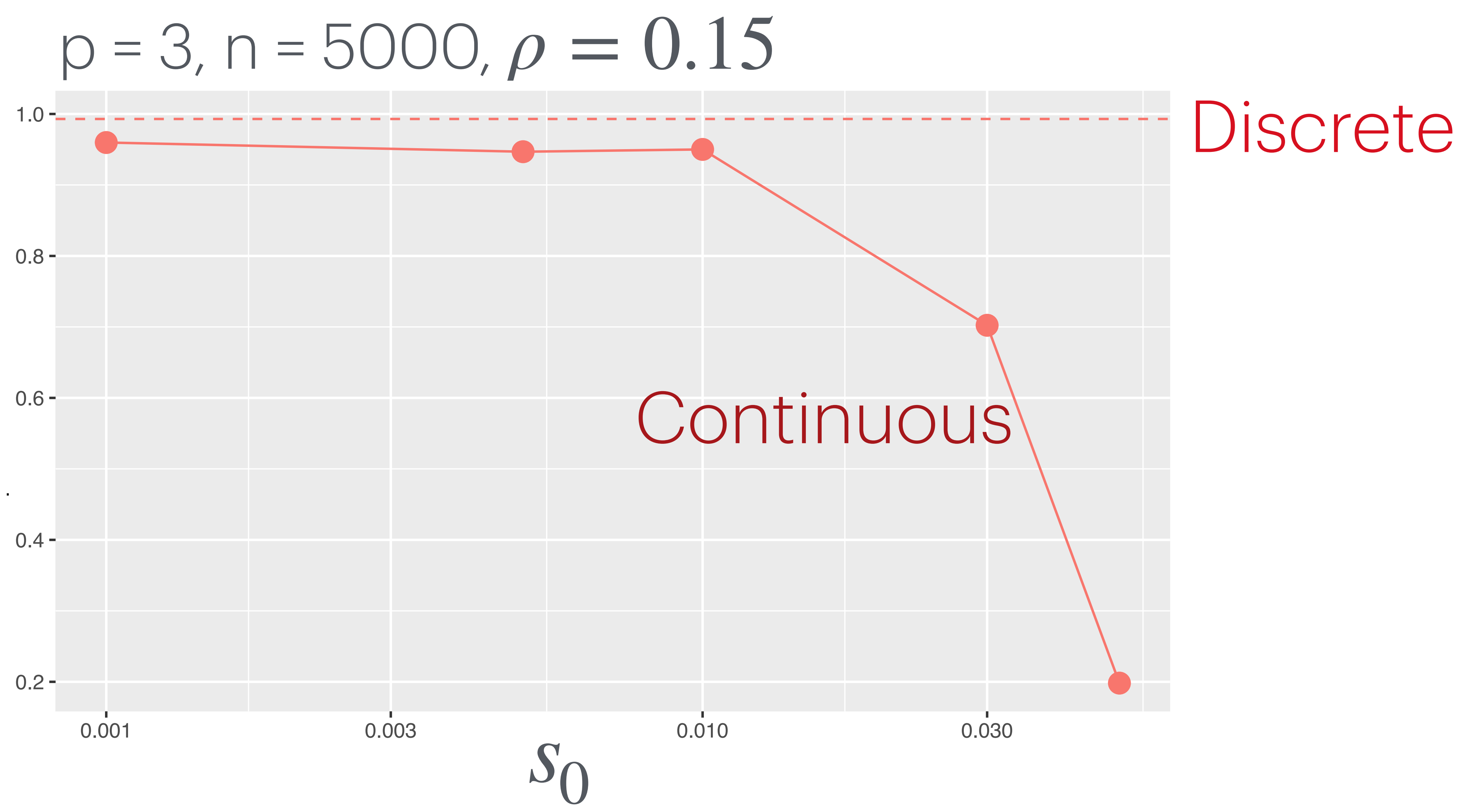


Continuous vs Discrete SAS

- Consistency if $s_0 \to 0$

Posterior proba. on true model

$\Pi(Z^0 \mid \mathbf{Y})$

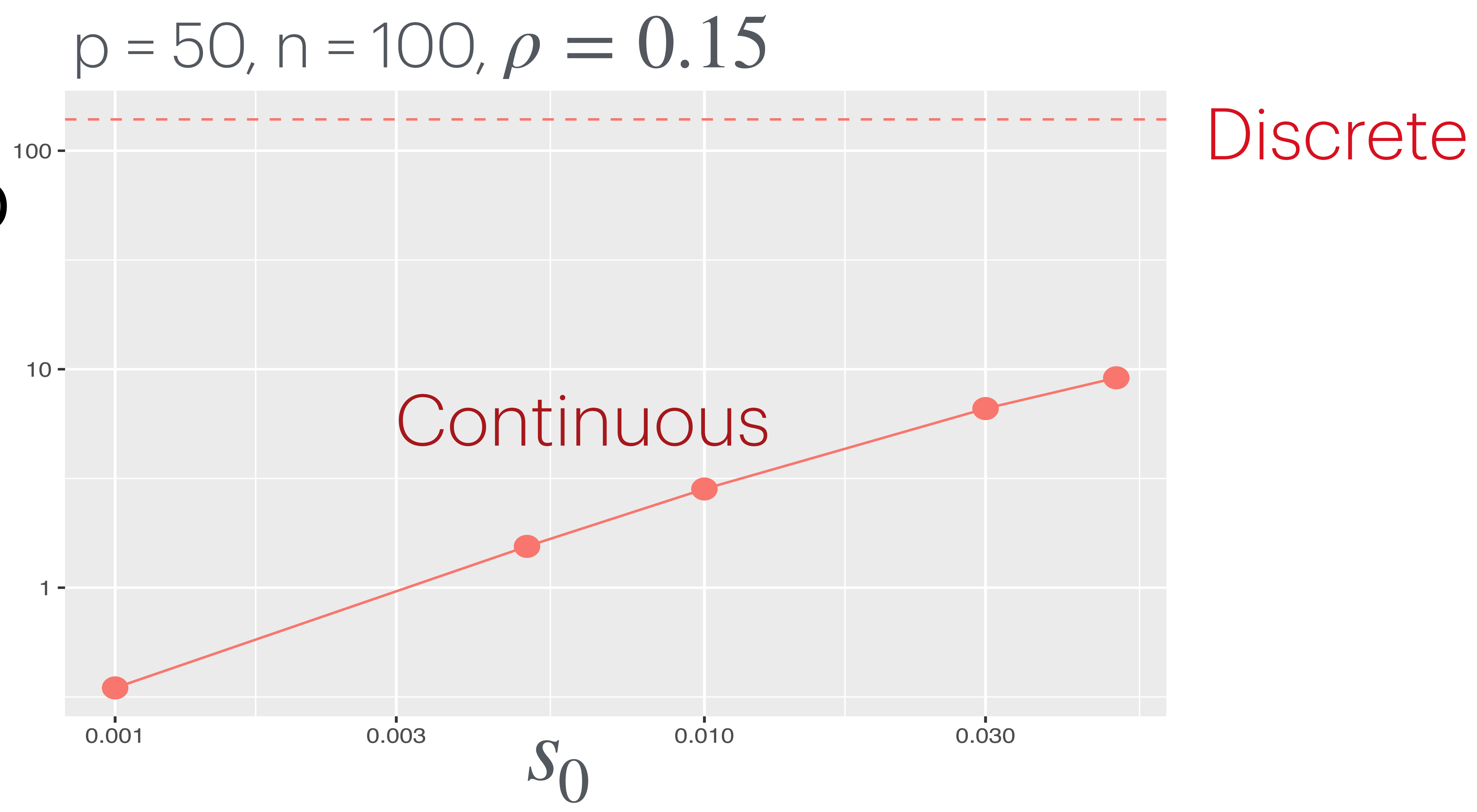


- Poor MCMC mixing if $s_0 < 0.01$

[George & McCulloch 1993, Wang 2015]

  - Expected jump distance (EJD) = average number of inclusion variable flips per iteration

EJD



➡ The continuous SAS is not scalable to p larger than 200

Our contributions

- Efficient Monte-Carlo Markov Chain (MCMC) algorithms targeting the posterior $\Pi(\Omega \mid \mathbf{Y})$ for the "discrete" spike-and-slab prior.
- We propose 2 types of Markov steps: local or globally-informed moves.
- We empirically show that it is computationally faster than state-of-the-art (fully) Bayesian algorithms for GGM.
- We analyse the mixing times of our MCMC, prove that it can be "dimension-free" under some sparsity conditions (in preprint soon)

Local/Global Metropolis-within-Gibbs

In a nutshell

- The conditional posterior of one column of Ω given the graphical model can be made tractable under some re-parametrisation.

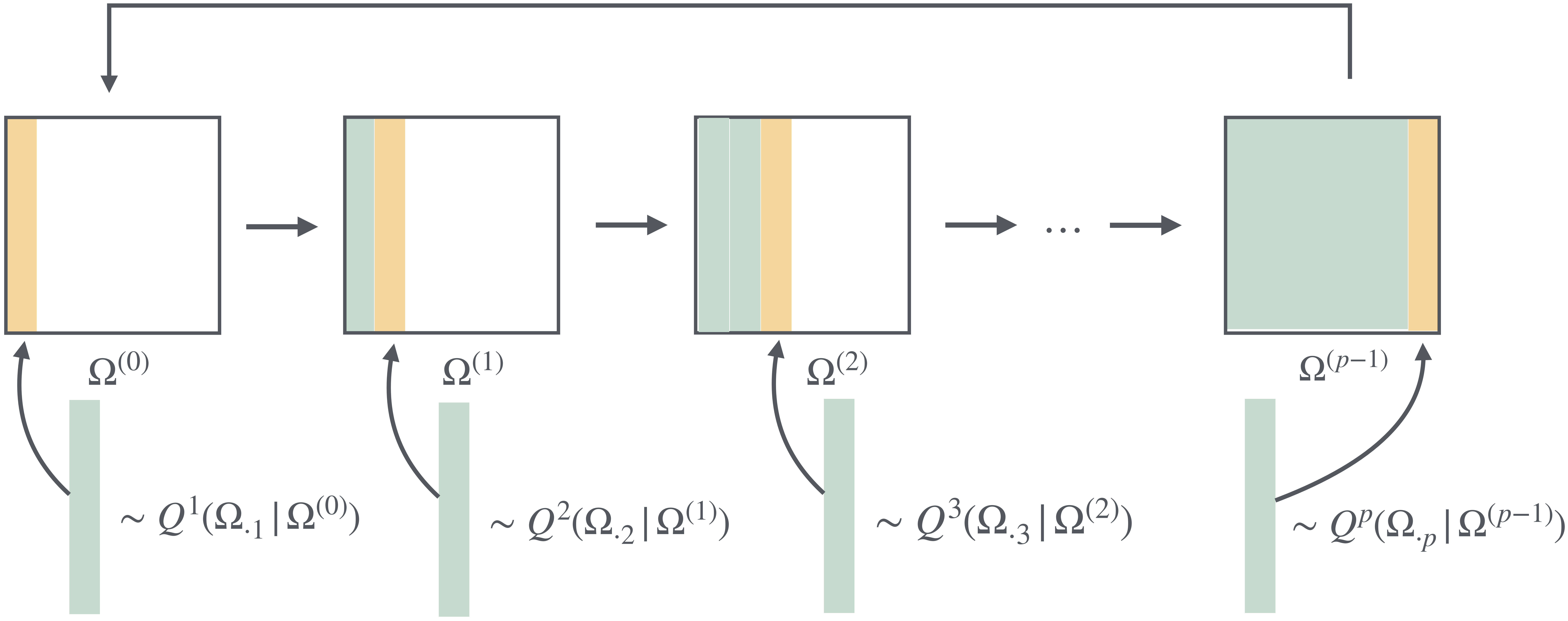
  ➡ This permits to design a block Metropolis-within-Gibbs sampler updating one column of Ω at a time.

- Sampling one column of the graphical model shares some similarity with variable selection in linear regression

  ➡ Local moves via Gibbs or Metropolis-Hastings schemes such as Birth-Death-Swap [Yang et al., 2016], Locally-Informed and Thresholded [Zhou et al., 2022] can be used.

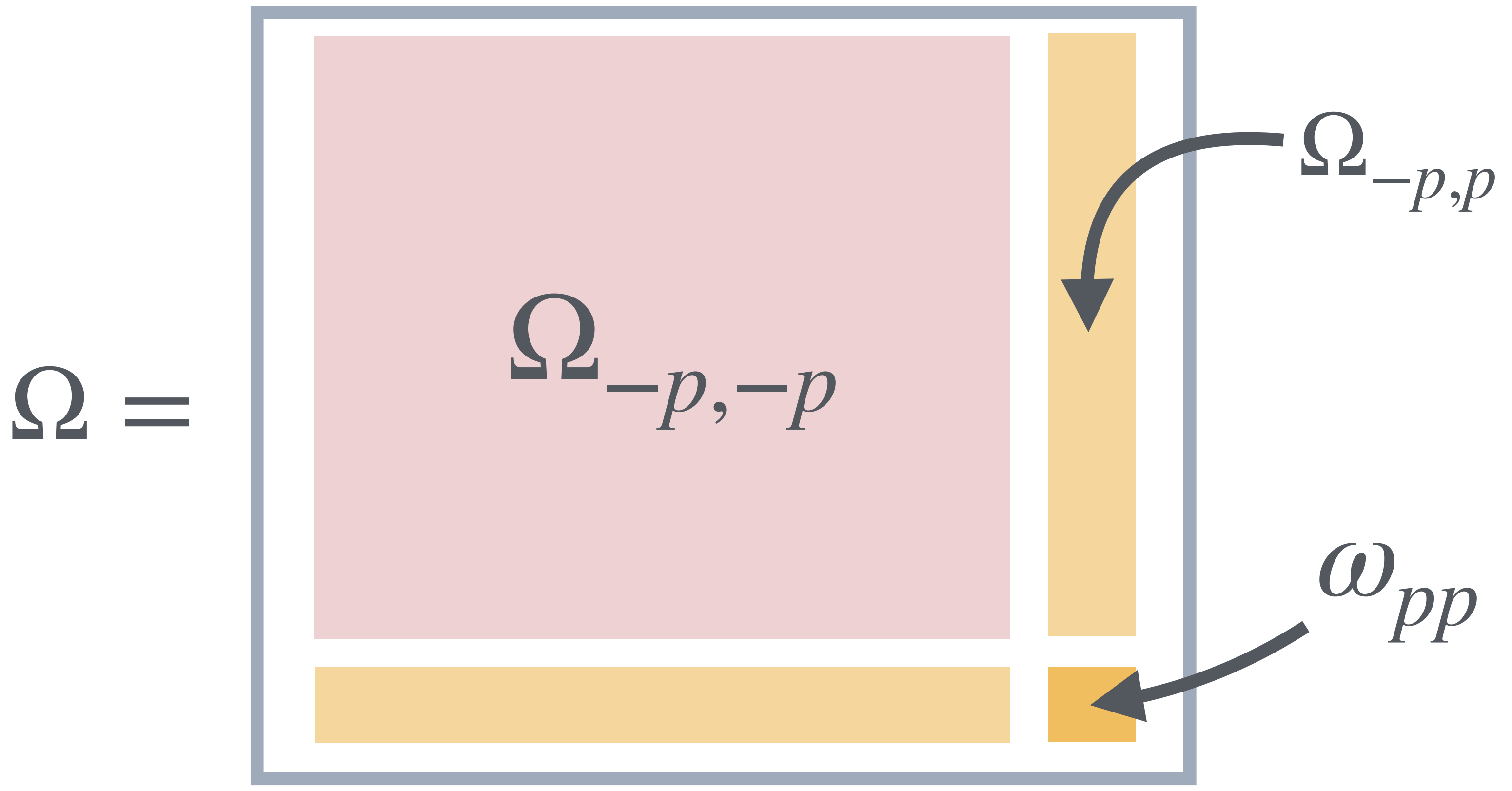
  ➡ Related linear regression models can be exploited to do global moves.

Metropolis-within-block Gibbs



Conditional posterior distribution

- Re-parametrise $\Pi(\Omega_{\cdot j} \mid \Omega_{-j-j}, \mathbf{Y})$ to a "tractable" form.
    - Consider $j = p$
    - $z \in \{0,1\}^{p-1}$ with $z_k = 0 \iff \Omega_{kp} = 0$
    - $u = -\Omega_{-p,p,z}$
    - $v_p = \omega_{pp} - u^T \Omega^{-1}_{-p-p,z} u$



- **Proposition:**

  Under the DSAS prior and the re-parametrisation we have

$$\Pi(u, v, z \mid \Omega_{-p,-p}, \mathbf{Y}) = \mathrm{N}(u; \cdot, \cdot)\mathrm{Ga}(v, \cdot, \cdot)\Pi(z \mid \Omega_{-p,-p}, \mathbf{Y})$$

with $\boxed{\Pi(z \mid \Omega_{-p-p}, \mathbf{Y}) \propto \frac{e^{\frac{m_z^T U_z m_z}{2}}}{s_1^{|z|_0} |U_z|^{\frac{1}{2}}} \theta^{|z|_0} (1-\theta)^{p-1-|z|_0}.}$ ➡ approximated via Metropolis-Hastings

Metropolis-Hastings (MH) 1: local moves

- $\Pi(z \mid \Omega_{-p-p}, \mathbf{Y})$ is intractable but not dissimilar to other variable selection problems
- One can use (local) SOTA Metropolis-Hasting Markov kernel to approximate it

➡ Gibbs [George and McCulloch, 1993]: $z_k \,|\, z_{-k}, \Omega_{-p-p}, \mathbf{Y} \sim \text{Ber}(\,\cdot\,)$

➡ Birth-Death-Swap [Yang et al., 2016]: $Q(z^* \mid z)$ where z^* is obtained from z by either:

  * adding a 1
  * removing a 1
  * swapping a 1 with a 0

➡ Locally-Informed and Thresholded proposal [Yang et al., 2016]: same as Birth-Death-Swap but each move is weighted by a function of the posterior and the ratios of weights are bounded.

MH 2: globally-informed moves

- The "model" z for $\Omega_{\cdot p}$ is also the "model" of β in the linear regression problem:

$$Y_p = \beta^T Y_{-p} + \epsilon$$

Proof: with $Y = (Y_1, \ldots, Y_p) \sim \textbf{MVN}(\mathbf{0}_p, \Omega^{-1})$ and denoting $Y_{-p} = \{Y_k, k \neq p\}$, then

$$p(Y_p \mid Y_{-p}) = \mathrm{N}\left(\frac{-\Omega^T_{-p,p}}{\omega_{pp}} Y_{-p}, \omega_{pp}^{-1}\right) \quad \text{or, equivalently,} \quad Y_p = \frac{-\Omega^T_{-p,p}}{\omega_{pp}} Y_{-p} + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \omega_{pp}^{-1}).$$

- A DSAS prior in the linear regression $Y_p = \beta^T Y_{-p} + \epsilon$ and with $z_k = 0 \iff \beta_k = 0$ leads to the posterior

$$\tilde{\Pi}(z \mid \mathbf{Y}_p, \mathbf{Y}_{-p}) \propto \frac{\theta^{|z|_0}(1-\theta)^{p-1-|z|_0}}{s_1^{-|z|_0} |W_z|^{\frac{1}{2}}} \left(\frac{1}{b + s_{pp} - \mu_z^T W_z \mu_z}\right)^{\frac{n}{2}+1}$$

➡ Many efficient samplers targeting this distribution

➡ $\tilde{\Pi}(z \mid \mathbf{Y}_p, \mathbf{Y}_{-p})$ is independent of Ω_{-p-p} while still being informed (globally) by the prior and data

MH 2: globally-informed moves + Tempering

- The proposal "linear regression" posterior

$$\tilde{\Pi}(z \,|\, \mathbf{Y}_p, \mathbf{Y}_{-p}) \propto \frac{\theta^{|z|_0}(1-\theta)^{p-1-|z|_0}}{s_1^{-|z|_0} \,|\, W_z \,|^{\frac{1}{2}}} \left(\frac{1}{b + s_{pp} - \mu_z^T W_z \mu_z} \right)^{\frac{n}{2}+1}$$

can be more concentrated than the target $\Pi(z \,|\, \Omega_{-p-p} \,.\, \mathbf{Y})$ and cause mixing issues.

➡ Tempering reduces over-concentration and improves mixing:

$$Q_\beta(z) \propto \tilde{\Pi}(z \mid \mathbf{Y}_p, \mathbf{Y}_{-p})^\beta$$

with $\beta \in (0,1]$.

Serial MCMC with local moves

<u>Algorithm:</u>

- <u>Input:</u> $\mathbf{Y}, T, \Omega^{(0)}$
- for each $t = 1,2,\ldots,T$:
  - for each $j = 1,2,\ldots,p$:
    - Let $\Omega^{(t,j-1)} = (\Omega_{-j-j}, \Omega_{\cdot j})$, $z_j = \mathbf{1}(\Omega_{\cdot j} \neq 0)$
    - MH step: propose $z_j^* \sim Q(z \,|\, z_j)$ (Gibbs/BDS/LIT) and accept with probability α
    - $u_j \,|\, z_j^* \sim$ MVN($\cdot$, $\cdot$)
    - $v_j \sim$ Ga($\cdot$, $\cdot$) and set $\omega_{jj}^{(t)} = v_j + u_j^T \Omega_{-j-j,z_j^*}^{-1} u_j$
    - Update $\Omega^{(t,j)} = (\Omega_{-j-j}, \Omega_{\cdot j}^{(t)} = (-u_j, \omega_{jj}^{(t)}))$
- <u>Output:</u> samples $\{\Omega^{(t)}\}_{t \leq T}$

Almost-parallel MCMC with global moves

Algorithm:

- Input: $\mathbf{Y}, T, \Omega^{(0)}, \beta$
- for each $j = 1,2,\ldots,p$ in parallel, for each $t = 1,2,\ldots,T$,
$$\tilde{z}_j^{(t)} \sim Q_\beta^j(z)$$
p independent MCMC chains
- for each $t = 1,2,\ldots,T$:
  - for each $j = 1,2,\ldots,p$:
    - let $\Omega^{(t,j-1)} = (\Omega_{-j-j}, \Omega_{\cdot j})$, $z_j = \mathbf{1}(\Omega_{\cdot j} \neq 0)$
    - MH step: propose $z_j^* = \tilde{z}_j^{(t)}$ (global) and accept with probability α
    - $u_j \,|\, z_j^* \sim$ MVN$(\,\cdot\,,\,\cdot\,)$
    - $v_j \sim$ Ga$(\,\cdot\,,\,\cdot\,)$ and set $\omega_{jj}^{(t)} = v_j + u_j^T \Omega_{-j-j,z_j^*}^{-1} u_j$
    - Update $\Omega^{(t,j)} = (\Omega_{-j-j}, \Omega_{\cdot j}^{(t)} = (-u_j, \omega_{jj}^{(t)}))$
- Output: samples $\{\Omega^{(t)}\}_{t \leq T}$

Numerical results

Comparison to state-of-the-art

- Gibbs sampler for continuous SAS [Wang , 2015]: SSGraph
- MCMC for G-Wishart prior:
  - BDGraph [Mohammadi et al., 2021]: continuous-time Birth-Death
  - BDGraph.MPL[1] [Mohammadi et al., 2023]: use pseudo-likelihood within BD steps
- Approximate Bayesian method:
  - Quasi-posterior [Atchade, 2021]: $Q(\Omega \mid \mathbf{Y}) \propto \prod_j Q(\Omega_{.j} \mid \mathbf{Y})$ with $Q(\Omega_{.j} \mid \mathbf{Y})$ is the posterior obtained by considering regressing Y_j onto Y_{-j}

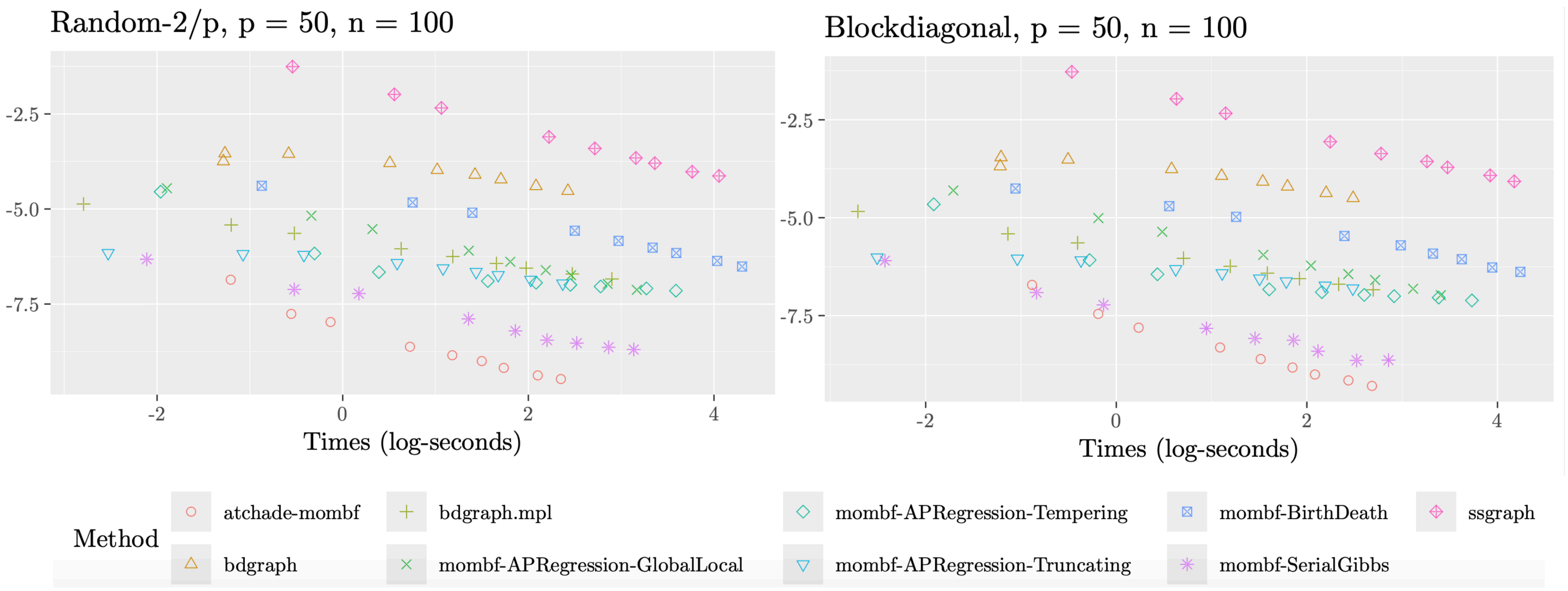
    ➡ Can also compute $Q(\Omega_{.j} \mid \mathbf{Y})$ in parallel for each j

    ➡ Does not guarantee symmetric positive definite Ω
- Our methods: implemented in mombf[2]
  - Serial with Gibbs or Birth-Death-Swap: Metropolis-wtihin-Gibbs with local moves
  - Almost Parallel-β: MwG with global moves and tempering (and other variants)

[1]https://cran.r-project.org/web/packages/BDgraph/index.html
[2]https://github.com/davidrusi/mombf

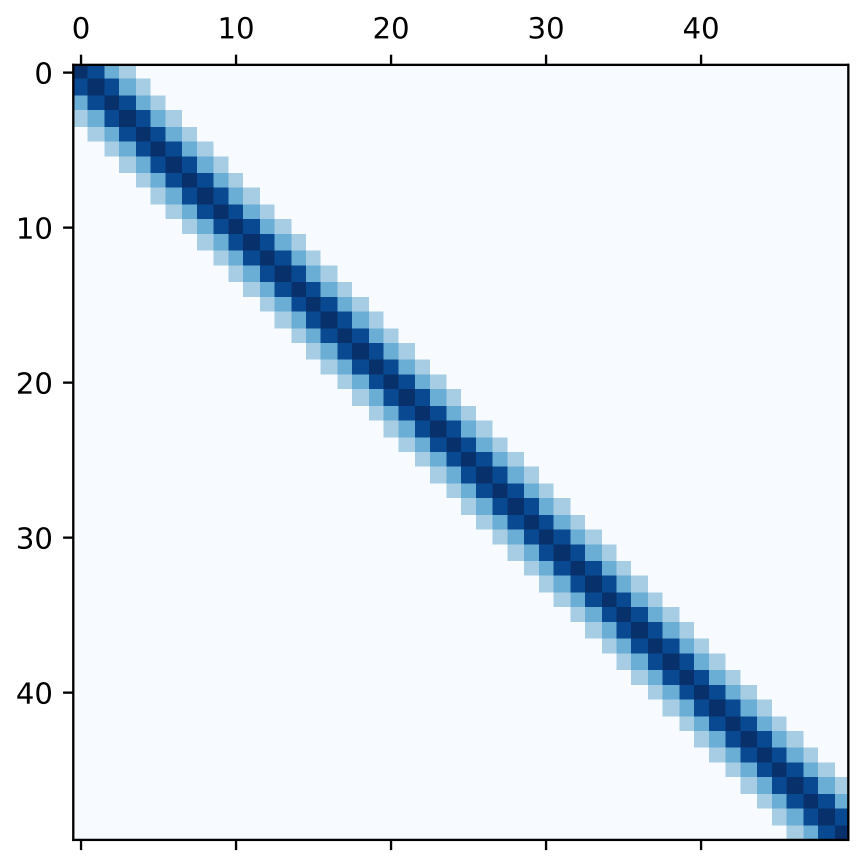
Algorithmic efficiency

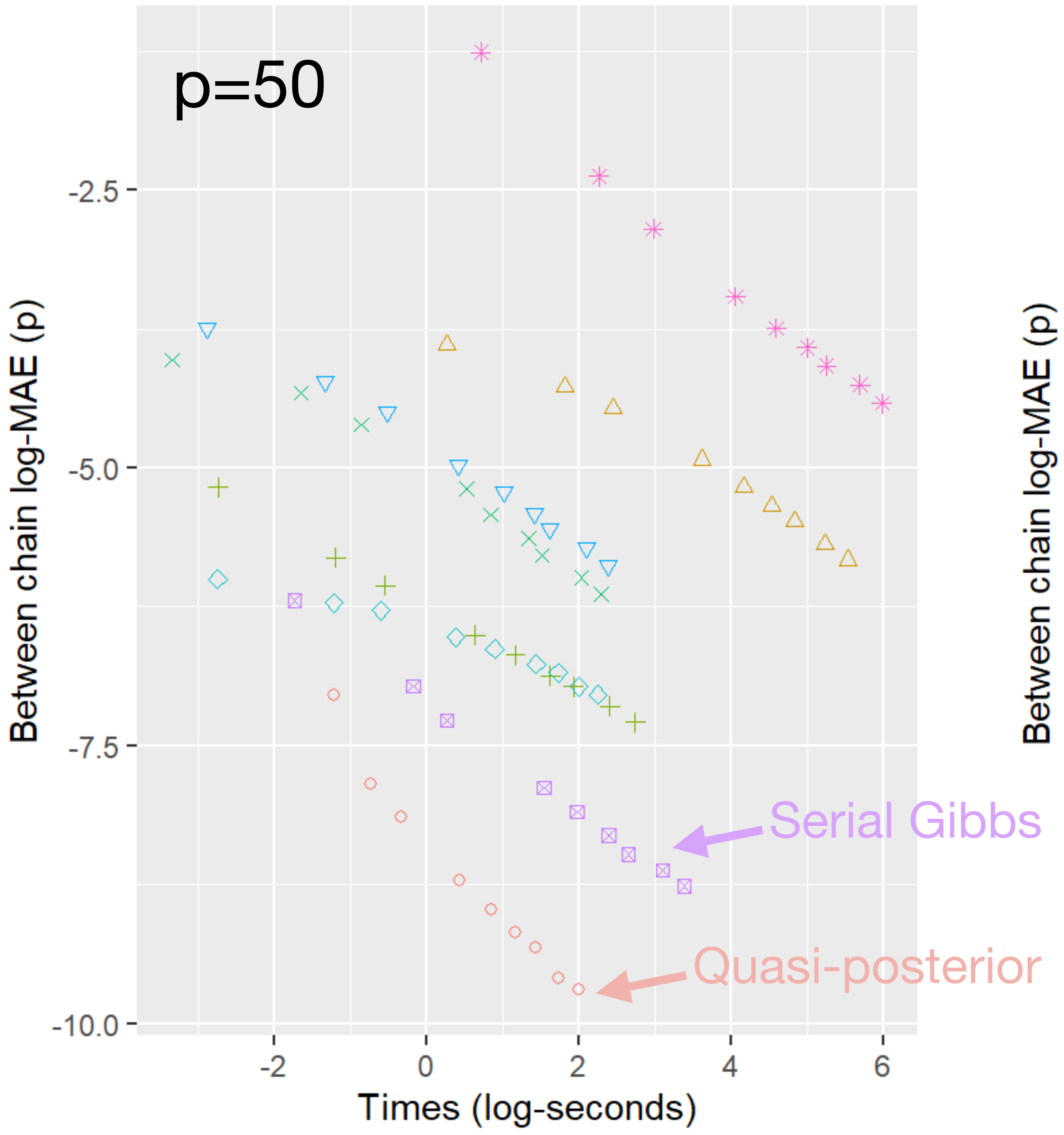
Mixing time vs clock time

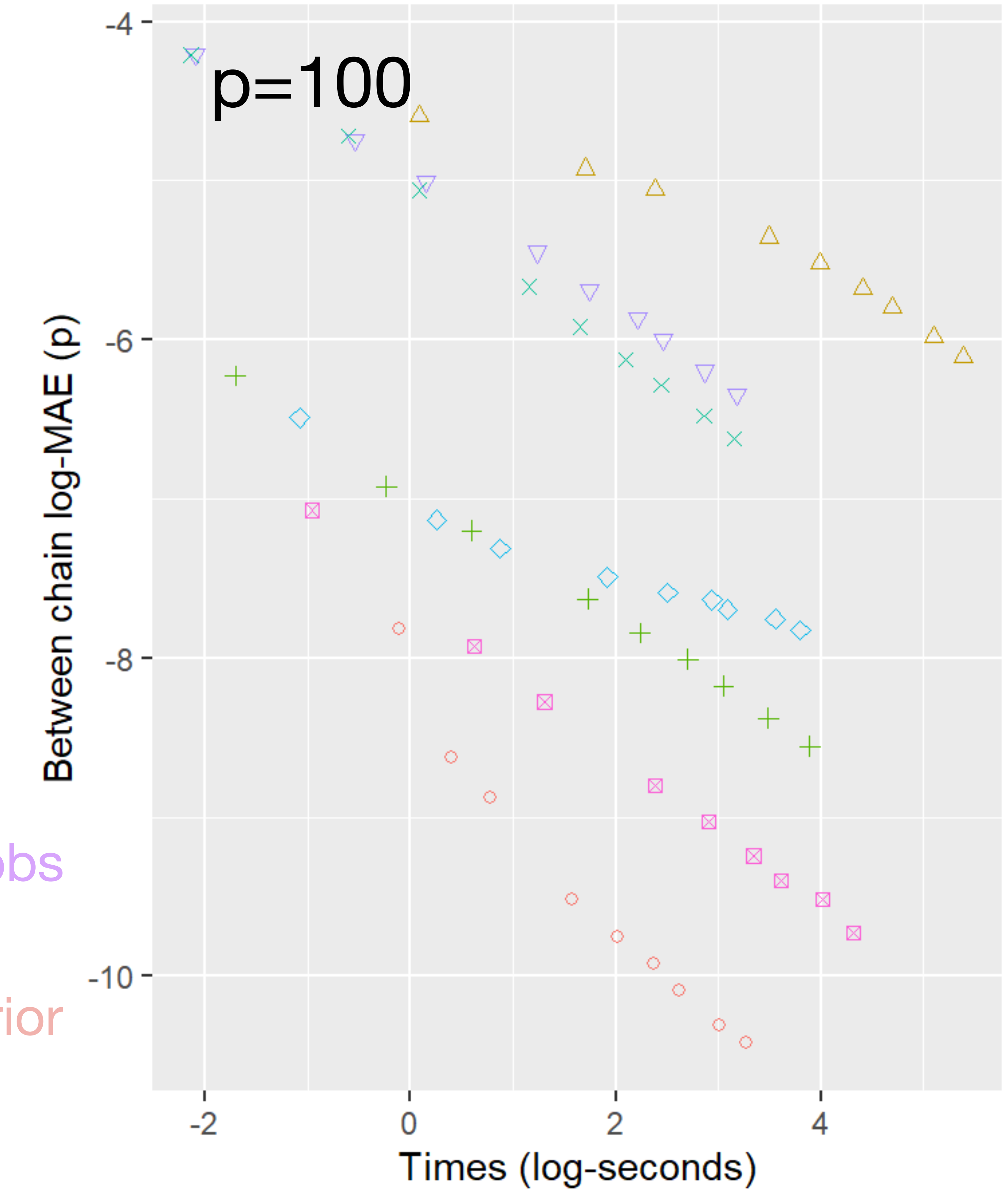


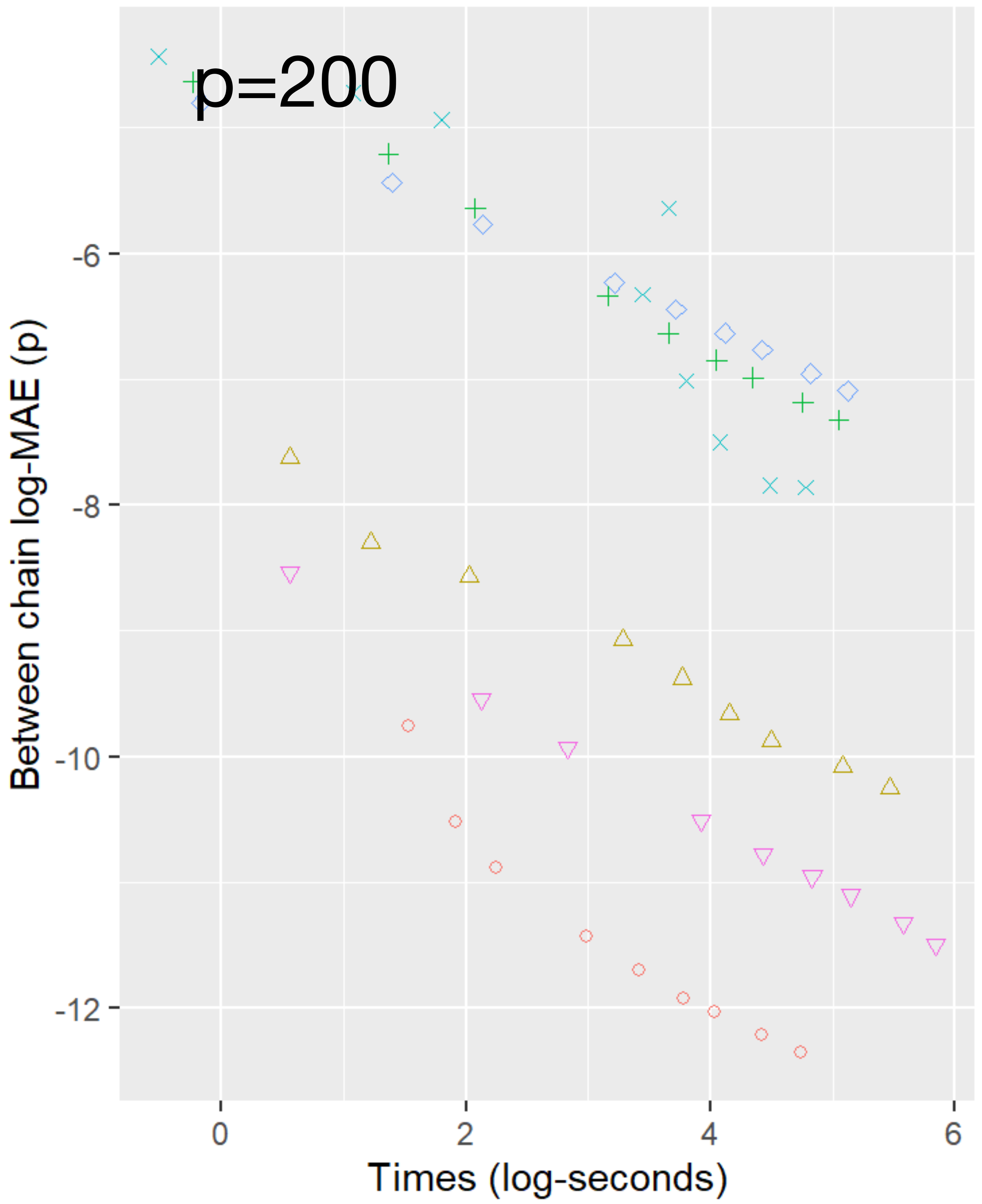
Algorithmic efficiency

Mixing time vs clock time: p=50,100, 200, n=2p

$\Omega_0 =$ 

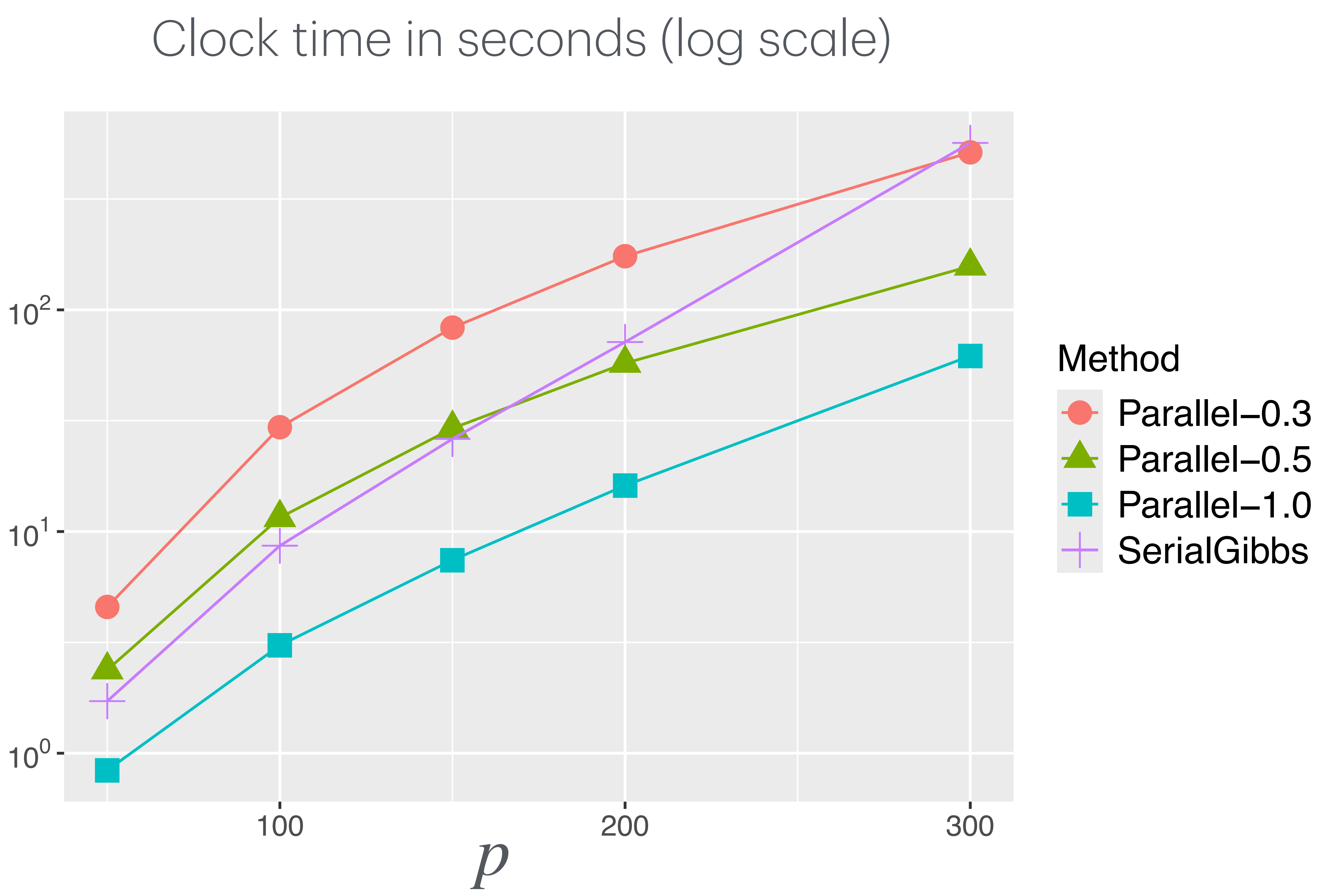


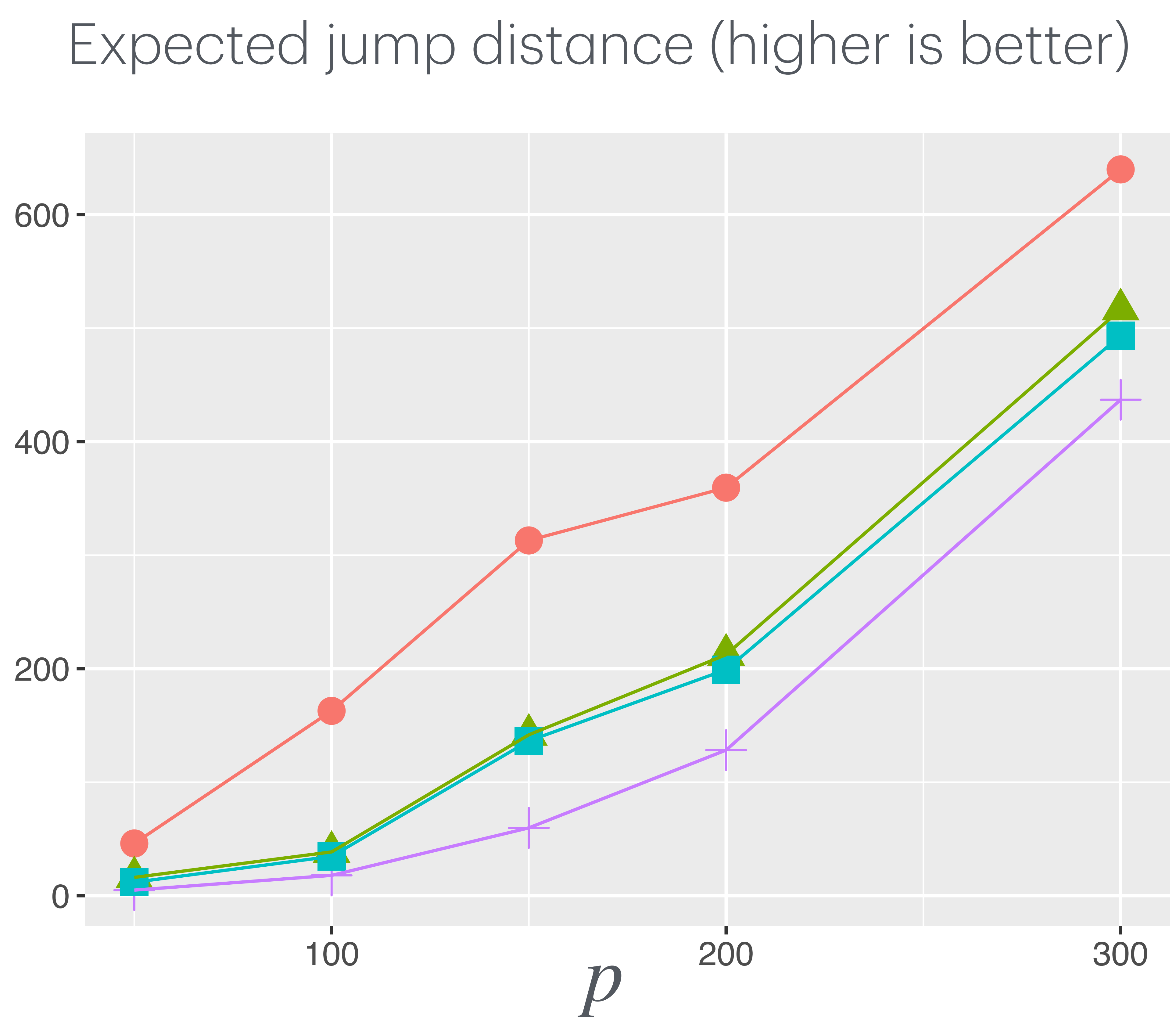




Comparison local/global steps

Clock time and Mixing quality in an ill-conditioned case





- Parallelisation can reduce clock time
- Global moves & tempering can improve mixing for particular (non-concentrated) posterior

Summary

- Novel MCMC algorithms for high-dimensional GGM with discrete spike-and-slab prior, implemented in the mombf[1] R package.
- We propose local and globally-informed steps, leveraging a toolbox of algorithms for linear regression and allowing an almost-parallel algorithm.
- We analyse the mixing times of the MH steps: can be "dimension-free" under sparsity conditions for
  - the local moves with Locally-Informed and Thresholded samplers
  - the global moves with tempering

    ➡ see preprint soon

Thank you for your attention!

[1] https://github.com/davidrusi/mombf

References

- Sayantan Banerjee and Subhashis Ghosal. Bayesian structure learning in graphical models. Journal of Multivariate Analysis, 136:147–162, 2015.
- Giacomo Zanella and Gareth Roberts. Scalable importance tempering and Bayesian variable selection. Journal of the Royal Statistical Society Series B: Statistical Methodology, 81(3):489–517, 2019.
- Hao Wang. Bayesian graphical lasso models and efficient posterior computation. Bayesian Analysis, 7(4):867–886, 2012.
- Hao Wang. Scaling it up: Stochastic search structure learning in graphical models. Bayesian Analysis, 10(2):351–377, 2015.
- Willem van den Boom, Alexandros Beskos, and Maria De Iorio. The G-Wishart weighted proposal algorithm: Efficient posterior computation for Gaussian graphical models. Journal of Computational and Graphical Statistics, 31(4): 1215–1224, 2022.
- Reza Mohammadi, Helene Massam, and Gerard Letac. Accelerating Bayesian structure learning in sparse gaussian graphical models. Journal of the American Statistical Association, pages 1–14, 2021.
- Edward I George and Robert E McCulloch. Variable selection via Gibbs sampling. Journal of the American Statistical Association, 88(423):881–889, 1993.
- Mohammadi, R., Schoonhoven, M., Vogels, L., and Birbil, S. I. High-Dimensional Bayesian Structure Learning in Gaussian Graphical Models using Marginal Pseudo-Likelihood. arXiv preprint arXiv:2307.00127, 2023.
- Narisetty, N. N., & He, X. Bayesian variable selection with shrinking and diffusing priors. Annals of Statistics, 42(2), 789-817, 2014.
- Zhou, Q., Yang, J., Vats, D., Roberts, G.O. and Rosenthal, J.S.. Dimension-free mixing for high-dimensional Bayesian variable selection. Journal of the Royal Statistical Society Series B: Statistical Methodology, 84(5), pp.1751-1784, 2022.