# The $G$-Wishart Weighted Proposal Algorithm: Efficient Posterior Computation for Gaussian Graphical Models

**Willem van den Boom**
Yong Loo Lin School of Medicine, National University of Singapore

2022 World Meeting of the
International Society for Bayesian Analysis

June 29, 2022

# Coauthors

Alexandros Beskos
UCL, ATI

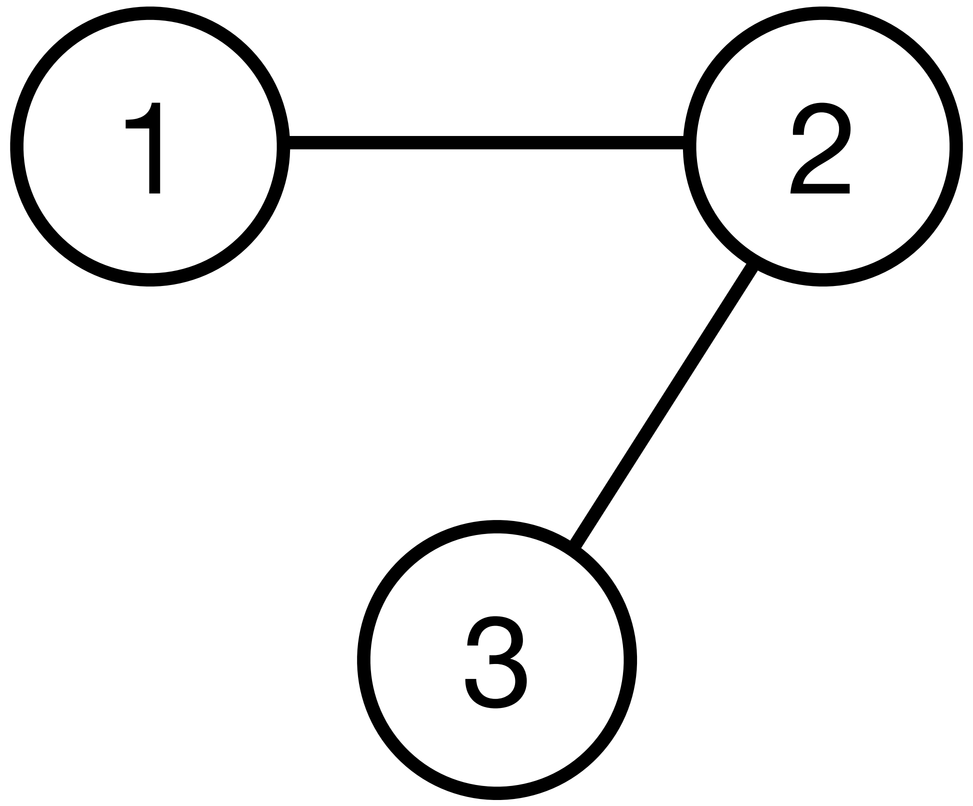
Maria De Iorio
NUS, A*STAR, UCL

# Motivation

- Inference in graphical models is hard: $2^{p(p-1)/2}$ possible graphs on $p$ nodes
- Variety of approaches
  - Estimating the precision through penalized likelihood: Graphical lasso (Friedman et al., 2007), nodewise regression (Zhou et al., 2011), precision factor model (Chandra et al., 2022)
  - Continuous spike-and-slab prior (Wang, 2015; Li et al., 2020)
  - Subset of graph space: Decomposable graphs (Giudici and Green, 1999)
  - Full graph space with the $G$-Wishart prior (Giudici, 1996), e.g., BDgraph (Mohammadi and Wit, 2019) and the double conditional Bayes factor sampler (Hinne et al., 2014)
- The $G$-Wishart weighted proposal algorithm (WWA): Speed up through advances in MCMC
  - Informed proposal (Zanella, 2019)
  - Delayed Metropolis-Hastings acceptance (Christen and Fox, 2005)

# Gaussian graphical models

- $Y_i \mid K, G \sim \mathcal{N}(0, K^{-1})$ independently for $i = 1, \ldots, n$
  - $Y_i \in \mathbb{R}^p$
- Graph $G$ specifies conditional independencies.
  - $G = \{(1,2), (2,3)\}$
  - $Y_{i1} \perp Y_{i3} \mid Y_{i2}$ if and only if $K_{13} = 0$
  - $(i,j) \notin G \Rightarrow K_{ij} = 0$




$Y_i \in \mathbb{R}^3$

$Y_{i1} \perp Y_{i3} \mid Y_{i2}$

# Prior specification

- Hierarchical prior $p(K, G) = p(K \mid G)\, p(G)$
- $G$-Wishart distribution as prior on the precision matrix (Giudici, 1996)
  - Wishart distribution truncated to matrices with $K_{ij} = 0$ for $(i, j) \notin G$
  - $p(K \mid G) = \mathscr{W}_G(K \mid \delta, D) = \dfrac{1}{I_G(\delta, D)} \, |K|^{\delta/2-1} \exp\left\{ -\frac{1}{2}\mathrm{tr}(K^\top D) \right\}$
- $K \mid G, Y \sim \mathscr{W}_G(\delta^\star, D^\star)$ where $\delta^\star = \delta + n$ and $D^\star = D + Y^\top Y$

# Goal

- The objective is inference on the graph $G$.

$$p(G \mid Y) \propto p(G) \int p(K \mid G)\, p(Y \mid K)\, dK = \frac{p(G)\, I_G(\delta^\star, D^\star)}{(2\pi)^{np/2}\, I_G(\delta, D)}$$

- The normalizing constant $I_G(\delta, D)$ is intractable.

- Posterior computation is challenging and of interest.
  - Roverato, 2002; Dellaportas et al., 2003; Atay-Kayis and Massam, 2005; Moghaddam et al., 2009; Lenkoski and Dobra, 2011; Wang and Li, 2012; Cheng and Lenkoski, 2012; Lenkoski, 2013; Hinne et al., 2014; Mohammadi and Wit, 2015, 2019; Mohammadi et al., 2021

# WWA's speedups

**Problem:** Expensive to resample the full precision matrix $K$

**Solution:** Gibbs update based on the Cholesky decomposition

**Problem:** Exploring the large discrete space of graphs

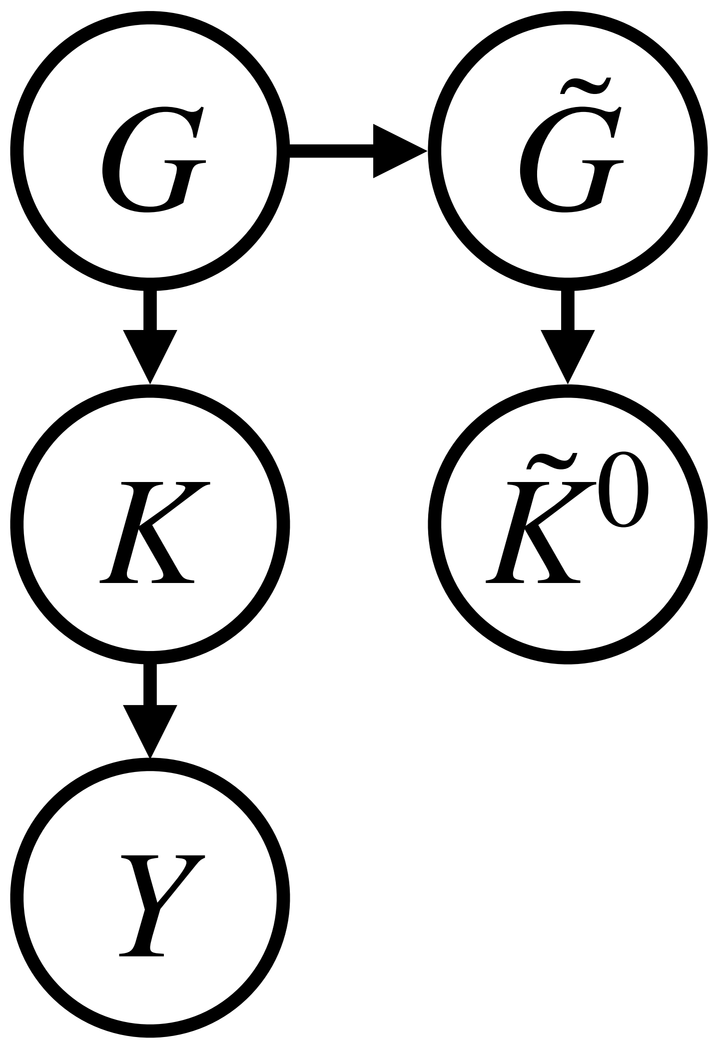
**Solution:** Informed proposal (Zanella, 2019) using a posterior approximation (Mohammadi et al., 2021) resulting in more efficient exploration

**Problem:** Most computational effort on proposed graphs that are rejected

**Solution:** Reject fast and cheaply using delayed Metropolis-Hastings acceptance (Christen and Fox, 2005)

# Review: Exchange algorithm

- The exchange algorithm (Murray et al, 2006) avoids evaluating $I_G(\delta, D)$ (Wang and Li, 2012).



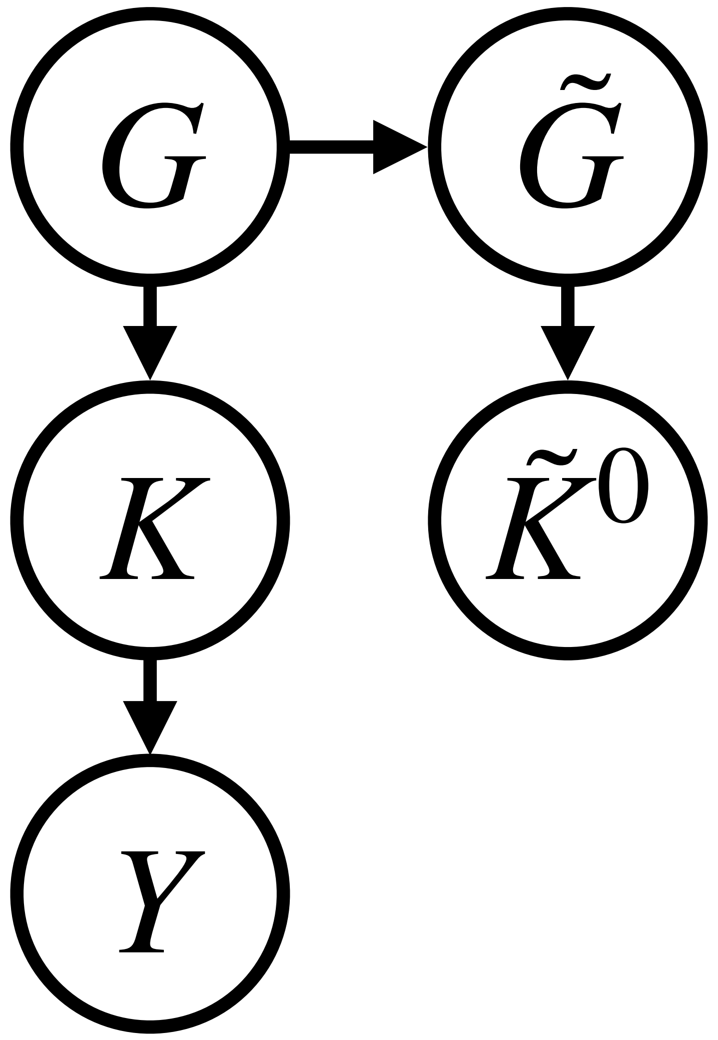

- Augmented state space $(K, G, \tilde{K}^0, \tilde{G})$:
  - Propose $\tilde{G}$ from $G$ by changing one edge.
  - "Prior" precision $\tilde{K}^0 \mid \tilde{G} \sim \mathscr{W}_{\tilde{G}}(\delta, D)$
  - Metropolis-Hastings acceptance ratio:

$$\frac{p(\tilde{K}, \tilde{G}, K^0, G)}{p(K, G, \tilde{K}^0, \tilde{G})} = \frac{p(\tilde{K} \mid \tilde{G})\, p(\tilde{G})\, p(Y \mid \tilde{K})\, p(K^0 \mid G)\, q(G \mid \tilde{G})}{p(K \mid G)\, p(G)\, p(Y \mid K)\, p(\tilde{K}^0 \mid \tilde{G})\, q(\tilde{G} \mid G)}$$

Exchange

# Review: Exchange algorithm

- The exchange algorithm (Murray et al, 2006) avoids evaluating $I_G(\delta, D)$ (Wang and Li, 2012).




- Augmented state space $(K, G, \tilde{K}^0, \tilde{G})$:
  - Propose $\tilde{G}$ from $G$ by changing one edge.
  - "Prior" precision $\tilde{K}^0 \mid \tilde{G} \sim \mathscr{W}_{\tilde{G}}(\delta, D)$
  - Metropolis-Hastings acceptance ratio:

$$\frac{p(\tilde{K}, \tilde{G}, K^0, G)}{p(K, G, \tilde{K}^0, \tilde{G})} = \frac{\boxed{p(\tilde{K} \mid \tilde{G})}\, p(\tilde{G})\, p(Y \mid \tilde{K})\, \boxed{p(K^0 \mid G)}\, q(G \mid \tilde{G})}{\boxed{p(K \mid G)}\, p(G)\, p(Y \mid K)\, \boxed{p(\tilde{K}^0 \mid \tilde{G})}\, q(\tilde{G} \mid G)}$$

Exchange

$I_G(\delta, D)$ and $I_{\tilde{G}}(\delta, D)$ cancel out.

**Problem:** Expensive to resample the full precision matrix $K$

**Solution:** Gibbs update based on the Cholesky decomposition

- Augmenting the state space by sampling $K \mid G \sim \mathscr{W}_G(\delta^\star, D^\star)$ is the main computational bottleneck.
- If $\tilde{G}$ is accepted, update $K$ such that it is distributed as $\mathscr{W}_{\tilde{G}}(\delta^\star, D^\star)$.
- Sampling one element of a Cholesky decomposition suffices.
  - Distribution of the Cholesky decomposition of a permuted $K$ differs in only one element for $G$ and $\tilde{G}$.
  - WWA samples this element from its Gaussian full conditional.

**Problem:** Exploring the large discrete space of graphs

**Solution:** Informed proposal

- Informed proposal puts more mass on regions with higher posterior probability (Zanella, 2019):

$$Q(\tilde{G} \mid G) \propto g\left\{\frac{p(\tilde{G} \mid Y)}{p(G \mid Y)}\right\} q(\tilde{G} \mid G)$$

  - Embarrassingly parallel scan over all $\tilde{G}$ that differ from $G$ by one edge.
  - WWA uses balancing function $g(t) = t/(1+t)$.

- WWA uses a modified version because $p(G \mid Y)$ is intractable:

$$Q(\tilde{G} \mid K, G) \propto g\left\{\frac{p(K, \tilde{G} \mid Y)}{p(K, G \mid Y)}\right\} q(\tilde{G} \mid G)$$

  - The ratio involves a ratio of intractable normalization constants $I_G(\delta, D)$.
  - Fast approximation of this ratio (Mohammadi et al., 2021)

**Problem:** Most computational effort on proposed graphs that are rejected

**Solution:** Reject fast and cheaply

- Delayed Metropolis-Hasting acceptance exploits a fast approximate acceptance probability (Christen and Fox, 2005).

- Approximation of ratio of $I_G(\delta, D)$ (Mohammadi et al., 2021) provides approximate acceptance without:
  - Exchange algorithm
  - The computational bottleneck of sampling $\tilde{K}^0 \mid \tilde{G} \sim \mathscr{W}_{\tilde{G}}(\delta, D)$

- The exchange algorithm only if $\tilde{G}$ is "promoted" to be considered for delayed acceptance

# Overview of WWA

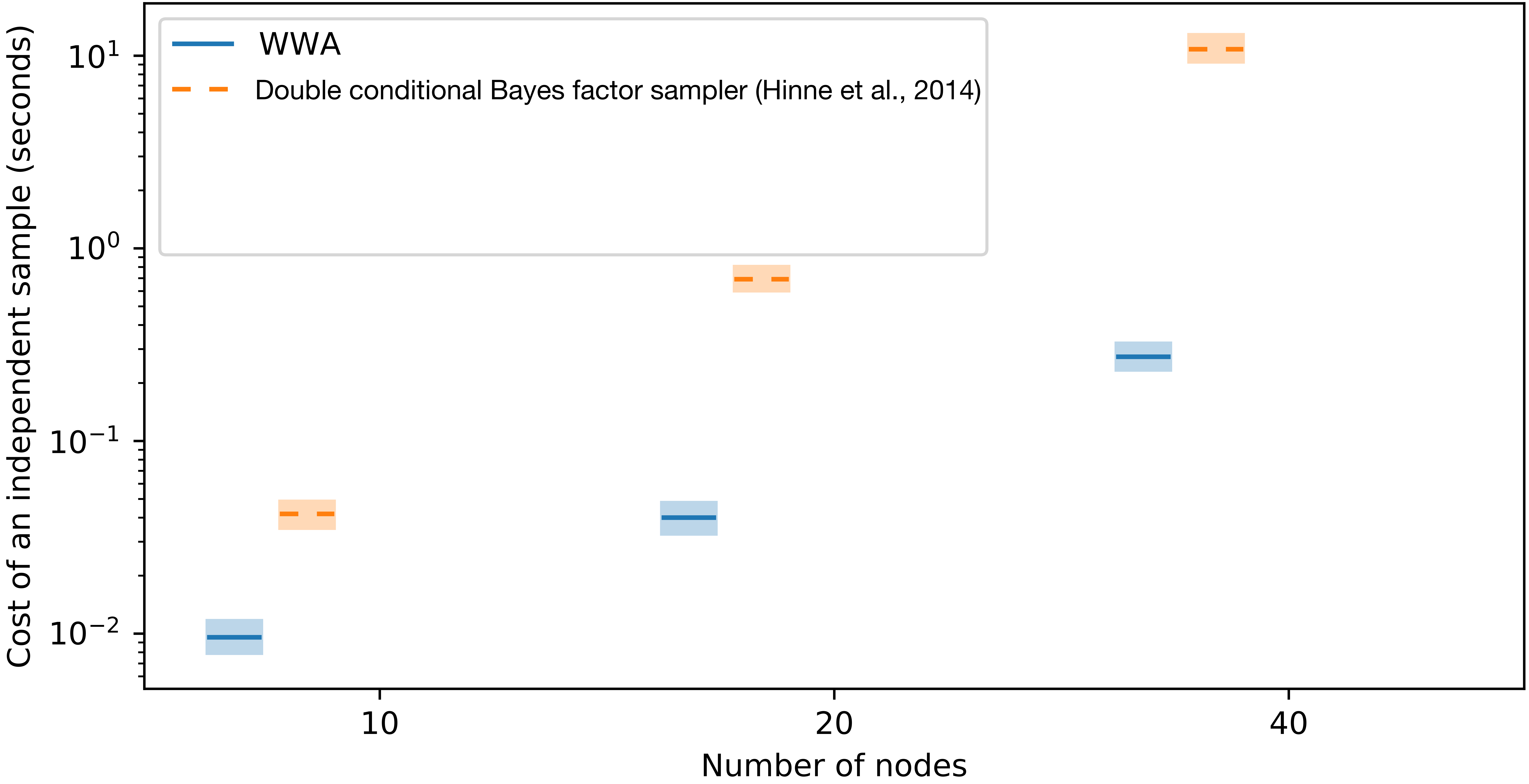
- MCMC step with $p(K, G \mid Y)$ as stationary distribution:
  - Propose $\tilde{G}$ from the informed proposal $Q(\tilde{G} \mid K, G)$.
  - Obtain $\tilde{K} \sim \mathscr{W}_{\tilde{G}}(\delta^\star, D^\star)$ by a Gibbs update and compute $Q(G \mid \tilde{K}, \tilde{G})$.
  - “Promote” $\tilde{G}$ using an approximate acceptance probability.
  - If promoted, sample $\tilde{K}^0 \mid \tilde{G} \sim \mathscr{W}_{\tilde{G}}(\delta, D)$ and set $(K, G) = (\tilde{K}, \tilde{G})$ with probability deriving from the exchange algorithm and delayed acceptance.
- Computation of $Q(\tilde{G} \mid K, G)$ is embarrassingly parallel.
- The computational bottleneck $\mathscr{W}_{\tilde{G}}(\delta, D)$ only occurs if $\tilde{G}$ is promoted.

# Empirical results

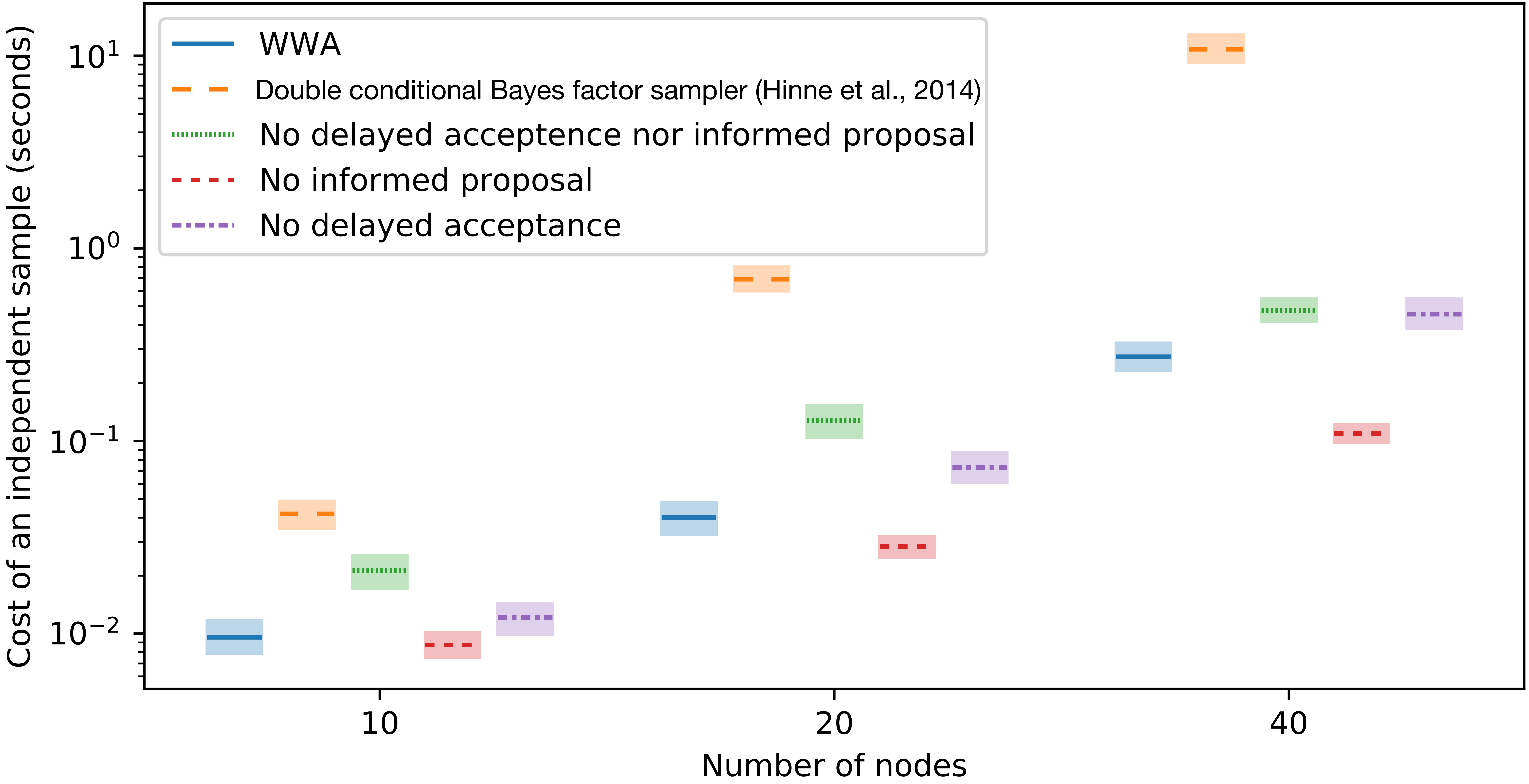
- Quantify MCMC efficiency:
  - Computational cost of a single MCMC iteration
  - MCMC mixing, e.g., effective sample size

$$\text{cost of an independent sample} = \frac{\text{number of MCMC steps}}{\text{effective sample size}} \times \text{cost per step}$$

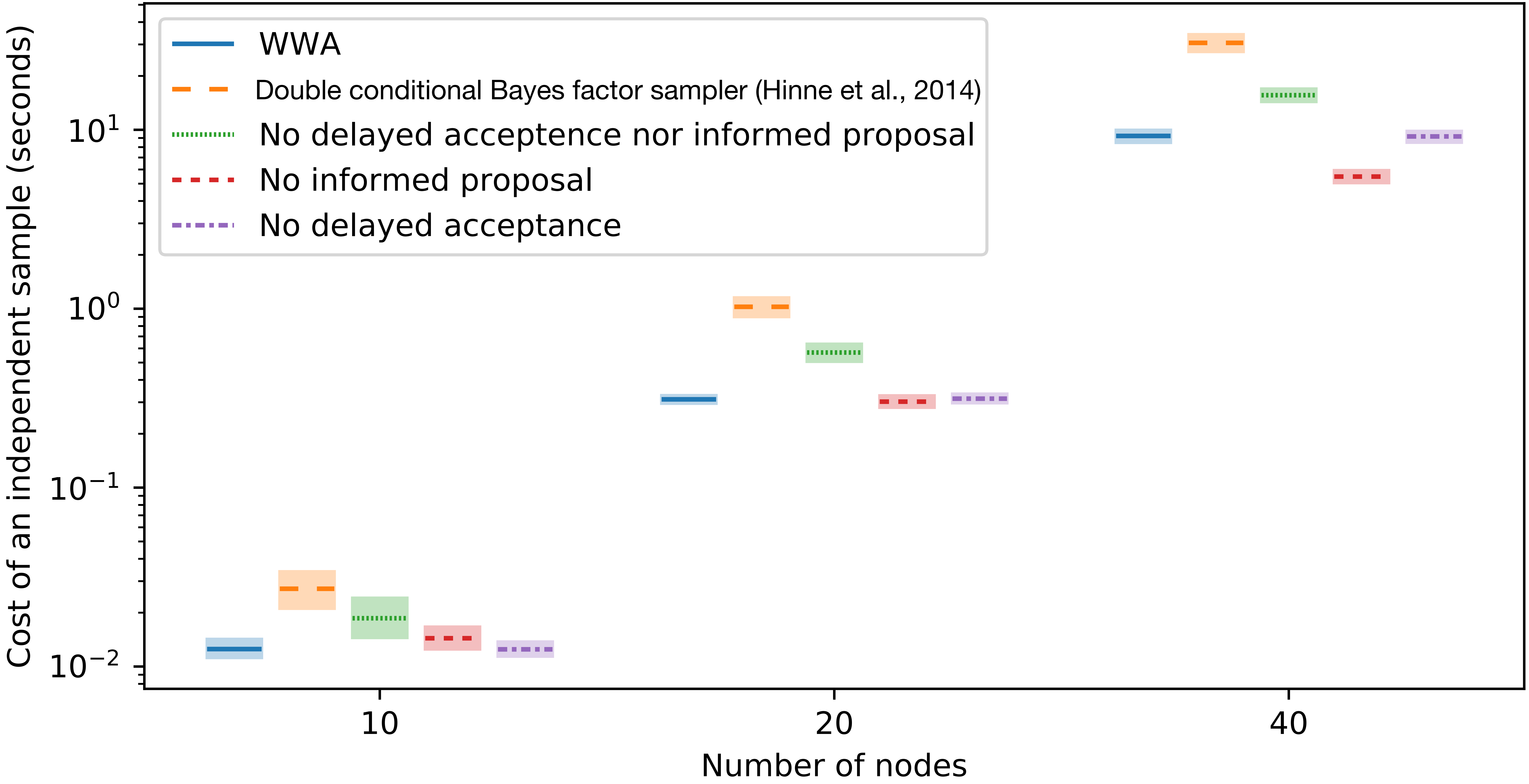
# Data simulated from cycle graphs



# Data simulated from cycle graphs



# Data simulated from uniformly sampled graphs



# Conceptual comparison of MCMC

| ***Feature*** | **(Exact) exchange algorithm** | **Gibbs update for *K*** | **Embarrassingly parallel** |
|---|---|---|---|
| *Benefit* | Stationary distribution equals the posterior | Reduced computational cost | Exploit available compute power |
| Wang and Li (2012) | ✗ | ✗ | ✗ |
| Cheng and Lenkoski (2012) | ✗ | ✓ | ✗ |
| Hinne et al. (2014) | ✓ | ✗ | ✗ |
| Mohammadi and Wit (2015) | ✓ | ✗ | ✓ |
| WWA | ✓ | ✓ | ✓ |

# Conclusion

- WWA provides a major speedup in MCMC with the $G$-Wishart prior.
  - Avoid resampling of the precision matrix via a Gibbs update
  - Informed proposal (Zanella, 2019) on the graph space
  - Reject fast and cheaply via delayed acceptance (Christen and Fox, 2005)
- Single edge update limits performance.
- WWA enables inference on the whole graph space in more scenarios:
  - Sparse seemingly unrelated regression (work in progress)
  - Learning of graph substructures (van den Boom et al., 2022)

# Related article

van den Boom, W., Beskos, A., and De Iorio, M. (2022). The G-Wishart weighted proposal algorithm: Efficient posterior computation for Gaussian graphical models. *Journal of Computational and Graphical Statistics*, advance online publication.

# References

Atay-Kayis, A., and Massam, H. (2005), A Monte Carlo Method for Computing the Marginal Likelihood in Nondecomposable Gaussian Graphical Models, Biometrika, 92:317–335.

Bhadra, A., and Mallick, B. K. (2013), Joint High-dimensional Bayesian Variable and Covariance Selection with an Application to eQTL Analysis, Biometrics, 69:47–457.

Chandra, N. K., Müller, P., and Sarkar, A. (2022), Bayesian Scalable Precision Factor Analysis for Massive Sparse Gaussian Graphical Models. arXiv:2107.11316v3

Cheng, Y., and Lenkoski, A. (2012), Hierarchical Gaussian Graphical Models: Beyond Reversible Jump, Electronic Journal of Statistics, 6:2309–2331.

Christen, J. A., and Fox, C. (2005), Markov Chain Monte Carlo Using an Approximation, Journal of Computational and Graphical Statistics, 14:795–810.

Dellaportas, P., Giudici, P., and Roberts, G. (2003), Bayesian Inference for Nondecomposable Graphical Gaussian Models, Sankhyā: The Indian Journal of Statistics, 65:43–55.

Friedman, J., Hastie, T., and Tibshirani, R. (2007), Sparse Inverse Covariance Estimation with the Graphical Lasso, Biostatistics, 9:432–441.

Giudici, P. (1996), Learning in Graphical Gaussian Models, in Bayesian Statistics 5, pp. 621–628.

Giudici, P., and Green, P. J. (1999), Decomposable Graphical Gaussian Model Determination, Biometrika, 86:785–801.

Hinne, M., Lenkoski, A., Heskes, T., and van Gerven, M. (2014), Efficient Sampling of Gaussian Graphical Models Using Conditional Bayes Factors, Stat, 3:326–336.

Lenkoski, A. (2013), A Direct Sampler for G-Wishart Variates, Stat, 2:119–128.

Lenkoski, A., and Dobra, A. (2011), Computational Aspects Related to Inference in Gaussian Graphical Models with the G-Wishart Prior, Journal of Computational and Graphical Statistics, 20:140–157.

# References

Li, Z. R., McCormick, T. H., and Clark, S. J. (2020), Using Bayesian Latent Gaussian Graphical Models to Infer Symptom Associations in Verbal Autopsies, Bayesian Analysis, 15:781–807.

Moghaddam, B., Khan, E., Murphy, K. P., and Marlin, B. M. (2009), “Accelerating Bayesian Structural Inference for Non-decomposable Gaussian Graphical Models,” in Advances in Neural Information Processing Systems (Vol. 22).

Mohammadi, R., Massam, H., and Letac, G. (2021), Accelerating Bayesian Structure Learning in Sparse Gaussian Graphical Models, Journal of the American Statistical Association, advance online publication.

Mohammadi, A., and Wit, E. C. (2015), Bayesian Structure Learning in Sparse Gaussian Graphical Models, Bayesian Analysis, 10:109–138.

Mohammadi, R., and Wit, E. C. (2019). BDgraph: An R package for Bayesian Structure Learning in Graphical Models. Journal of Statistical Software, 89(3).

Murray, I., Ghahramani, Z., and MacKay, D. J. C. (2006), MCMC for Doubly-intractable Distributions, in Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence, pp. 359–366.

Roverato, A. (2002). Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. Scandinavian Journal of Statistics, 29:391–411.

van den Boom, W., De Iorio, M., and Beskos, A. (2022). Bayesian Learning of Graph Substructures. arXiv:2203.11664

Wang, H. (2015), Scaling it Up: Stochastic Search Structure Learning in Graphical Models, Bayesian Analysis, 10:351–377.

Wang, H., and Li, S. Z. (2012), Efficient Gaussian Graphical Model Determination Under G-Wishart Prior Distributions, Electronic Journal of Statistics, 6:168–198.

Zanella, G. (2019), Informed Proposals for Local MCMC in Discrete Spaces, Journal of the American Statistical Association, 115:852–865.

Zhou, S., Rütimann, P., Xu, M., and Bühlmann, P. (2011), High-dimensional Covariance Estimation Based on Gaussian Graphical Models, The Journal of Machine Learning Research, 12:2975–3026.

NUS
National University of Singapore
Yong Loo Lin School of Medicine

# We are hiring!

Division of Biomedical Data Science
led by Prof. Maria De Iorio

Postdoc positions, PhD studentships

# Thank you!