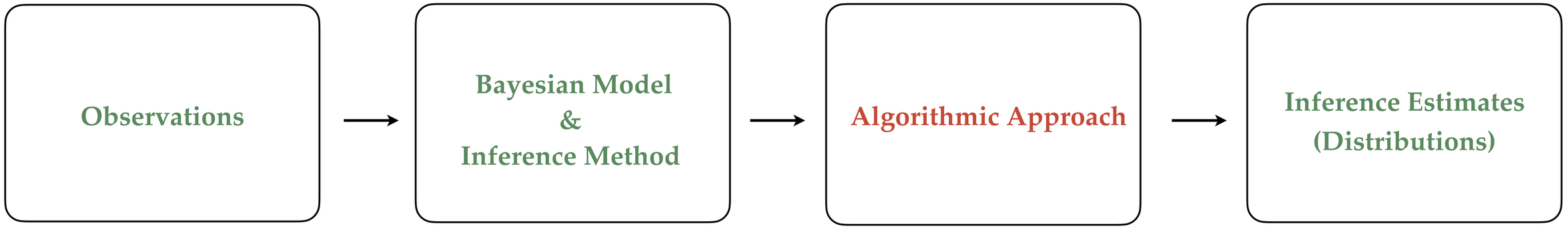
Sparse Algorithms for Large-Scale Bayesian Inference Problems

Lisa Gaedke-Merzhäuser[1]

Matthias Bollhöfer[2]
Håvard Rue[3]
Olaf Schenk[1]

[1]Università della Svizzera italiana, Lugano, Switzerland
[2]Technical University of Braunschweig
[3]King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

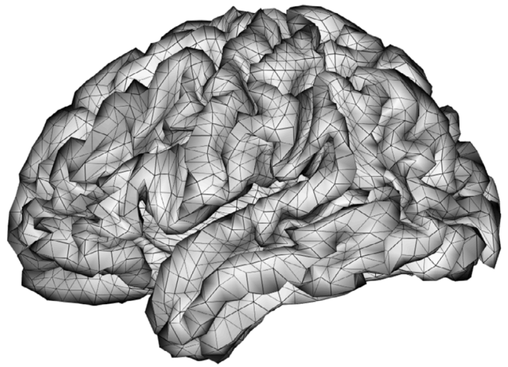
Overview

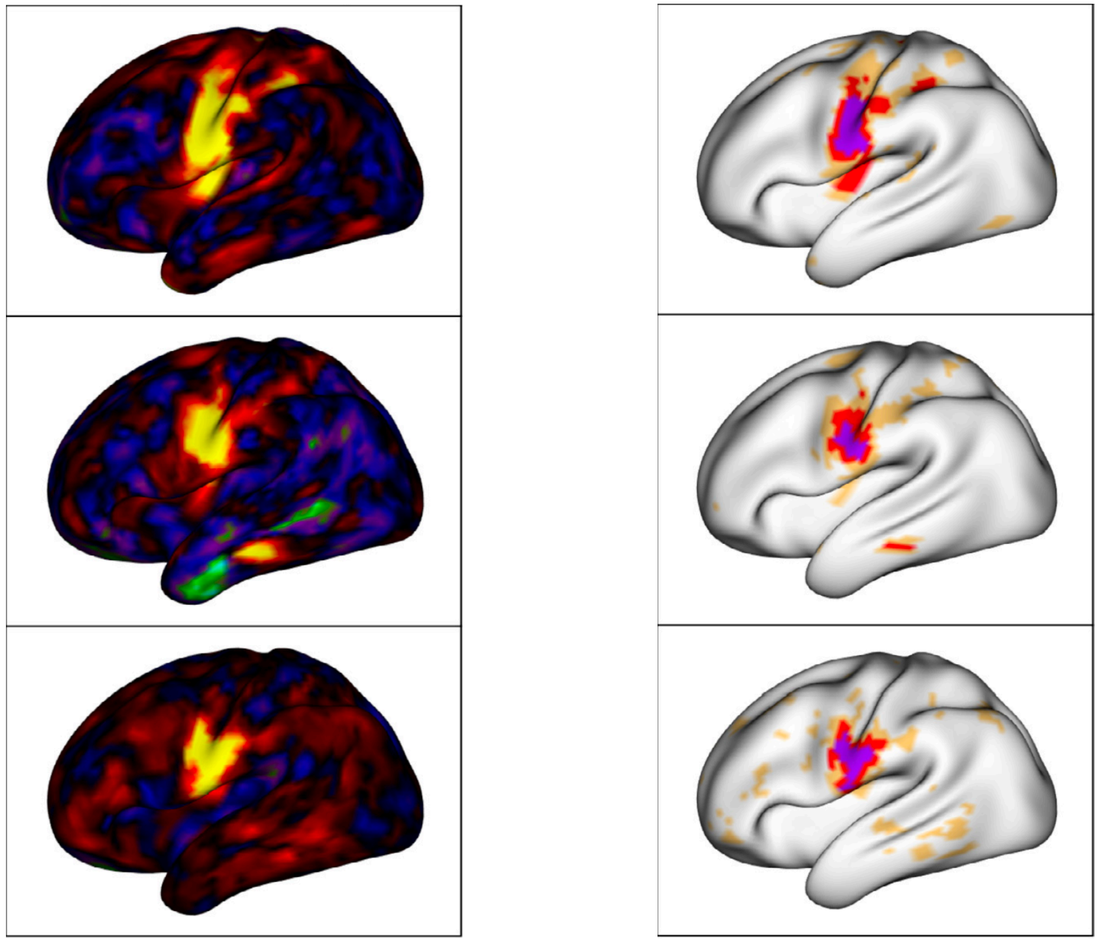


Linear Solvers

Cortical Surface Modeling of Human Brain Activation[1]

- fMRI data
- Estimate functional brain responses
- Spatial Bayesian generalized linear model

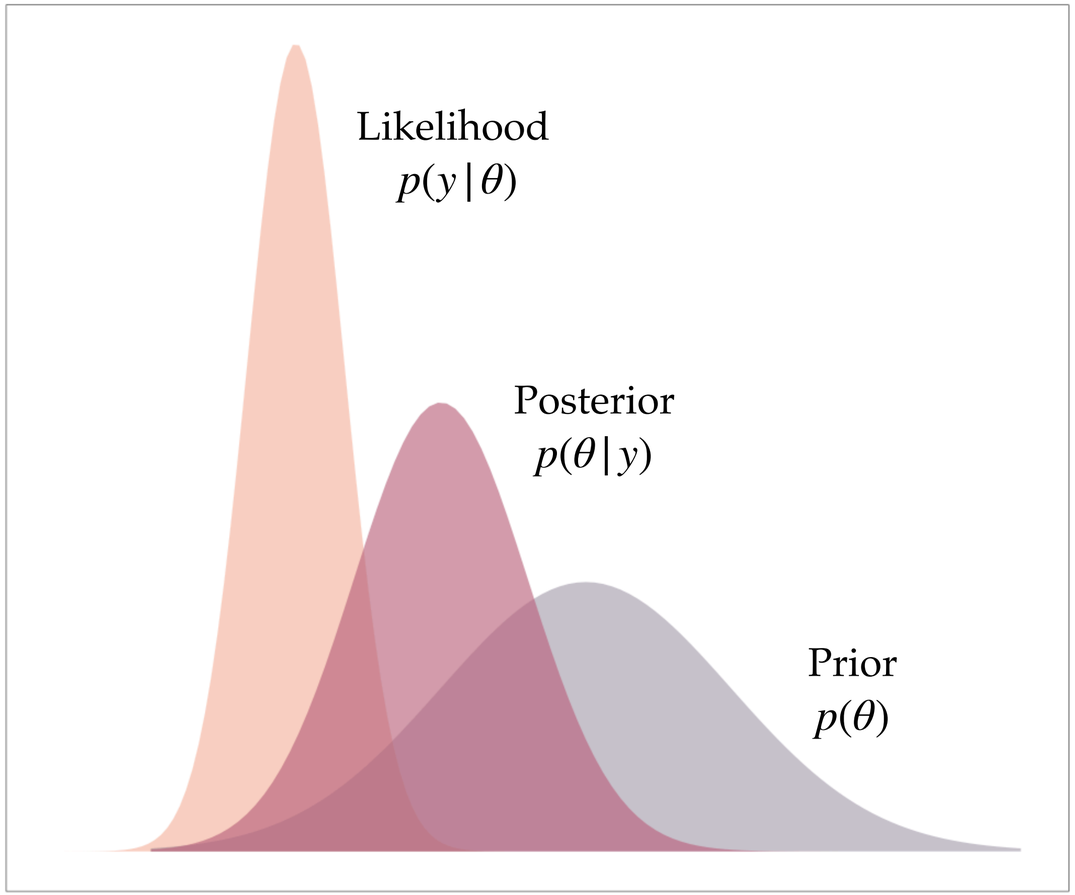




[1]Spencer, Daniel, et al. "Spatial Bayesian GLM on the cortical surface produces reliable task activations in individuals and groups." *NeuroImage* 249 (2022): 118908.

Bayesian Modeling

Aim: Update our beliefs upon observing data using Bayes' rule



Bayes Rule

$$p(\theta \mid y) = \frac{p(y \mid \theta)\ p(\theta)}{p(y)} \propto p(y \mid \theta)\ p(\theta)$$

Bayesian Hierarchical Modeling

***Aim**: Update our beliefs upon observing data using Bayes' rule*

Data

y

$p(y \mid x, \theta)$

Likelihood

Latent Parameters

$x \mid \theta \sim N(0, Q_x^{-1}(\theta))$

High-dimensional
Q sparse

$p(x \mid \theta)$

Prior

Hyperparameters

θ

Low-dimensional

$p(\theta)$

Prior

$$p(\theta \mid y) \propto \frac{p(\theta)\, p(x \mid \theta)\, p(y \mid x, \theta)}{p(x \mid \theta, y)}$$

WANT: $p(\theta_i \mid y),\ p(x_j \mid y)$

Integrated Nested Laplace Approximations[1] (INLA)

- Alternative to MCMC, variational inference
- Construct approx. to joint posterior $p(\theta \mid y)$
- Approximate $p(x \mid \theta, y)$ as Gaussian
  - w/ sparse precision matrix $Q_{x|y}(\theta)$

$$f(\theta) := -\log \tilde{p}(\theta \mid y) = -\log \frac{p(\theta)\, p(x \mid \theta)\, p(y \mid x, \theta)}{p_G(x \mid \theta, y)}$$

Approx. Posterior of the Hyperparameters

I. Nonlinear optimisation problem, find:

$$\min f(\theta)$$

II. Approximate marginal distributions $p(\theta_i \mid y)$

III. Approximate marginal distributions $p(x_j \mid y)$

- BFGS-algorithm, requires for k-th iterate
  - $f(\theta^k)$
  - $\nabla f(\theta^k)$: approx. w. finite differences
- Evaluate $f(\theta^*), f(\theta^{*1}), \ldots, f(\theta^{*K})$
- Compute variances
  - $Q^{-1}_{x|y}(\theta^*), Q^{-1}_{x|y}(\theta^{*1}), \ldots, Q^{-1}_{x|y}(\theta^{*K})$

[1] *Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the royal statistical society.*

Computational Bottleneck Operations INLA

- Step I & II: Evaluate $f(\theta)$ for different θ

$$f(\theta) := -\log \tilde{p}(\theta \mid y)$$

Approx. Posterior of the Hyperparameters

requires →

- **Cholesky factorization** of large s.p.d. matrix $Q_x(\theta), Q_{x|y}(\theta)$
- **Solve** $Q_{x|y}(\theta)\mu = b(\theta)$

- Step III: Compute inverse

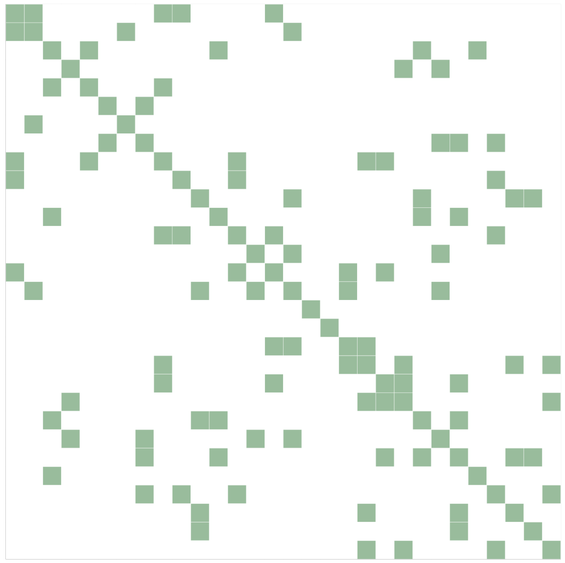
$$Q_{x|y}^{-1}(\theta*), Q_{x|y}^{-1}(\theta^{*1}), \ldots, Q_{x|y}^{-1}(\theta^{*K})$$

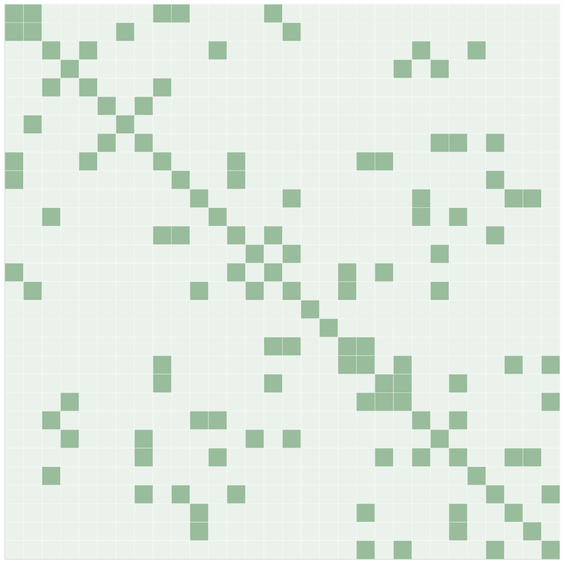
requires →

- **Selected matrix inversion** of large s.p.d. matrix $Q_{x|y}(\theta)$

$$\Sigma_{ij}(\theta) = (Q_{x|y}^{-1}(\theta))_{ij} \text{ for which } (Q_{x|y}(\theta))_{ij} \neq 0$$

Selected Matrix Inversion





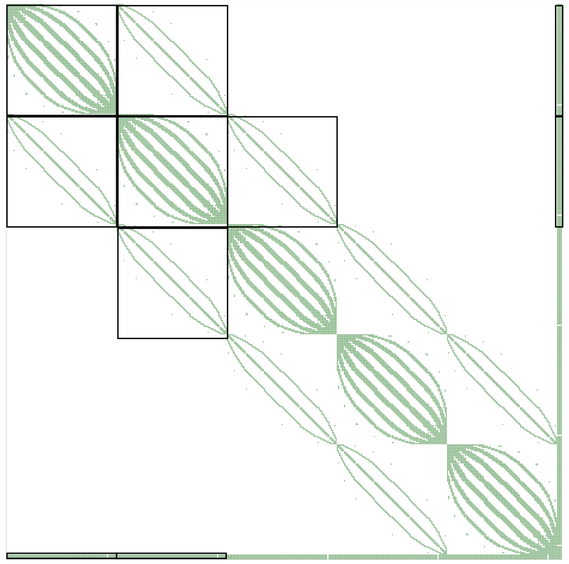
Sparse Matrix → Dense Inverse → Selected Inverse Entries

Selected Inversion

Direct Cholesky Factorization
Spatio-temporal Models

DENSE BLOCKS (BTA)

- Use “natural” block structure
- Treat each block as dense
- GPU-based (MAGMA, CUDA)



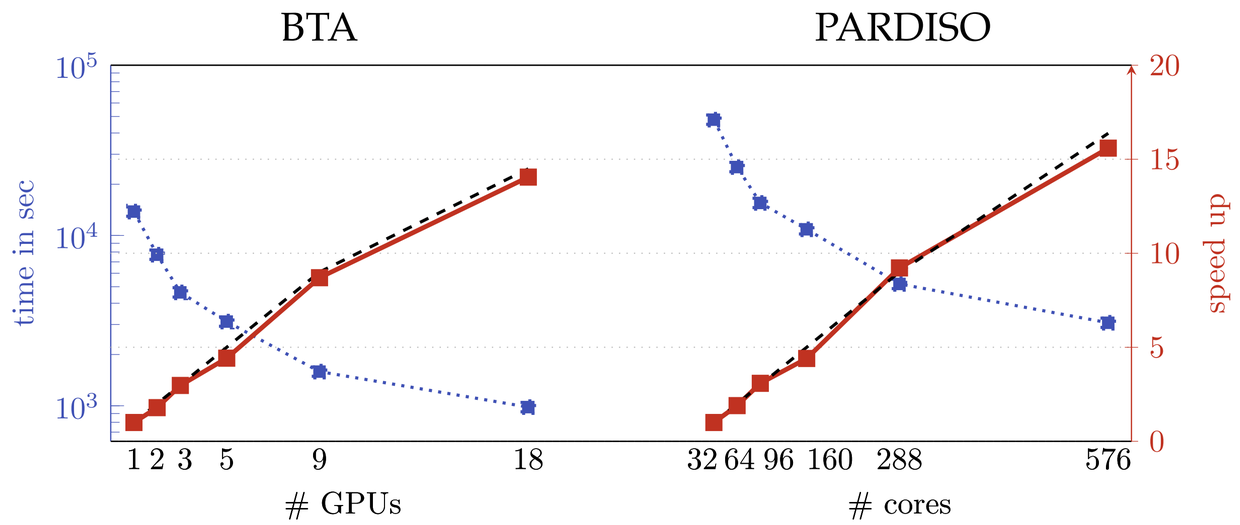
SPARSE

- Symbolic factorization
- Compute Reordering 1x
- CPU-based (PARDISO)

G.-M. L, Krainski E, Janalik R, Rue H, Schenk O. **Integrated Nested Laplace Approximations for Large-Scale Spatial-Temporal Bayesian Modeling**. arXiv preprint. (2023).

Performance Results — INLA$_{DIST}$
Spatio-temporal Models

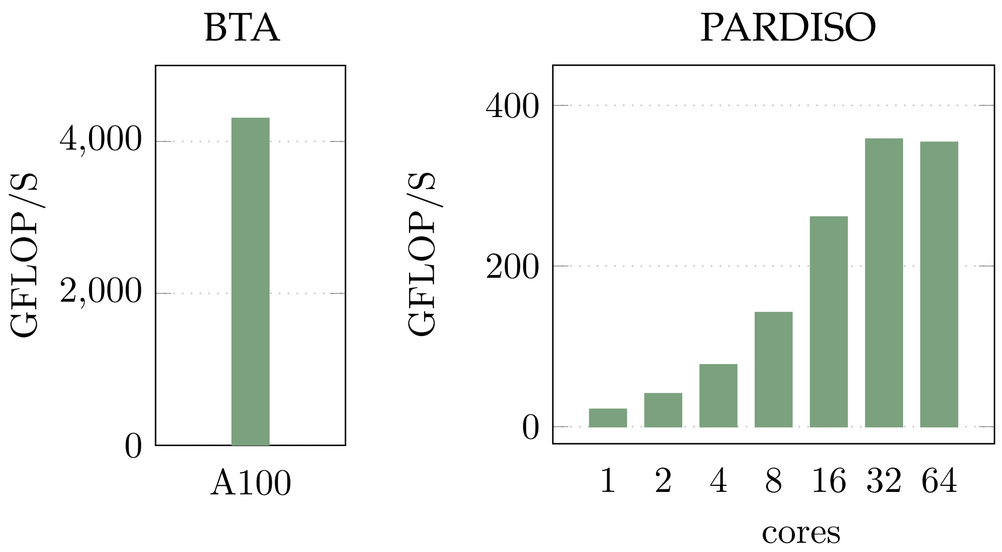
Multi Node - INLA$_{DIST}$



- BTA (GPU) : 1x factorization per GPU, A100
- PARDISO (CPU) : 1x factorization per node "Ice Lake" processors
- 18 processes for 18 parallel factorizations : $f(\theta^1), \ldots, f(\theta^K)$
  - ➡ saturates outer parallelism

- **Parallelize further within computational bottlenecks?**

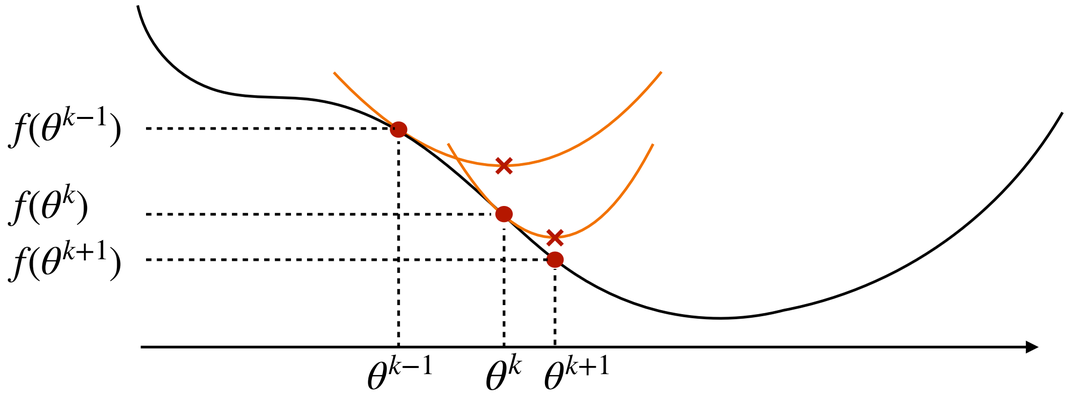
Single Node - Cholesky Factorization



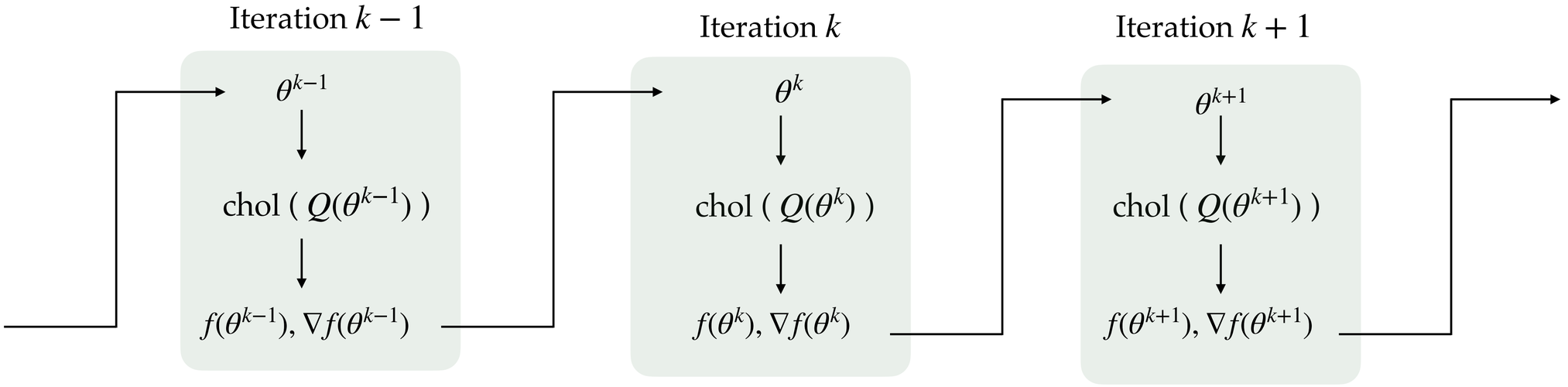
- No speed up after 32 cores
  - ➡ Many architectures have >32 cores

- Can we leverage this?
- Flexibly adapt to architecture?
- Do we have additional structural information?

BFGS Algorithm



- For every evaluation of $f(\theta^k)$: $L(\theta^k) = \text{chol}\,(\,Q(\theta^k)\,)$
- Recurrent sparsity pattern for all $Q(\theta^k)$
- Values in $Q(\theta^k)$ continuously vary with θ^k
- $Q(\theta^k)$ similar to $Q(\theta^{k+1})$, $L(\theta^k)$ similar to $L(\theta^{k+1})$



Iterative Cholesky Factorization

$$L_{ii} = \sqrt{A_{ii} - \sum_{k=1}^{i-1} (L_{ik})^2} \quad \text{for } i = j$$

$$L_{ij} = \frac{A_{ij} - \sum_{k=1}^{i-1} L_{ik} L_{jk}}{L_{ii}} \quad \text{for } i < j$$

$$L_{ii}^{(m+1)} = \sqrt{A_{ii} - \sum_{k=1}^{i-1} (L_{ik}^{(m)})^2} \quad \text{for } i = j$$

$$L_{ij}^{(m+1)} = \frac{A_{ij} - \sum_{k=1}^{i-1} L_{ik}^{(m)} L_{jk}^{(m)}}{L_{ii}} \quad \text{for } i < j$$

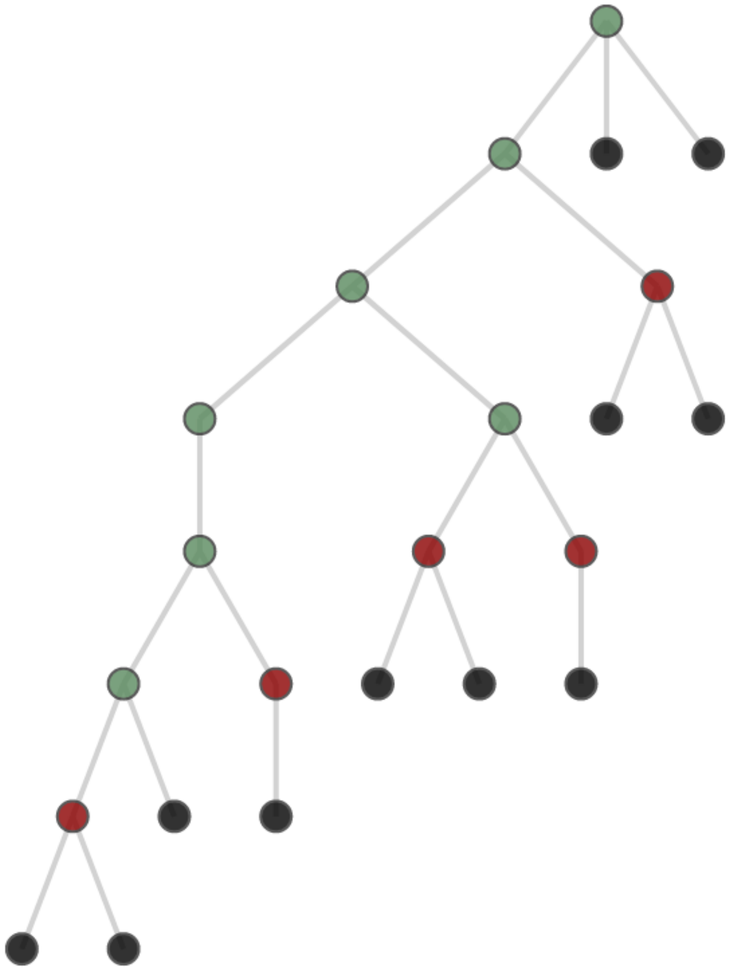
- Reordering and direct factorization of $L(\theta^0)(L(\theta^0))^T = Q(\theta^0)$
  - ✤ Known sparsity pattern, known first factor $L(\theta^0)$
- Use $L(\theta^{i-1})$ to compute $L(\theta^i) = \text{chol}(Q(\theta^i))$
  - ✤ $L^{(0)}(\theta^i) := L(\theta^{i-1})$

- Guaranteed Convergence: In each iteration at least one new correct entry
- All updates can be computed in parallel
- Similar to *Fine-grained parallel incomplete LU factorization* [Chow & Patel '15]

Iterative Cholesky Factorization

Implementation

- LDL^T factorization for numeric stability
- 1st Iteration: direct factorization
- Reorder
  → obtain independent supernodes
- Multi-threading
- Large matrices: no complete parallel update
  → architecture dependent
  → sequential : direct solver
- Different strategies for updating
  → start with leaves of elimination tree



Iterative Cholesky Factorization
Application

- Spatio-temporal model
- Optimization problem
  - ✤ BFGS-Iteration
- Recurrent factorization sparse s.p.d matrix $Q(\theta)$
  - ✤ $n \approx 160\,000$
  - ✤ nnz $\approx 1 \cdot 10^7$

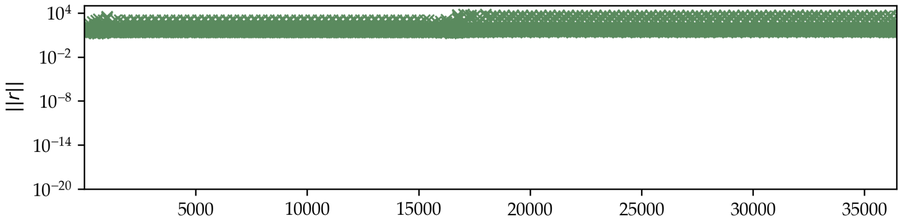
	DIRECT	**ITERATIVE**
BFGS Iter 1 ✤ compute $f(\theta^1)$	direct factorization $L_1 L_1^T = Q(\theta^1)$	direct factorization $L_1 L_1^T = Q(\theta^1)$
BFGS Iter 2 ✤ compute $f(\theta^2)$	direct factorization $L_2 L_2^T = Q(\theta^2)$	iteratively compute $L_2 = \text{iter}(Q(\theta^2), L_1)$
BFGS Iter 3 ✤ compute $f(\theta^3)$	direct factorization $L_3 L_3^T = Q(\theta^3)$	iteratively compute $L_3 = \text{iter}(Q(\theta^3), L_2)$

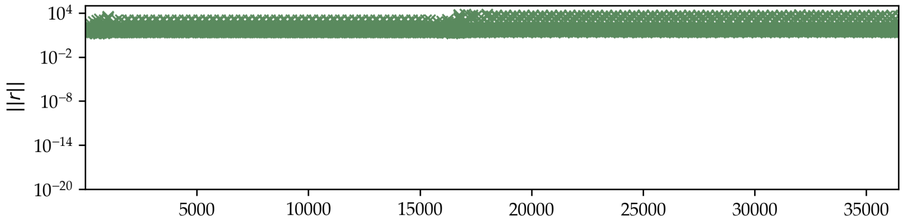
Update Strategy

BFGS-Iteration 5 — 1024 threads

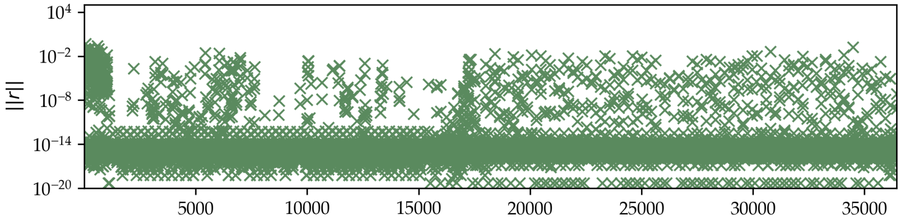
dynamic *etree*

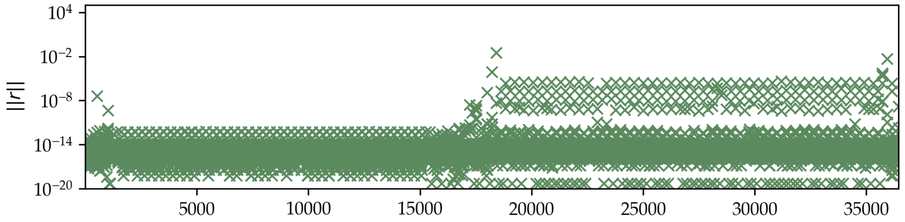
Cholesky-Iter 0



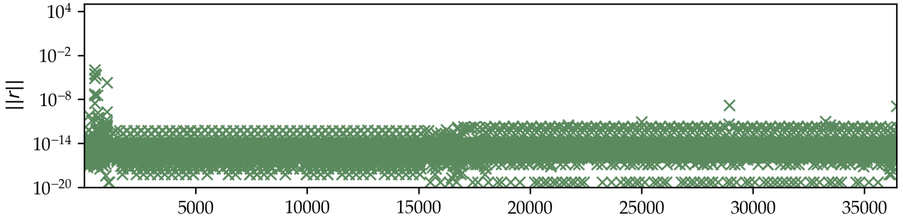


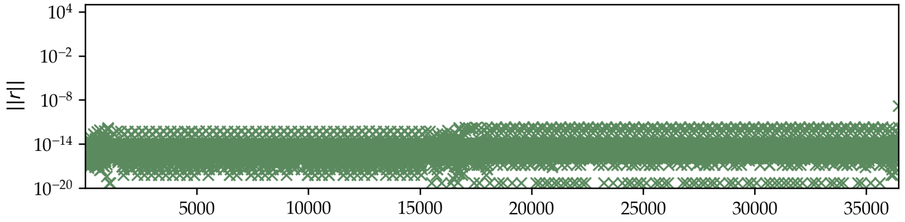
Cholesky-Iter 1



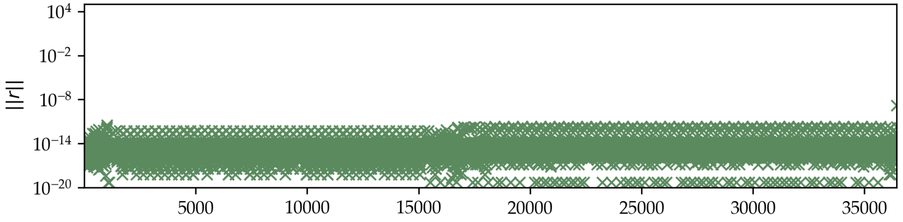


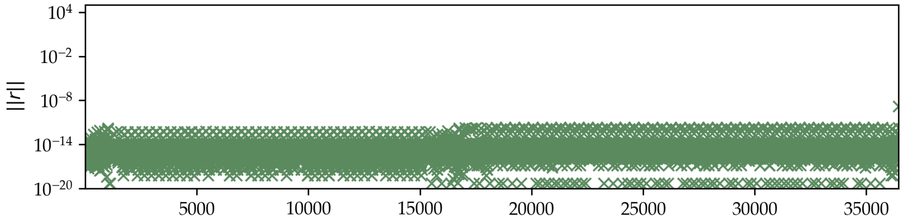
Cholesky-Iter 2





Cholesky-Iter 3



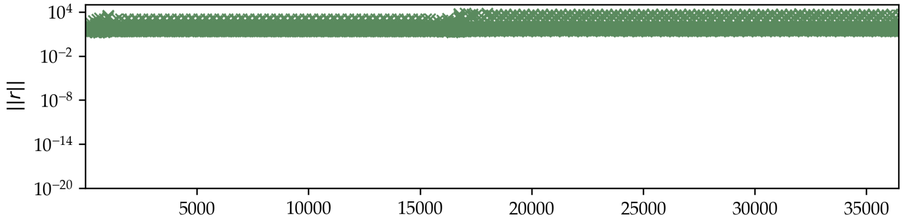


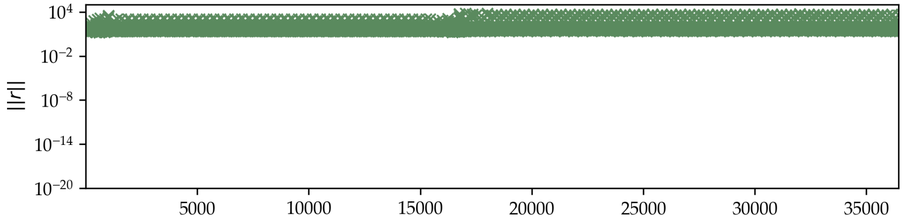
Number of Threads

BFGS-Iteration 5 — "etree" updating strategy

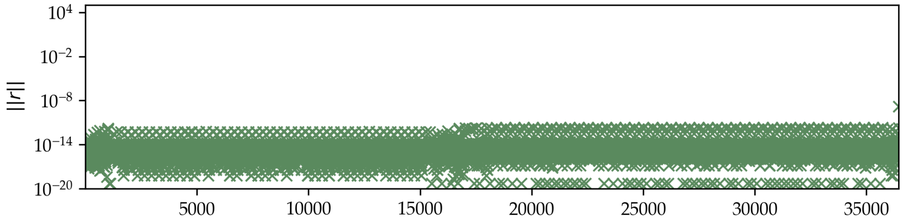
32 threads *1024 threads*

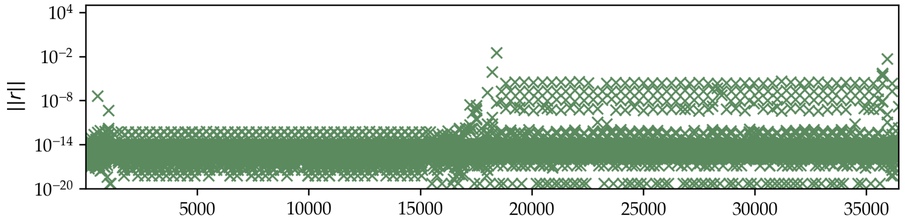
Cholesky-Iter 0



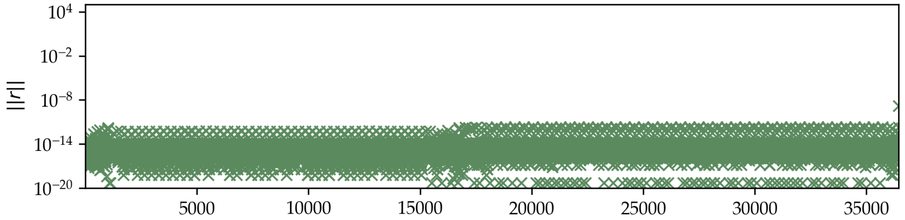


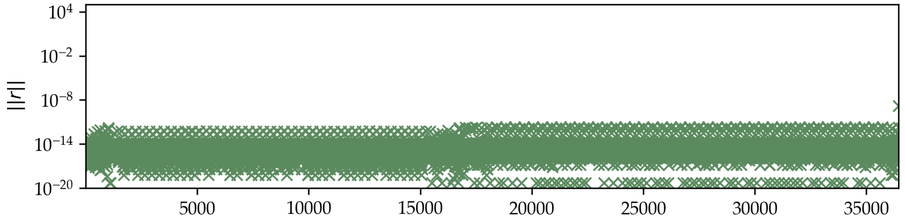
Cholesky-Iter 1



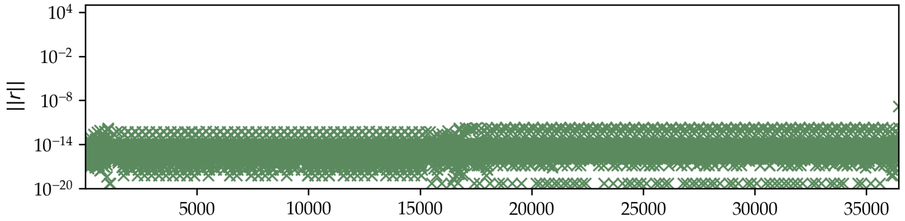


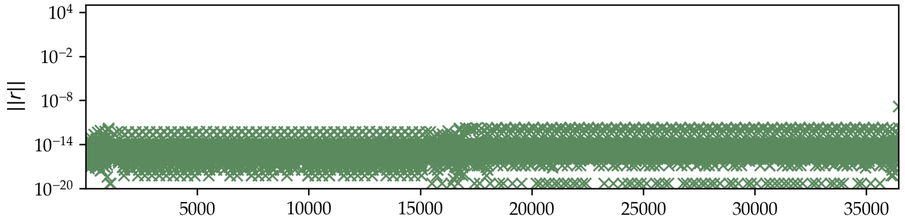
Cholesky-Iter 2





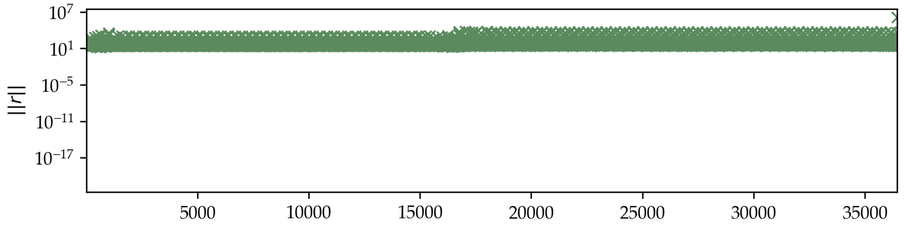
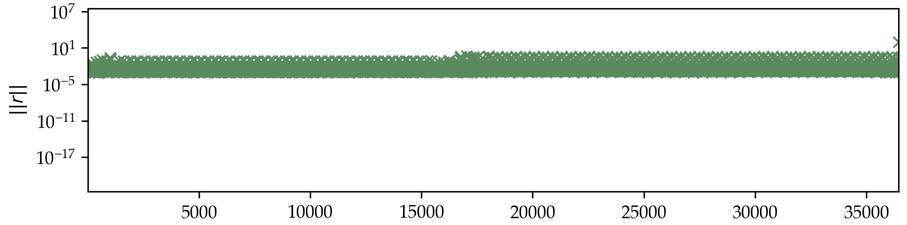
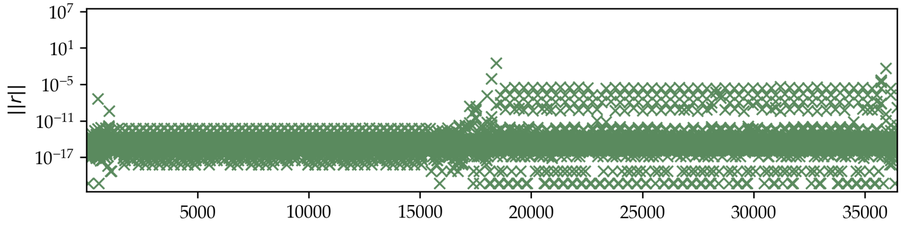
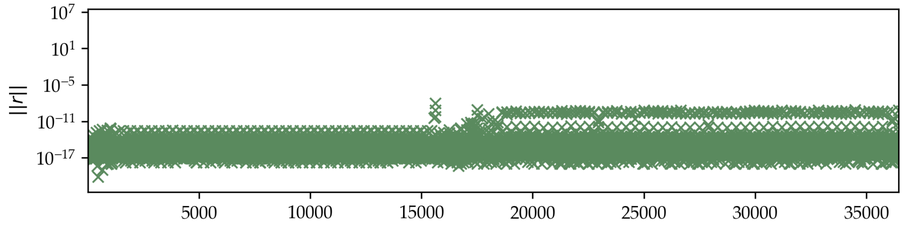
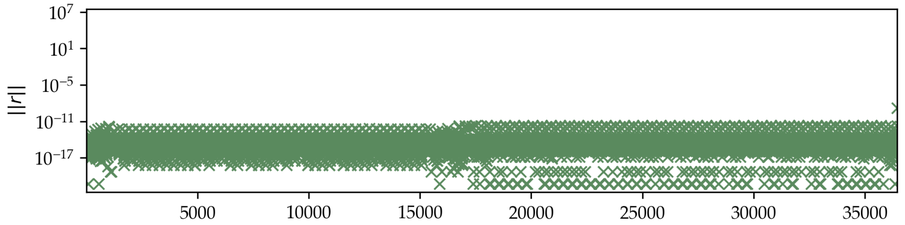
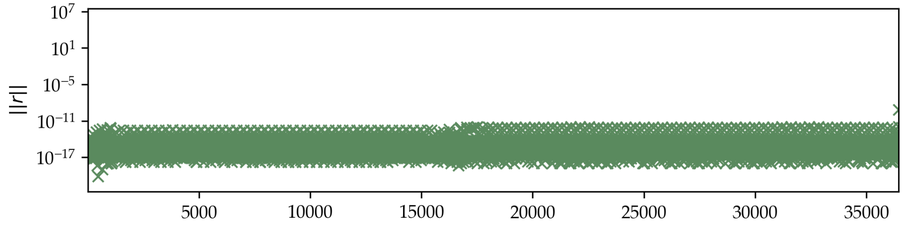
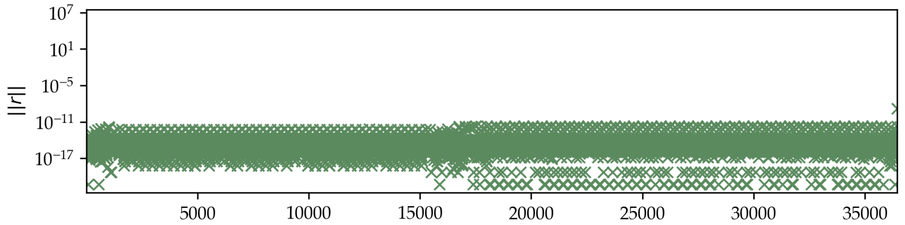
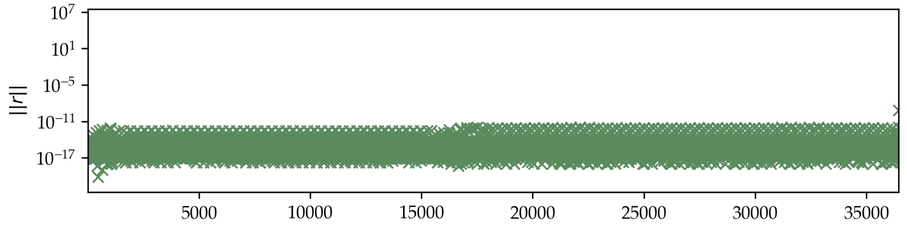
Cholesky-Iter 3





Accuracy Initial Guess

1024 threads — "etree" updating strategy

	BFGS-Iter 5	*BFGS-Iter 43*
Cholesky-Iter 0		
Cholesky-Iter 1		
Cholesky-Iter 2		
Cholesky-Iter 3		

Summary

- Large-scale spatio-temporal Bayesian inference
  - ✤ Integrated Nested Laplace Approximations
- Sparse Linear Algebra with recurring sparsity patterns
- Computational Bottleneck: Optimization Problem
  - ✤ Cholesky decomposition
  - ✤ Forward-Backward Substitution
  - ✤ Selected matrix inversion
- Direct sparse and direct block solvers

Ongoing Work

- Sequence of matrices to be factorized
- Iterative Cholesky Factorization
  - ✤ Massively parallel updates
  - ✤ Known Sparsity Structure
  - ✤ “Good” initial guess

G.-M. L, Krainski E, Janalik R, Rue H, Schenk O. **Integrated Nested Laplace Approximations for Large-Scale Spatial-Temporal Bayesian Modeling**. arXiv preprint arXiv:2303.15254. (2023).

G.-M. L, van Niekerk, J, Schenk O, Rue H. **Parallelized integrated nested Laplace approximations for fast Bayesian inference**. *Stat Comput* **33**, 25 (2023).

Thank you for your attention.

Questions?