Pathfinder.jl: approximate inference for the statistical workflow

Seth Axen[1] and Lu Zhang[2]
1. University of Tübingen 2. University of Southern California

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Summary

Goal Explore uses for Pathfinder in statistical workflow

Approach

- Benchmark 3 ways of initializing HMC with Pathfinder
- Benchmark Pathfinder variations

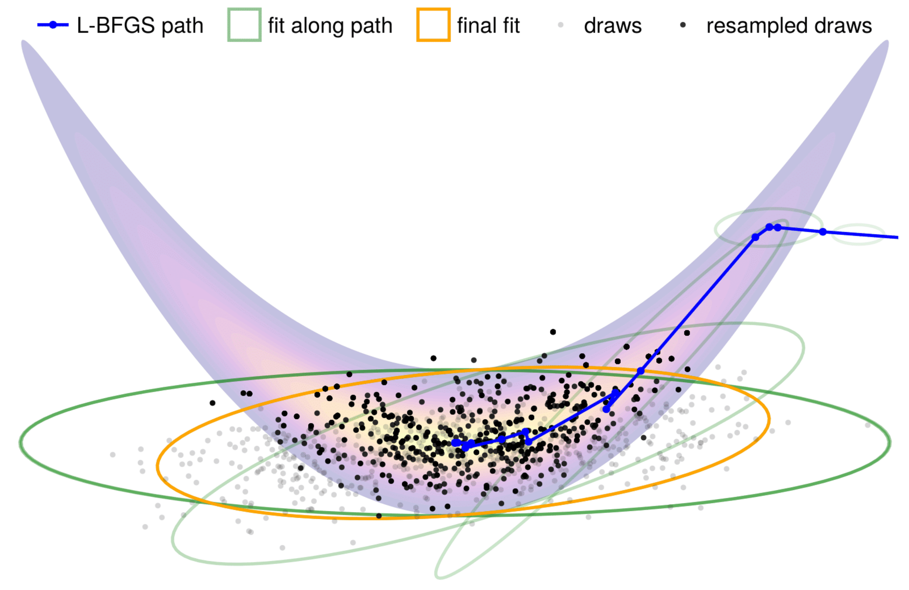
Contributions

- A Pathfinder implementation for the Julia ecosystem
- Pathfinder is not yet a general Stan warm-up replacement
- Hager-Zhang line search improves performance over Moré-Thuente
- Returned quantities are useful as model diagnostics

Pathfinder: quasi-Newton variational inference

Algorithm [1]

1. Maximize log-density using L-BFGS
2. Construct sequence of multivariate normal variational approximations using inverse Hessian approximations Σ
3. Find approximation with highest ELBO estimate
4. (single-Pathfinder) Draw from variational approximation
5. (multi-Pathfinder) Run 1-4 in parallel and importance resample draws from runs



Example execution of the Pathfinder algorithm. 95% ellipses of approximations are shown.

Pathfinder.jl

Integrations

- Optimization.jl: replace L-BFGS with any optimizer
- LogDensityProblems.jl: support any PPL with a log-density function (Turing, Stan, etc)
- Transducers.jl: parallelize across multiple threads or cores
- Distributions.jl/PDMats.jl: use the variational model anywhere in the ecosystem
- InferenceObjects.jl (planned): use the draws with any package that recognizes ArviZ InferenceData

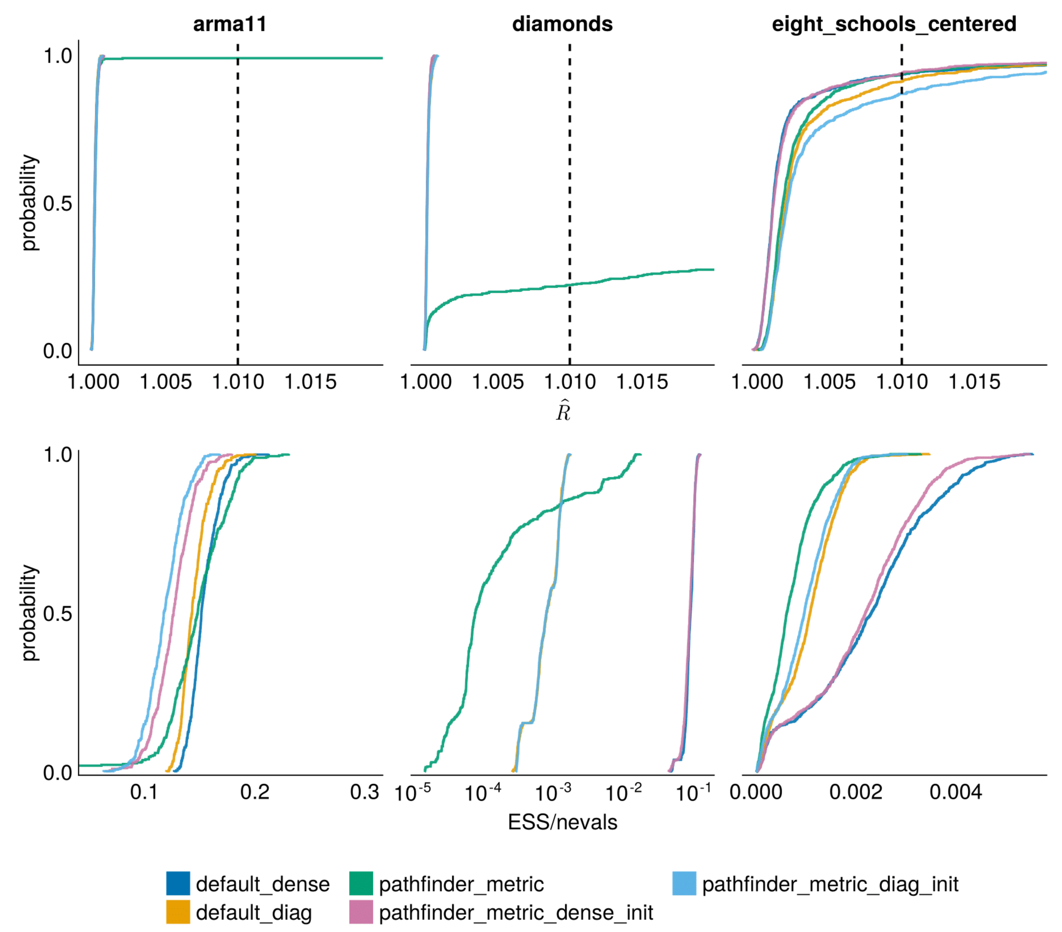
Initializing dynamic HMC with Pathfinder

Initialization strategies

1. `default_<metric>`: Stan tuning
2. Pathfinder point $\to$ Stan step-size/metric tuning *(not shown)*
3. `pathfinder_metric_<metric>_init`: Pathfinder point/metric $\to$ Stan step-size/metric tuning
4. `pathfinder_metric`: Pathfinder point/metric $\to$ Stan step-size tuning

Models from posteriordb

- `arma11` (4 parameters, simple)
- `diamonds` (26 parameters, high correlation)
- `eight_schools_centered` (10 parameters, funnel)



ECDF plots of convergence and performance diagnostics for HMC with different initialization strategies.

Variations of Pathfinder for dynamic HMC

Line search

Goal: With current point x_l and search direction p_l, maximize α_l where $x_{l+1} = x_l + \alpha_l p_l$ subject to conditions

Moré-Thuente (default) Wolfe conditions

Hager-Zhang[2] Approximate Wolfe conditions

Line search initialization

Static (default) Initial step size guess $\alpha_l = 1$

Static scaled $\alpha_l = \min\left(1, \|p_l\|^{-1}\right)$

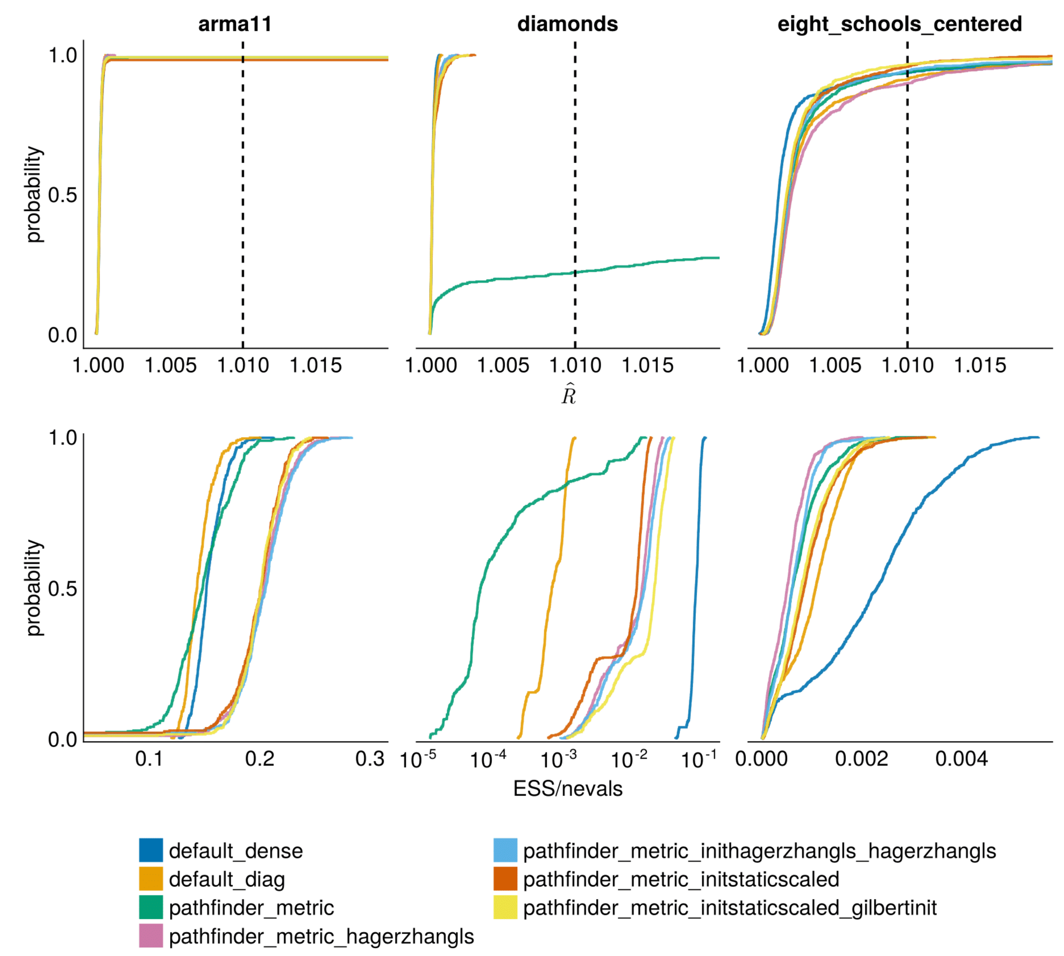
Hager-Zhang[2]

Preconditioning inverse Hessian

Initialize $\Sigma_l = \operatorname{diag}(b_l)$ using $s_l = x_l - x_{l-1}$ and $y_l = \nabla \log p(x_l) - \nabla \log p(x_{l-1})$

Nocedal-Wright (default) $(b_l)_i = \frac{\langle y_l, s_l \rangle}{\|y_l\|^2}$

Gilbert[3] $(b_l)_i = \frac{\langle y_l, s_l \rangle}{h((y_l)_i, (s_l)_i, (b_{l-1})_i)}$



ECDF plots of convergence and performance diagnostics for HMC using the initial point and metric from variations of Pathfinder.

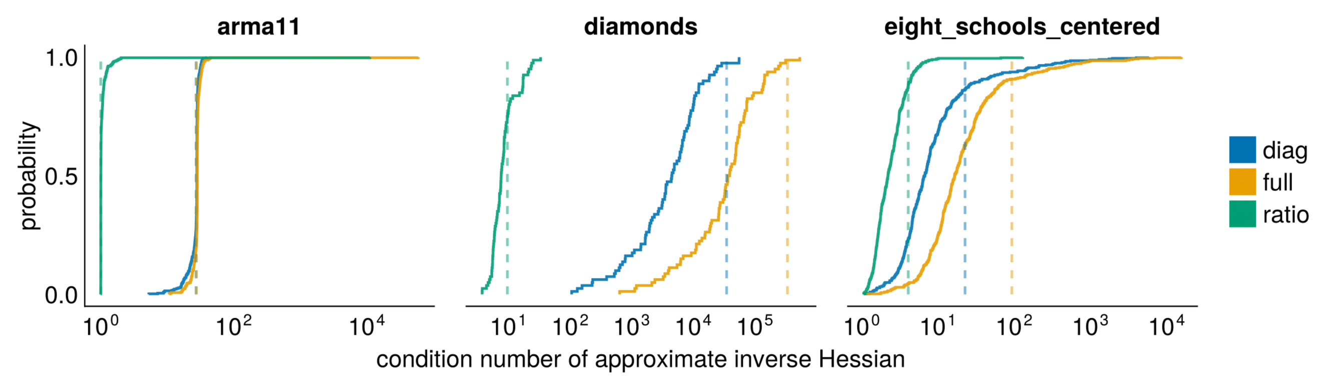
Diagnosing computational issues

Condition numbers

With $\lambda = \operatorname{eigvals}(\Sigma)$, $\operatorname{cond}(\Sigma) = \frac{\max_i |\lambda_i|}{\min_i |\lambda_i|}$

"Do the parameters have different variances?": $\operatorname{cond}(\Sigma \circ I)$

"Is a dense metric better than diagonal?": $\frac{\operatorname{cond}(\Sigma)}{\operatorname{cond}(\Sigma \circ I)}$

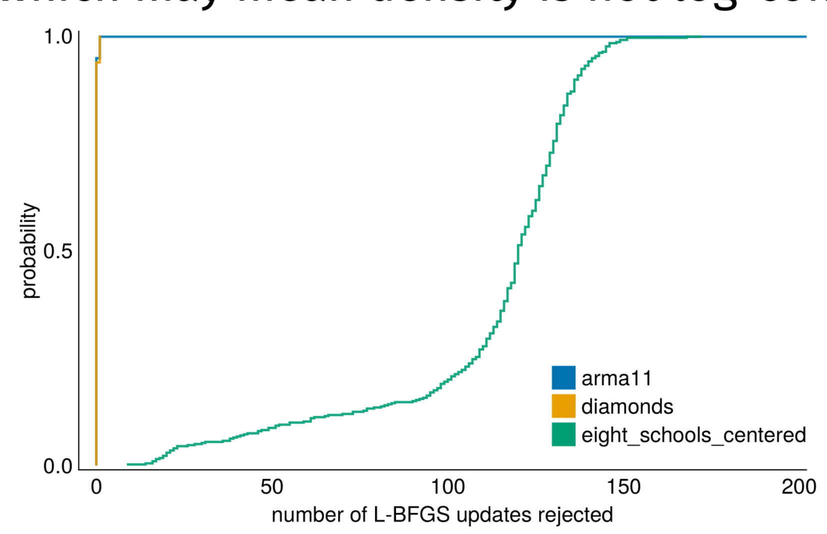


ECDF plots of condition numbers and their ratios compared with that of the reference posterior covariance (dash)

Number of rejected BFGS updates

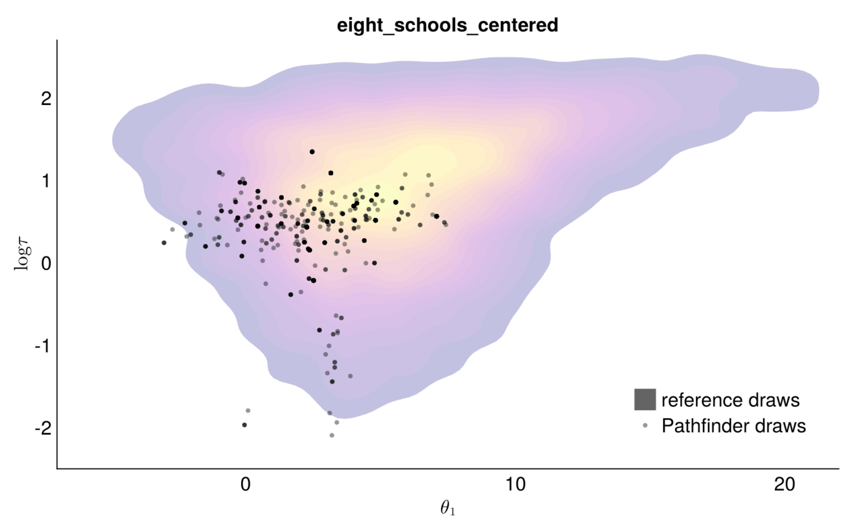
"Should the model be reparameterized?"

BFGS updates are rejected to keep Σ positive definite, which may mean density is not log-concave.



Plot importance resampled draws

"Are there any odd features?"



Pathfinder draws indicate the funnel geometry apparent in the reference posterior draws.

References

[1] L Zhang, B Carpenter, A Gelman, A Vehtari. (2021). *Pathfinder: Parallel quasi-Newton variational inference.* 2021. arXiv:2108.03782

[2] WW Hager and H Zhang. (2006). *Algorithm 851: CG_DESCENT, a conjugate gradient method with guaranteed descent.* TOMS, 32(1): 113–137.

[3] JC Gilbert, C Lemaréchal. (1989). *Some numerical experiments with variable-storage quasi-Newton algorithms.* Math. Program. 45, 407–435