From viral evolution to spatial contagion: a biologically modulated Hawkes model

Andrew J. Holbrook

UCLA Biostatistics

ICMAT-CSIC 17.6.2021

Brazil is facing a spike in Covid-19 deaths and an overwhelmed health care system. A more contagious variant of the virus may be part of the problem.
Dado Galdieri for The New York Times

Opinion

How to Protect Yourself Against Coronavirus Variants

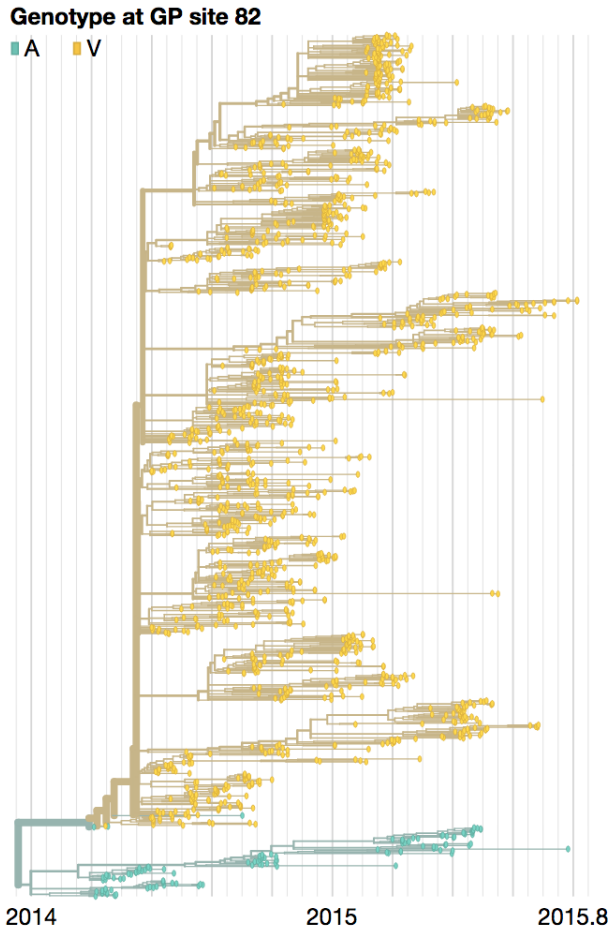
Upgrading your mask and staying vigilant are more important than ever.

By Abraar Karan

Dr. Karan is an internal medicine physician at the Brigham and Women's Hospital and Harvard Medical School who worked on the Massachusetts government response to Covid-19 last year.

March 3, 2021, 5:00 a.m. ET

A single mutation swept through the Ebola virus population in West Africa



Some questions

Question 1. How should we characterize, quantify and estimate rates of viral spread?

Some questions

Question 1. How should we characterize, quantify and estimate rates of viral spread?

Question 2. Can we develop a single model that relates genetic changes to enhanced spatial contagion?

Some questions

Question 1. How should we characterize, quantify and estimate rates of viral spread?

Question 2. Can we develop a single model that relates genetic changes to enhanced spatial contagion?

Question 3. Can we leverage both sequenced and unsequenced cases?

Some questions

Question 1. How should we characterize, quantify and estimate rates of viral spread?

Question 2. Can we develop a single model that relates genetic changes to enhanced spatial contagion?

Question 3. Can we leverage both sequenced and unsequenced cases?

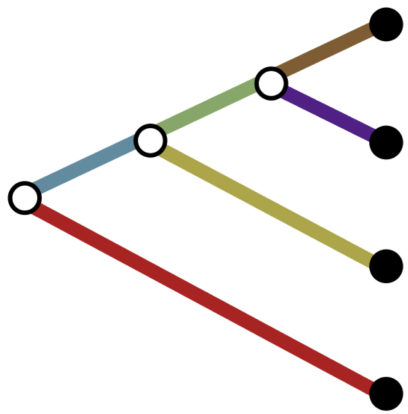
Question 4. How might we quantify our uncertainty?

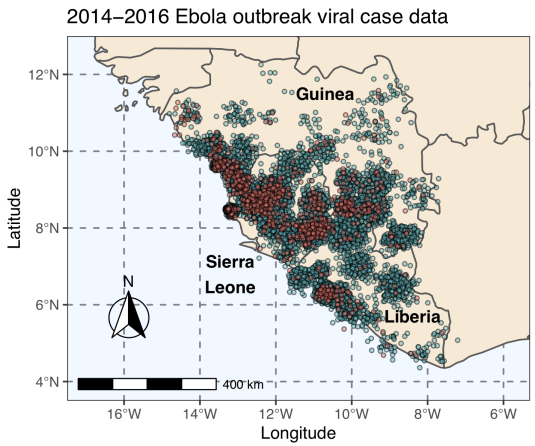
Part 1. Phylogenetic Hawkes process

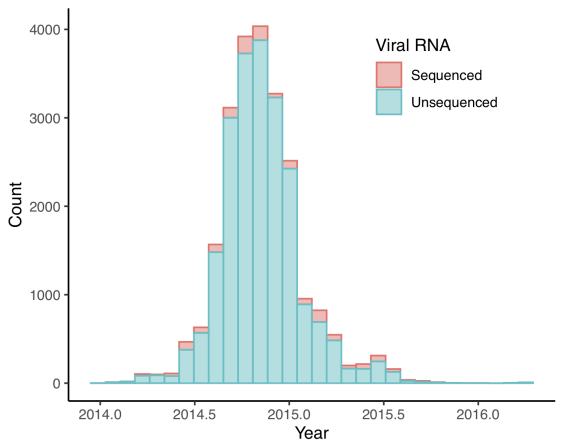
The challenge

You are given

1. spatiotemporal coordinates (x_n, t_n) for $n = 1, \ldots, N$ viruses;
2. the evolutionary history of a small subset of viruses.







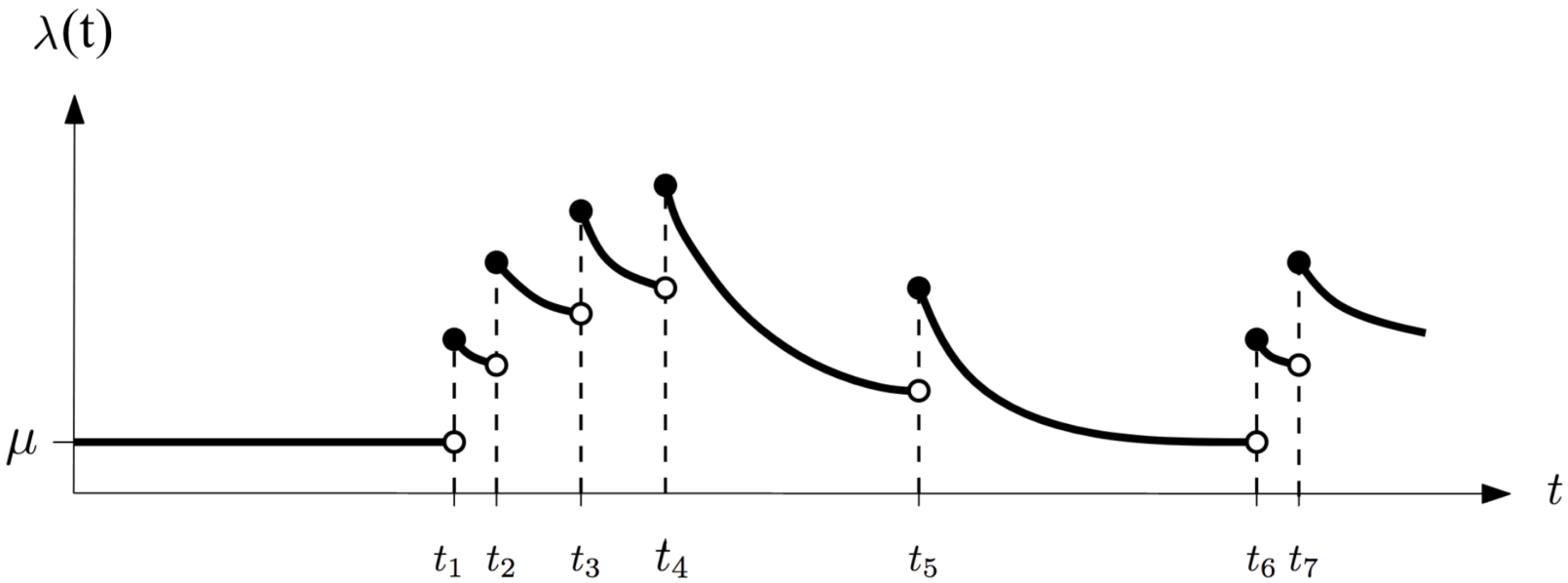
You are tasked to

1. use this data to discover whether certain branches have greater contagiousness;
2. quantify your uncertainty with respect to the relevant quantity.

Two paradigms

	Traditional Bayesian phylogenetics	Hawkes processes
Observational limit	N in low thousands	N in high tens-of-thousands
Biological insight	Evolutionary history	None
Genetic sequencing	Required	Not required
Spatiotemporal data	Not required	Required
Geographic spread	Not modeled	Modeled
Large-scale transport	Does not induce bias	Induces bias

Hawkes process



Laub et al. 2015

$$\lambda(t) = \mu + \xi(t) = \mu + \sum_{t_n < t} g(t - t_n)$$

Spatiotemporal Hawkes process

$$\lambda(\mathrm{x}, t) = \mu(\mathrm{x}) + \xi(\mathrm{x}, t) = \mu(\mathrm{x}) + \sum_{t_n < t} g(\mathrm{x} - \mathrm{x}_n, t - t_n)$$

Variable degrees of contagion

One can tailor the triggering function to change for each observation (Schoenberg et al., 2019):

$$\lambda(\mathrm{x}, t) = \mu(\mathrm{x}) + \sum_{t_n < t} g_n(\mathrm{x} - \mathrm{x}_n, t - t_n)\,.$$

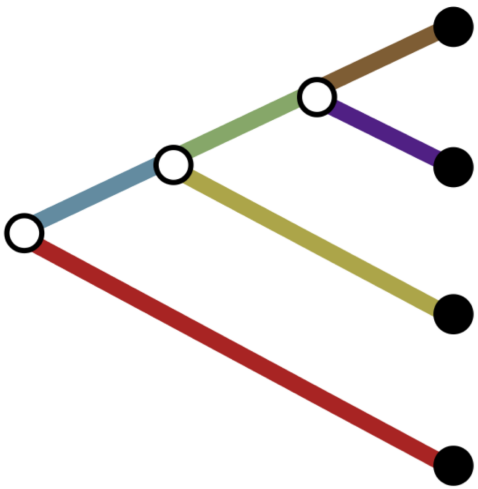
In the following, I specify

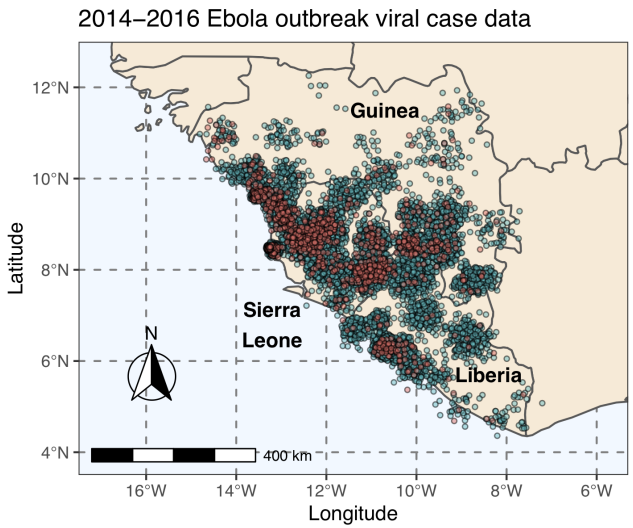
$$\mu(\mathrm{x}) = \frac{\mu_0}{\tau_x^D} \sum_{n=1}^{N} \phi\left(\frac{\mathrm{x} - \mathrm{x}_n}{\tau_x}\right) \mathcal{I}_{[x \neq x_n]}$$

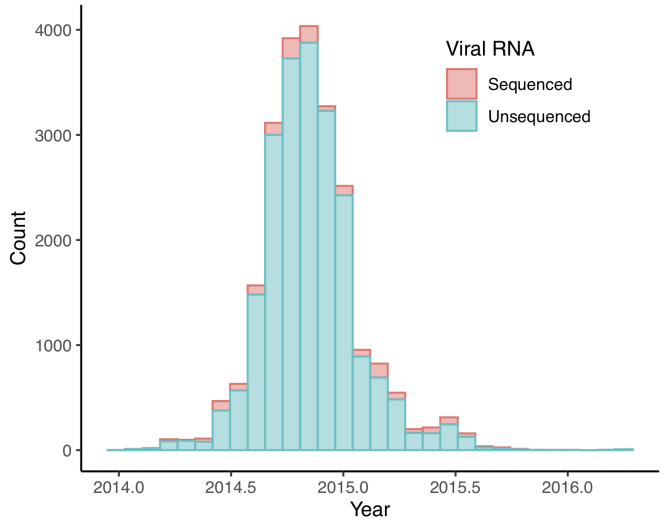
and

$$\xi(\mathrm{x}, t) = \frac{\theta_0 \omega}{h^D} \sum_{t_n < t} \theta_n \, e^{-\omega\,(t - t_n)} \phi\left(\frac{\mathrm{x} - \mathrm{x}_n}{h}\right)\,.$$

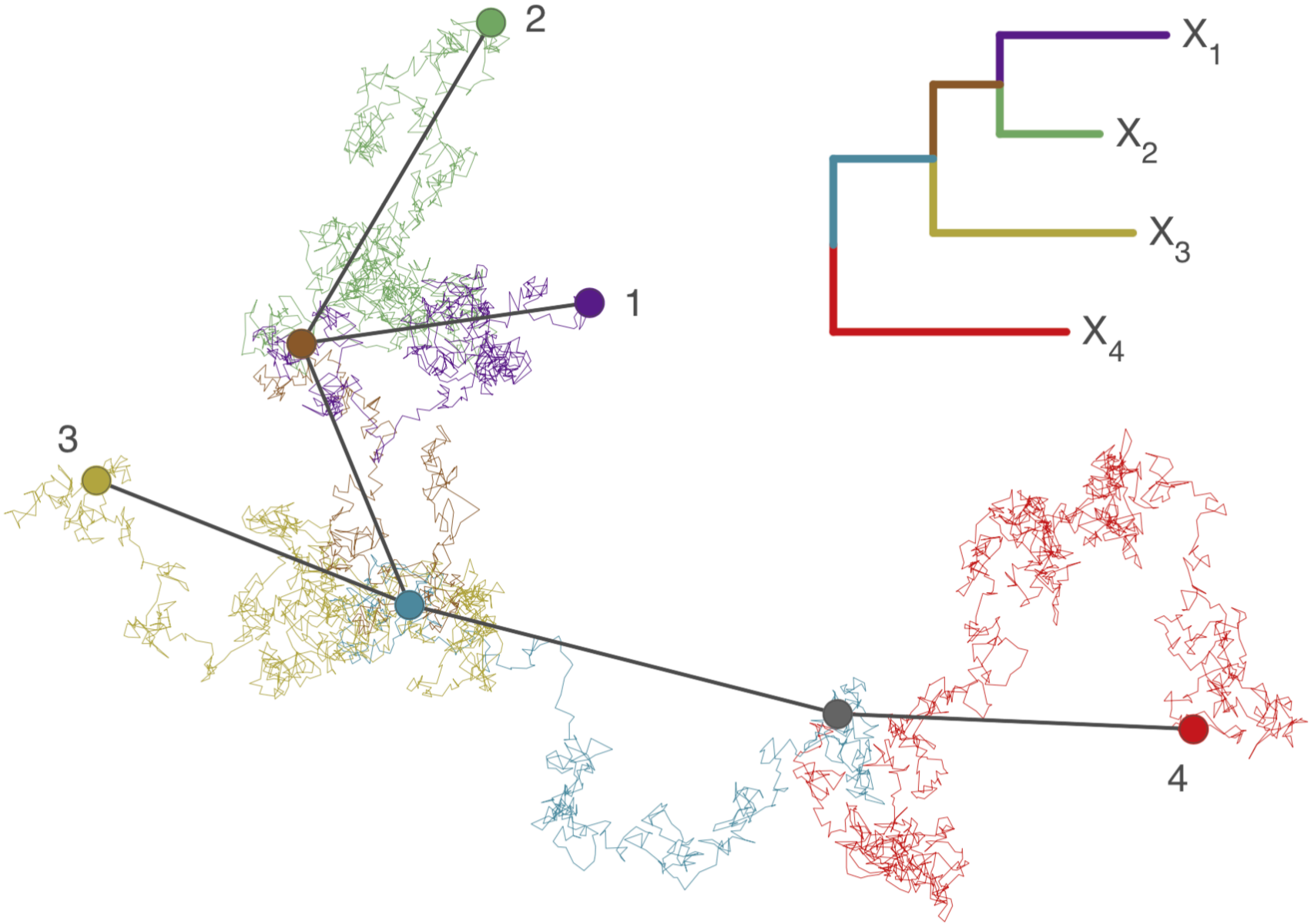
The challenge







Brownian phylogenetic diffusion



Brownian phylogenetic diffusion

Associate to each tip n of a rooted, M-tipped, binary tree a Brownian motion z_n, centered at its parent node $z_{pa(n)}$. Then

$$z_n | z_{pa(n)} \sim Normal(z_{pa(n)}, t_n \sigma^2) ,$$

for t_n the branch length of node n to its parent.

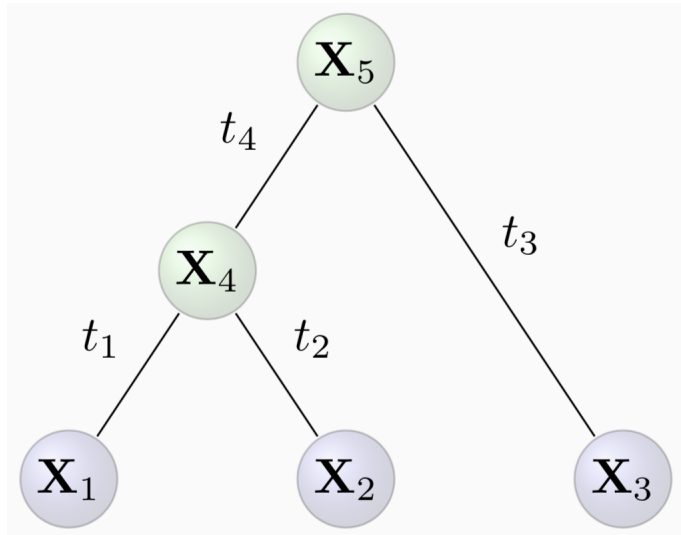
Write the joint distribution as

$$\mathrm{z} \sim Normal_M(0, \sigma^2 \mathrm{V})$$

for

$$[\mathrm{V}]_n = t_n + t_{pa(n)} + t_{pa(pa(n))+\ldots}$$

$$[\mathrm{V}]_{nn'} = \begin{cases} [\mathrm{V}]_n - t_n, & pa(n) = pa(n') \\ 0, & o/w \end{cases}$$



Zhang et al. 2019

Log link on productivities

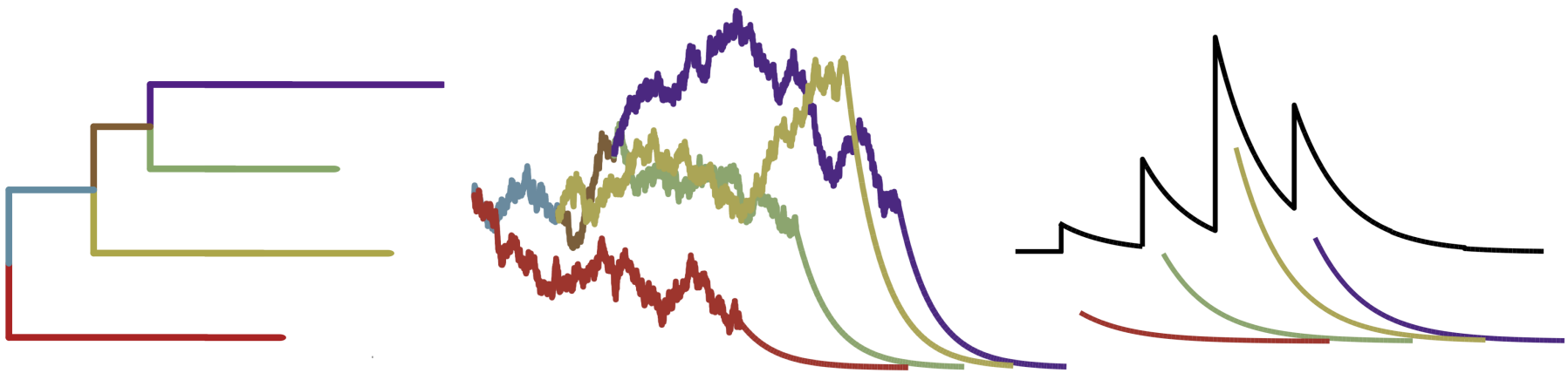
Recall our self-exciting rate function:

$$\xi(\mathrm{x}, t) = \frac{\theta_0 \omega}{h^D} \sum_{t_n < t} \theta_n \, e^{-\omega\,(t - t_n)} \phi\left(\frac{\mathrm{x} - \mathrm{x}_n}{h}\right) ,$$

and define

$$\begin{cases} \theta_n = \theta_n(z_n) = \exp\left(z_n + \beta t_n\right) & z_n \in \mathbb{R}\,, \quad n \in \mathcal{M} \\ \theta_n = 1 & n \notin \mathcal{M}\,. \end{cases}$$

Phylogenetic Hawkes process



Part 2. Bayesian inference

Likelihood based inference and Bayes

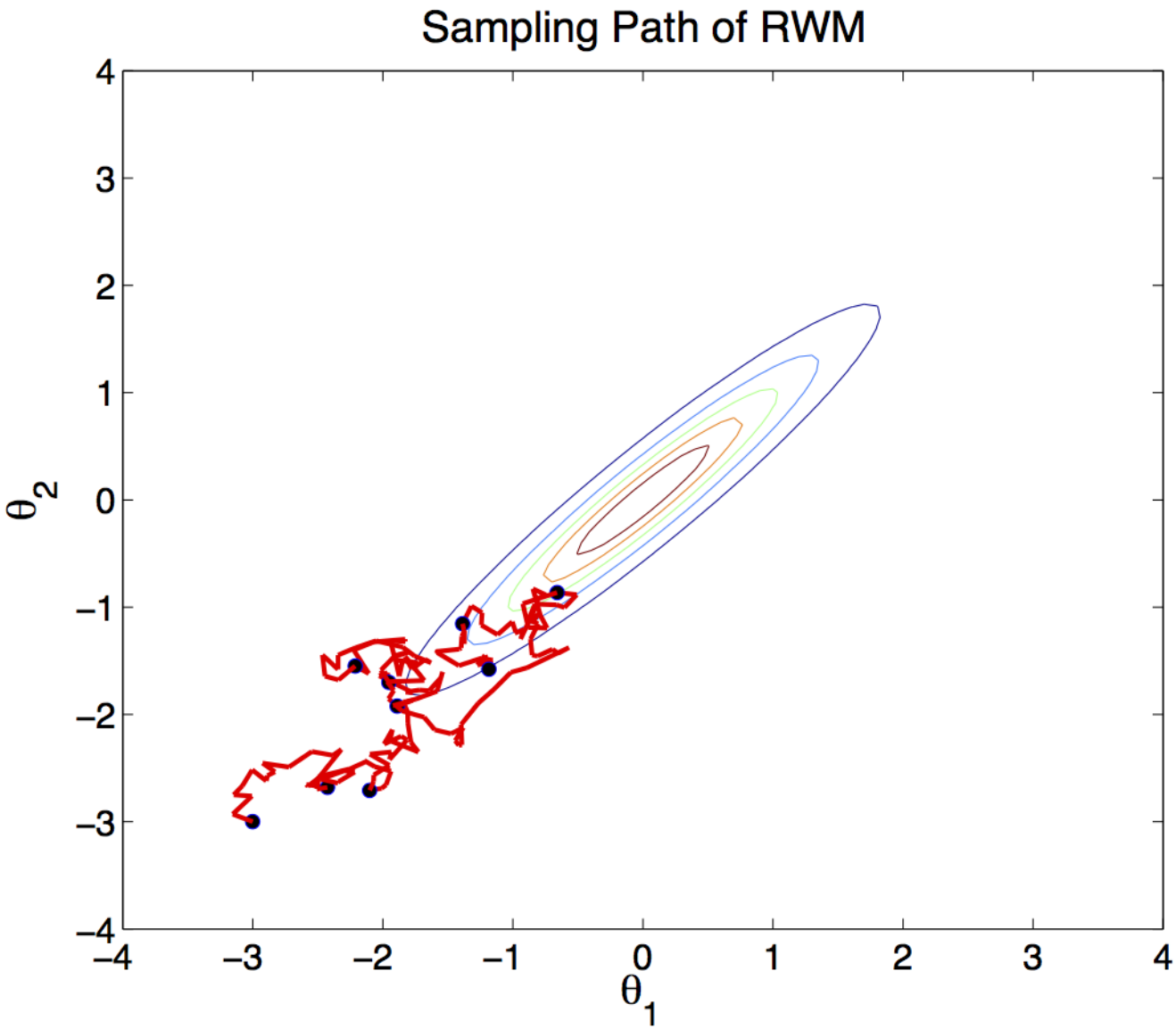
Assume data generated according to $\mathrm{y}_n \overset{\perp}{\sim} f(\mathrm{y}_n|\boldsymbol{\theta}, \mathrm{z}_n)$ with prior distributions $\boldsymbol{\theta} \sim p_\theta(\boldsymbol{\theta})$ and $(\mathrm{z}_1, \ldots, \mathrm{z}_N) = \mathrm{Z} \sim p_z(\mathrm{Z})$.

Bayes' theorem says:

$$p(\boldsymbol{\theta}|\mathrm{Y}) = \frac{f(\mathrm{Y}|\boldsymbol{\theta})\, p_\theta(\boldsymbol{\theta})}{f(\mathrm{Y})} = \frac{\int_{\mathrm{Z}} f(\mathrm{Y}|\mathrm{Z}, \boldsymbol{\theta}) p_z(\mathrm{Z})\mathrm{dZ}\; p_\theta(\boldsymbol{\theta})}{\int_\Theta \left(\int_{\mathrm{Z}} f(\mathrm{Y}|\mathrm{Z}, \boldsymbol{\theta}) p_x(\mathrm{Z})\mathrm{dZ}\right) p_\theta(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}},$$

where $f(\mathrm{Y}|\boldsymbol{\theta}, \mathrm{Z}) = \prod_n^N f(\mathrm{y}_n|\boldsymbol{\theta}, \mathrm{z}_n)$ is the *likelihood* function and $f(\mathrm{Y}|\boldsymbol{\theta})$ is the *marginal likelihood*.

Random walk Metropolis



RWM requires likelihood evaluations

Our Hawkes process likelihood scales $\mathrm{O}(N^2)$:

$$\ell(\mathrm{X}, \mathrm{t} | \mu_0, \tau_x, \theta_0, \boldsymbol{\theta}, \omega, h) = -\Lambda(t_N) + \sum_{n=1}^{N} \log \lambda_n$$

$$= \sum_{n=1}^{N} \left\{ \log \left[\sum_{n'=1}^{N} \left(\frac{\mu_0 \, \mathcal{I}_{[\mathrm{x}_n \neq \mathrm{x}_{n'}]}}{\tau_x^D} \phi \left(\frac{\mathrm{x}_n - \mathrm{x}_{n'}}{\tau_x} \right) + \frac{\theta_0 \theta_{n'} \, \omega \, \mathcal{I}_{[t_{n'} < t_n]}}{h^D} e^{-\omega \, (t_n - t_{n'})} \phi \left(\frac{\mathrm{x}_n - \mathrm{x}_{n'}}{h} \right) \right) \right] - \Lambda_n \right\}$$

$$:= \sum_{n=1}^{N} \left[\log \left(\sum_{n'=1}^{N} \lambda_{nn'} \right) - \Lambda_n \right] .$$

Statistics and Computing (2021) 31:4
https://doi.org/10.1007/s11222-020-09980-4

Scalable Bayesian inference for self-excitatory stochastic processes applied to big American gunfire data

Andrew J. Holbrook[1] · **Charles E. Loeffler**[2] · **Seth R. Flaxman**[3] · **Marc A. Suchard**[1,4,5]

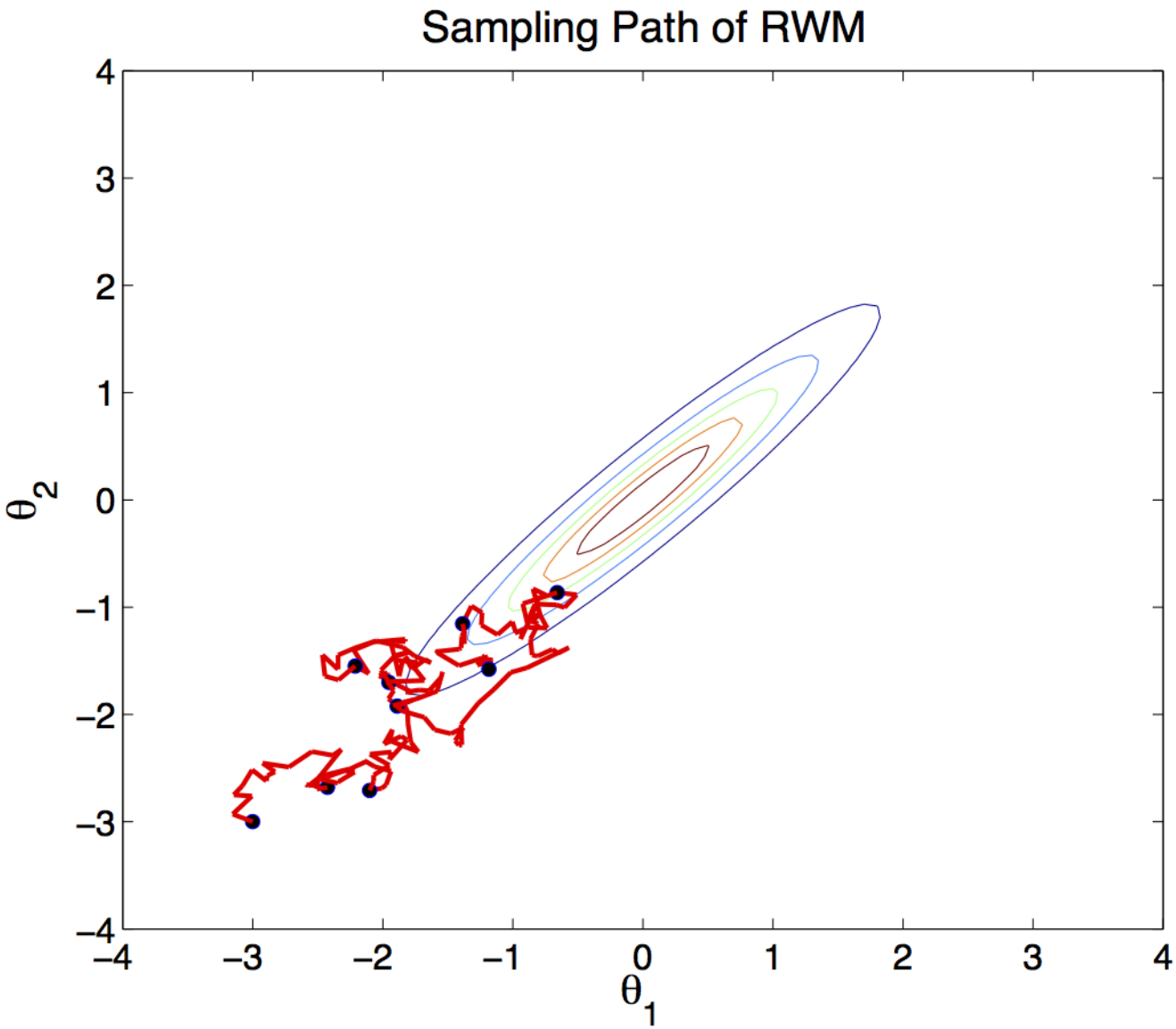
Received: 13 May 2020 / Accepted: 2 December 2020

Abstract

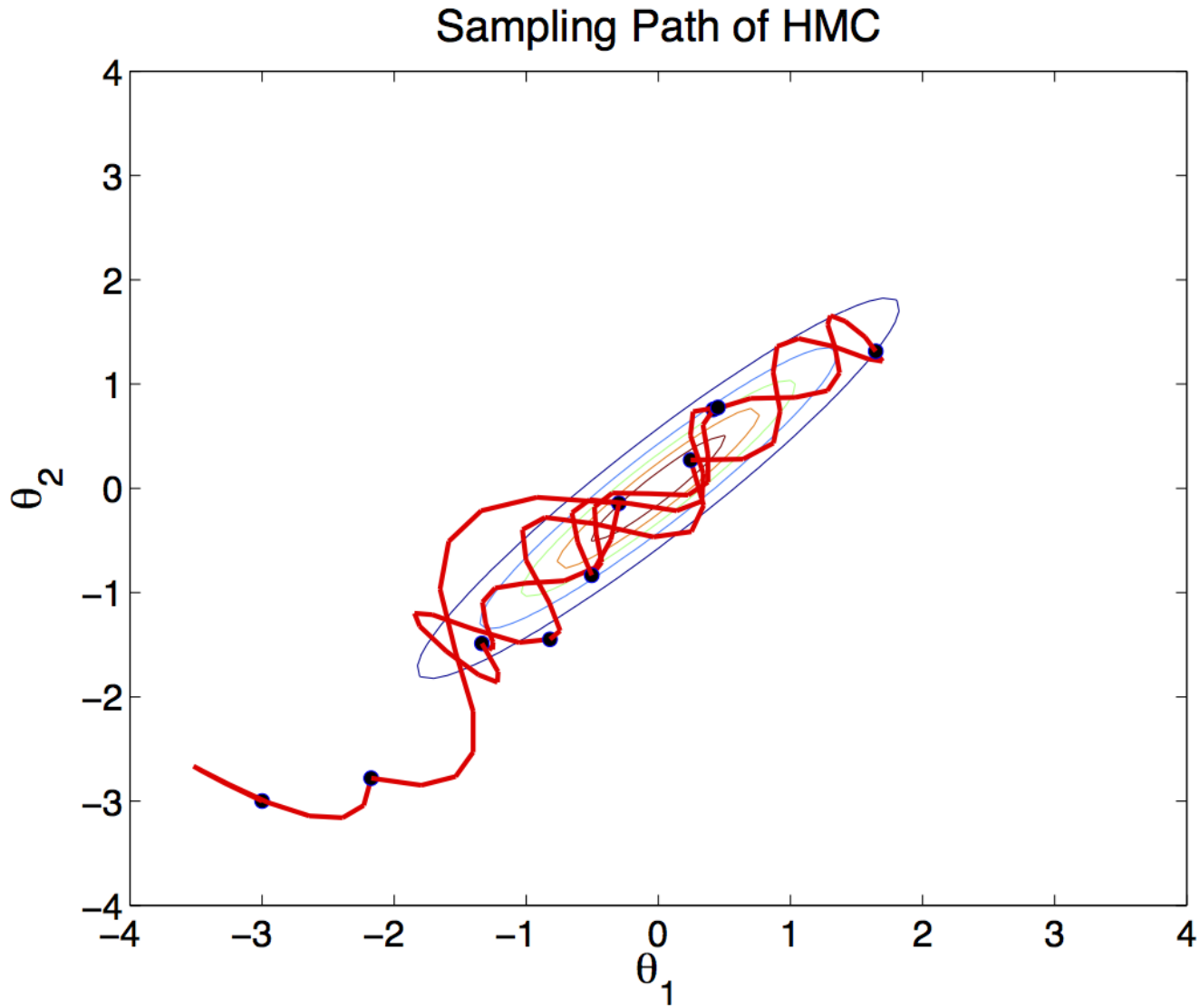
The Hawkes process and its extensions effectively model self-excitatory phenomena including earthquakes, viral pandemics, financial transactions, neural spike trains and the spread of memes through social networks. The usefulness of these stochastic process models within a host of economic sectors and scientific disciplines is undercut by the processes' computational burden: complexity of likelihood evaluations grows quadratically in the number of observations for both the temporal and spatiotemporal Hawkes processes. We show that, with care, one may parallelize these calculations using both central and graphics processing unit implementations to achieve over 100-fold speedups over single-core processing. Using a simple adaptive Metropolis–Hastings scheme, we apply our high-performance computing framework to a Bayesian analysis of big gunshot data generated in Washington D.C. between the years of 2006 and 2019, thereby extending a past analysis of the same data from under 10,000 to over 85,000 observations. To encourage widespread use, we provide HPHAWKES, an open-source R package, and discuss high-level implementation and program design for leveraging aspects of computational hardware that become necessary in a big data setting.

Keywords Massive parallelization · GPU · SIMD · Spatiotemporal Hawkes process

Random walk Metropolis



Hamiltonian Monte Carlo



Hamiltonian Monte Carlo

Augment parameter space with auxiliary Gaussian variable p and construct a Hamiltonian energy function:

$$\begin{aligned} H(\mathsf{z}, \mathsf{p}) &= -\log(\pi(\mathsf{z}) \times \phi(\mathsf{p})) \\ &\propto -\log \pi(\mathsf{z}) + \frac{1}{2}\mathsf{p}^T\mathsf{p}\,. \end{aligned}$$

New states of the Markov chain are proposed by forward integrating Hamilton's equations:

$$\begin{aligned} \frac{\mathrm{d}\mathsf{z}}{\mathrm{d}t} &= \frac{\partial H}{\partial \mathsf{p}} = \mathsf{p} \\ \frac{\mathrm{d}\mathsf{p}}{\mathrm{d}t} &= -\frac{\partial H}{\partial \mathsf{z}} = \nabla \log \pi(\mathsf{z})\,. \end{aligned}$$

Numerical simulation induces discretization error, which we correct with a Metropolis accept-reject step.

Hamiltonian Monte Carlo

Benefits. HMC computes high-dimensional integrals;
scales to 30,000+ parameters.

Challenges. HMC necessitates repeated computation of
log-likelihood and its gradient (best case $\mathcal{O}(N)$).

HMC for variable rates?

- The Hawkes likelihood scales $\mathcal{O}(N^2)$ ✓
- But the Hawkes log-likelihood gradient also scales $\mathcal{O}(N^2)$

$$\frac{\partial \ell}{\partial \theta_n} = -\frac{\partial \Lambda_n}{\partial \theta_n} + \sum_{t_n < t_{n'}} \frac{1}{\lambda_{n'}} \frac{\partial \lambda_{n'n}}{\partial \theta_n}$$

$$= \theta_0 \left(e^{-\omega\,(t_N - t_n)} - 1 \right) + \sum_{t_n < t_{n'}} \frac{1}{\lambda_{n'}} \frac{\theta_0 \omega}{h^D} e^{-\omega\,(t_{n'} - t_n)} \phi \left(\frac{x_{n'} - x_n}{h} \right)$$

Parallelization tools

Central processing unit (CPU):

1. Global parallelization: 2 to 60 cores (multi-core)
2. Local parallelization: single instruction multiple data (SIMD)

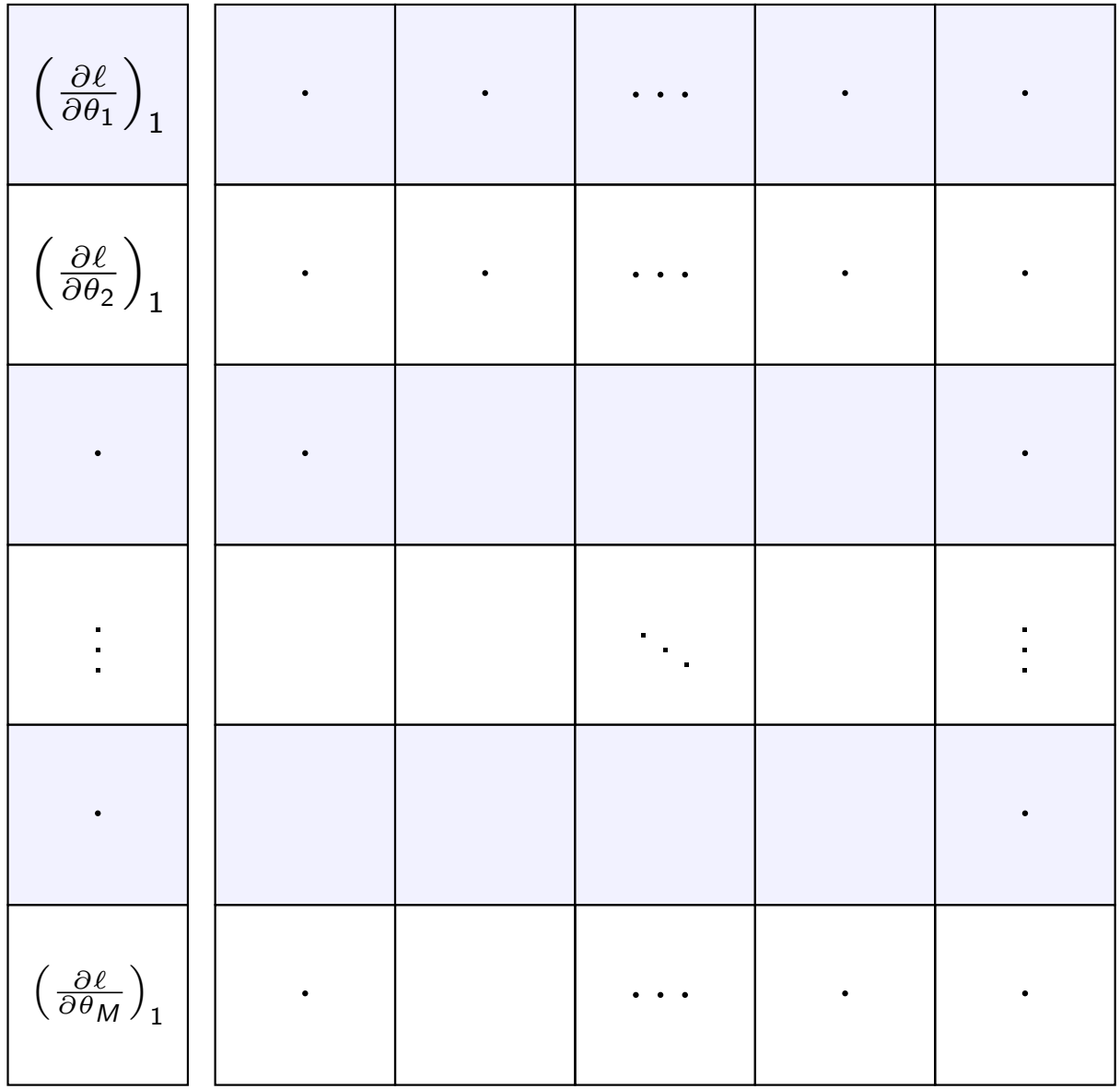
Graphics processing unit (GPU):

1. Thousands of cores (many-core)
2. Single instruction multiple threads (SIMT)
3. High memory bandwidth (not strictly maths anymore)

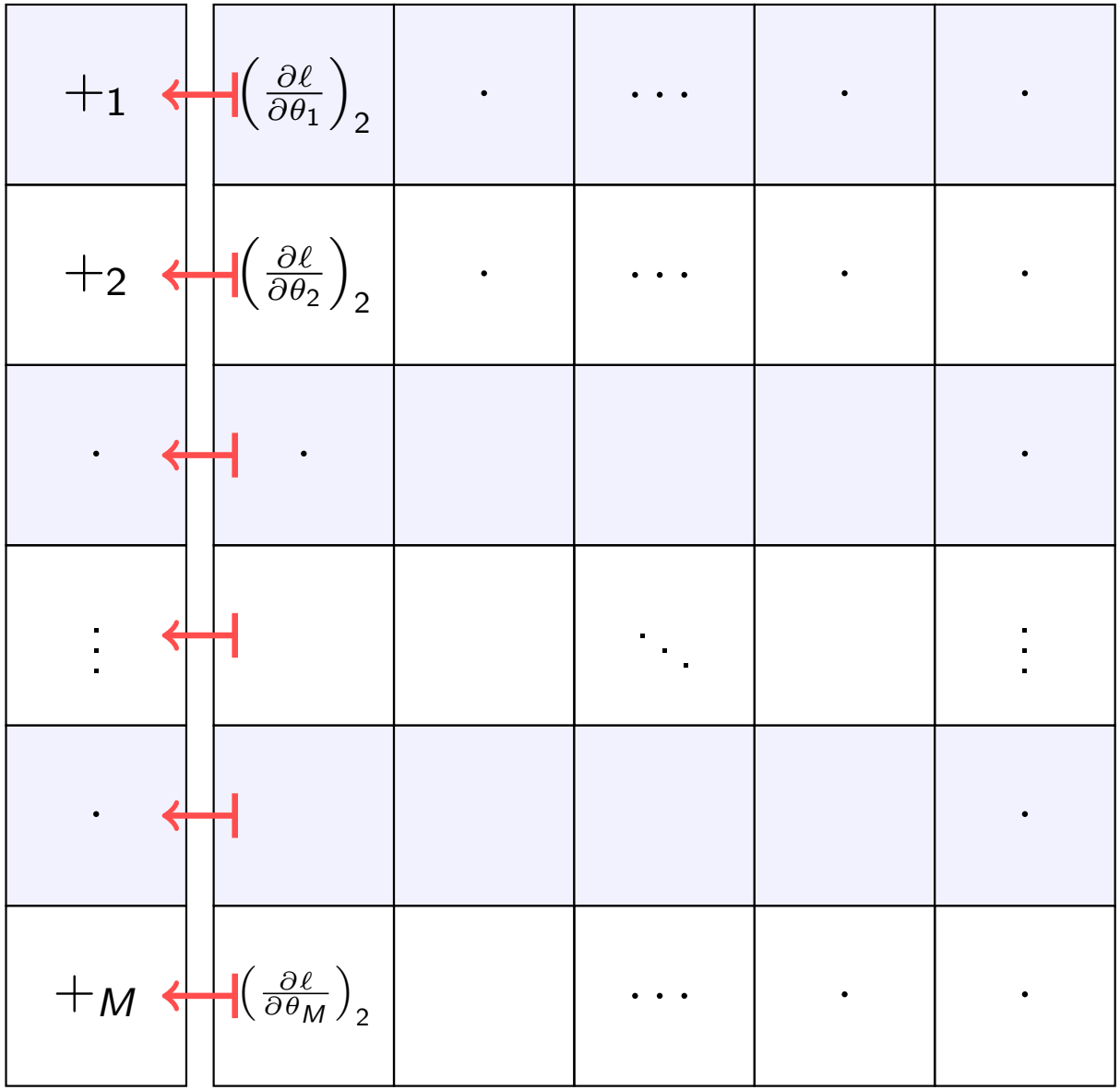
Parallel gradient calculations

$\left(\frac{\partial \ell}{\partial \theta_1}\right)_1$	$\left(\frac{\partial \ell}{\partial \theta_1}\right)_2$	.	$\cdots$	.	$\left(\frac{\partial \ell}{\partial \theta_1}\right)_N$	$\frac{\partial \ell}{\partial \theta_1}$
$\left(\frac{\partial \ell}{\partial \theta_2}\right)_1$	$\left(\frac{\partial \ell}{\partial \theta_2}\right)_2$	.	$\cdots$	.	$\left(\frac{\partial \ell}{\partial \theta_2}\right)_N$	$\frac{\partial \ell}{\partial \theta_2}$
.	.				.	.
$\vdots$			$\ddots$		$\vdots$	$\vdots$
.					.	.
$\left(\frac{\partial \ell}{\partial \theta_M}\right)_1$	.		$\cdots$	.	$\left(\frac{\partial \ell}{\partial \theta_M}\right)_N$	$\frac{\partial \ell}{\partial \theta_M}$

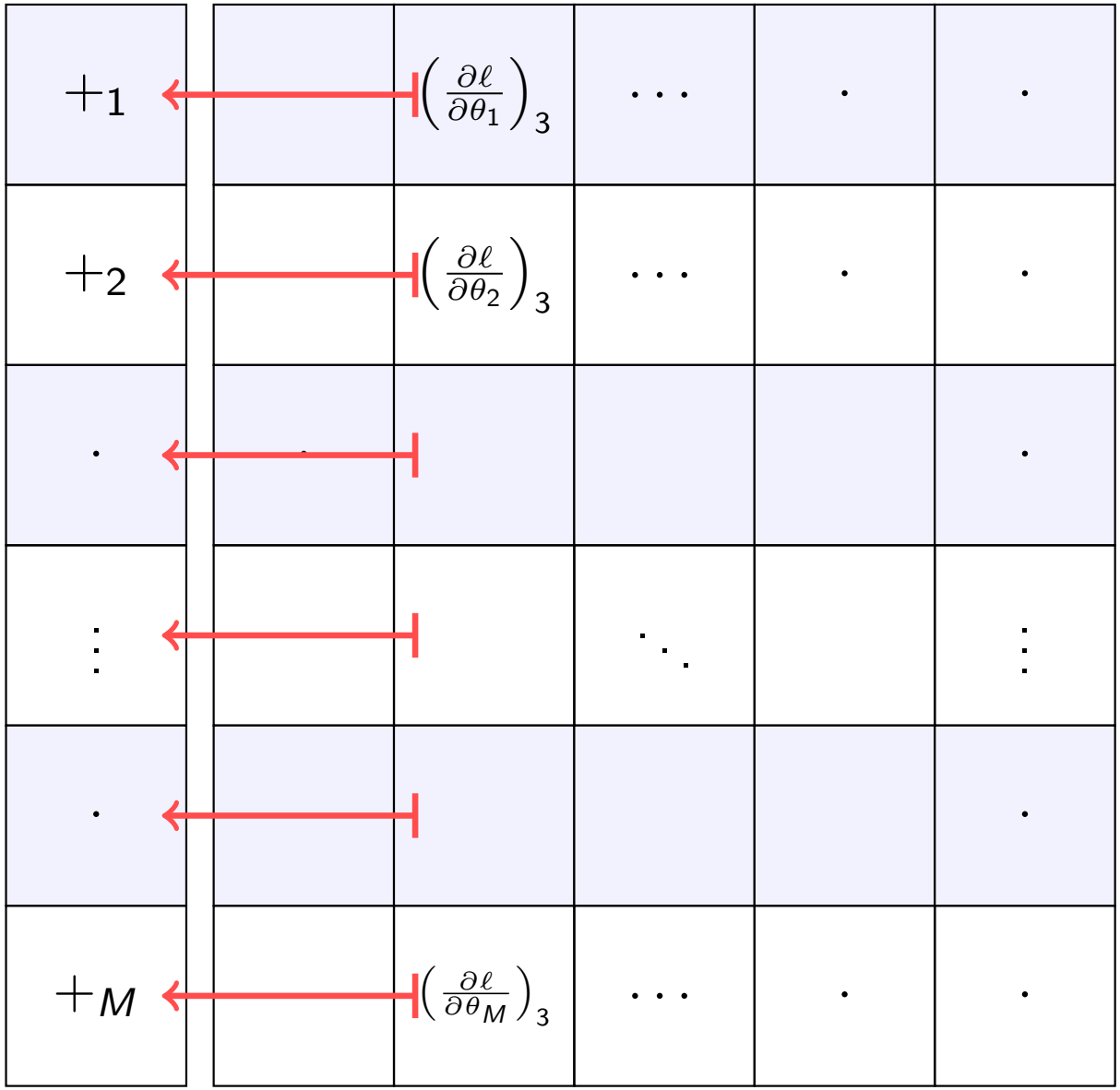
Parallel gradient calculations



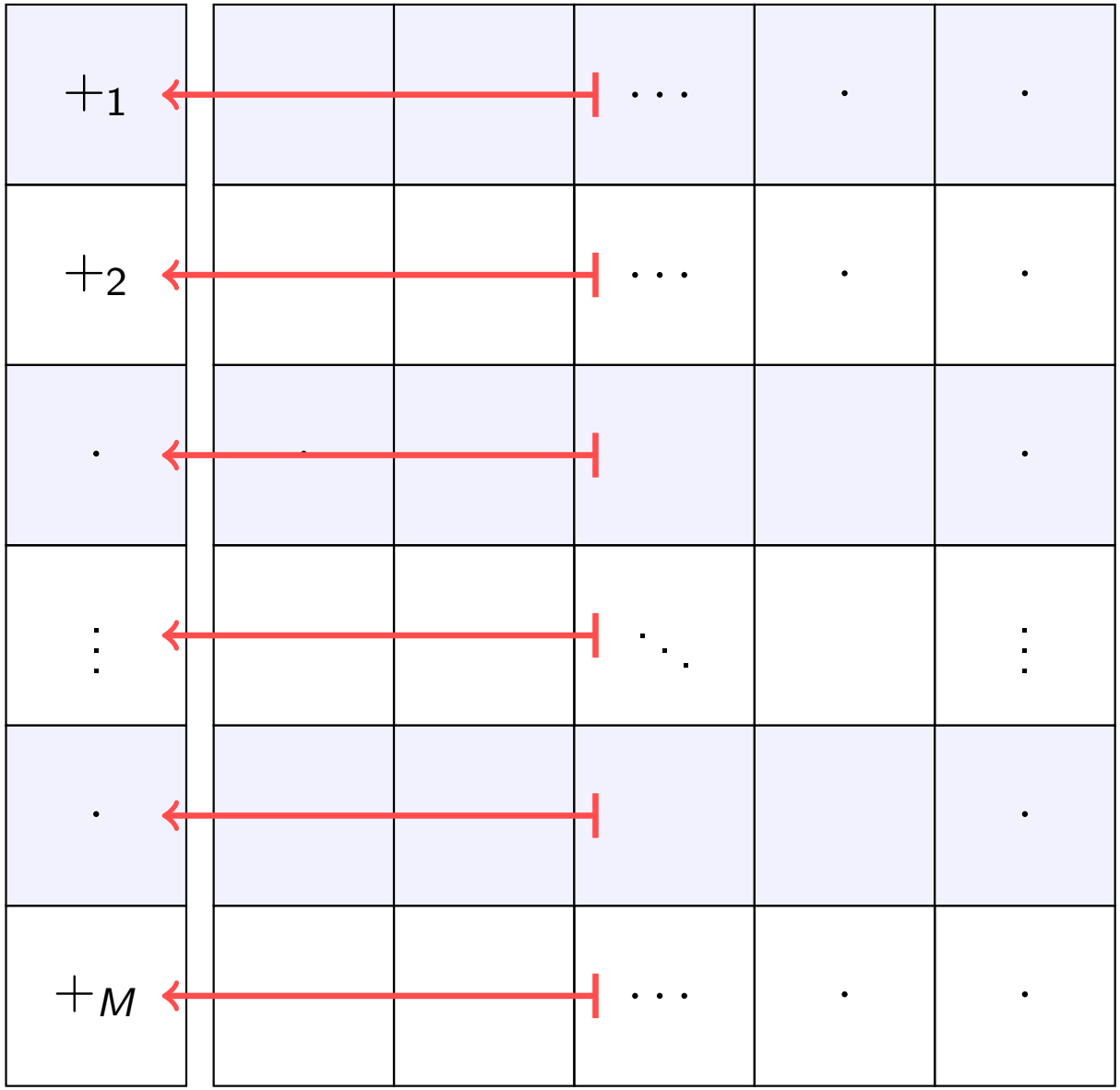
Parallel gradient calculations



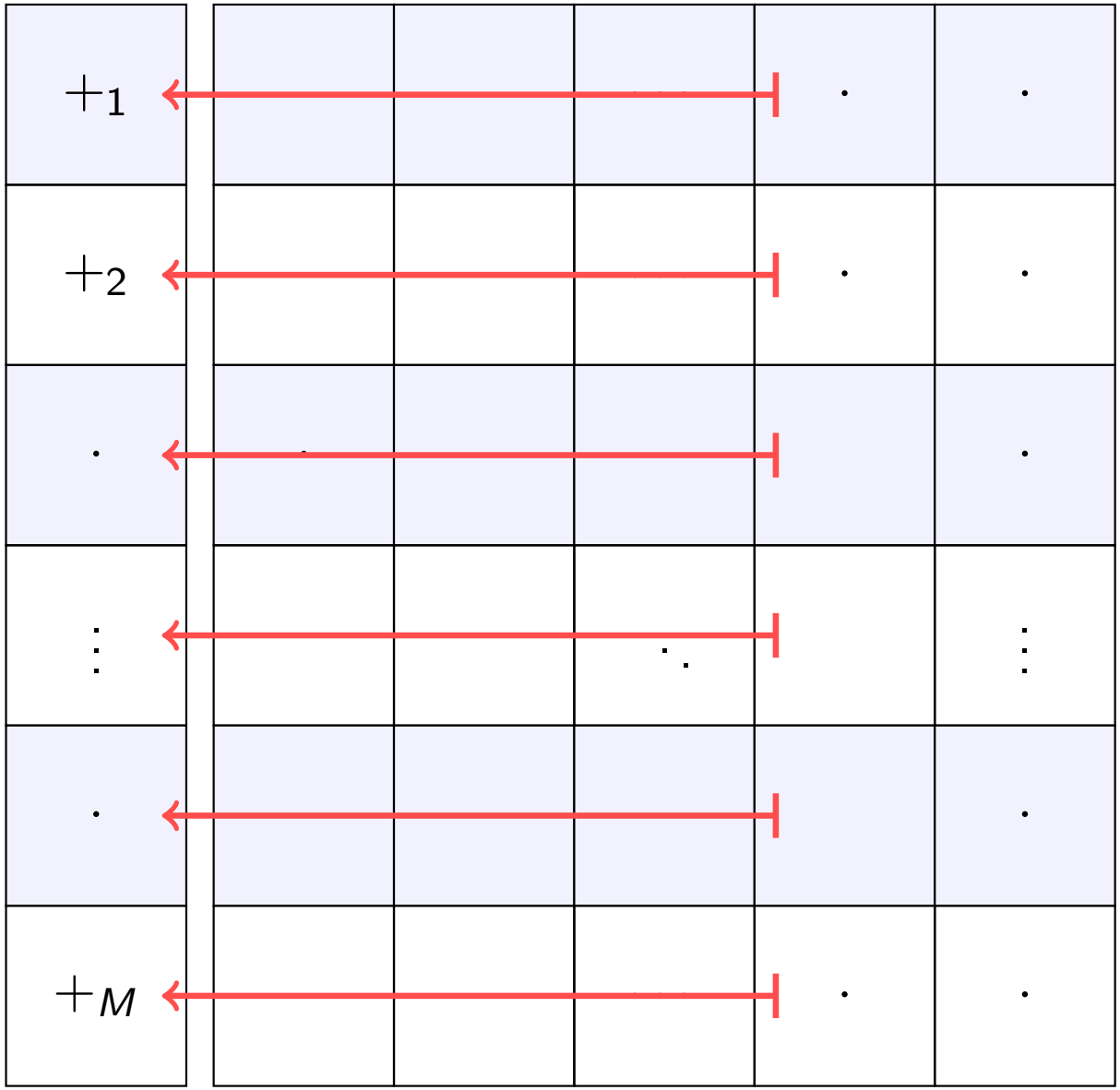
Parallel gradient calculations



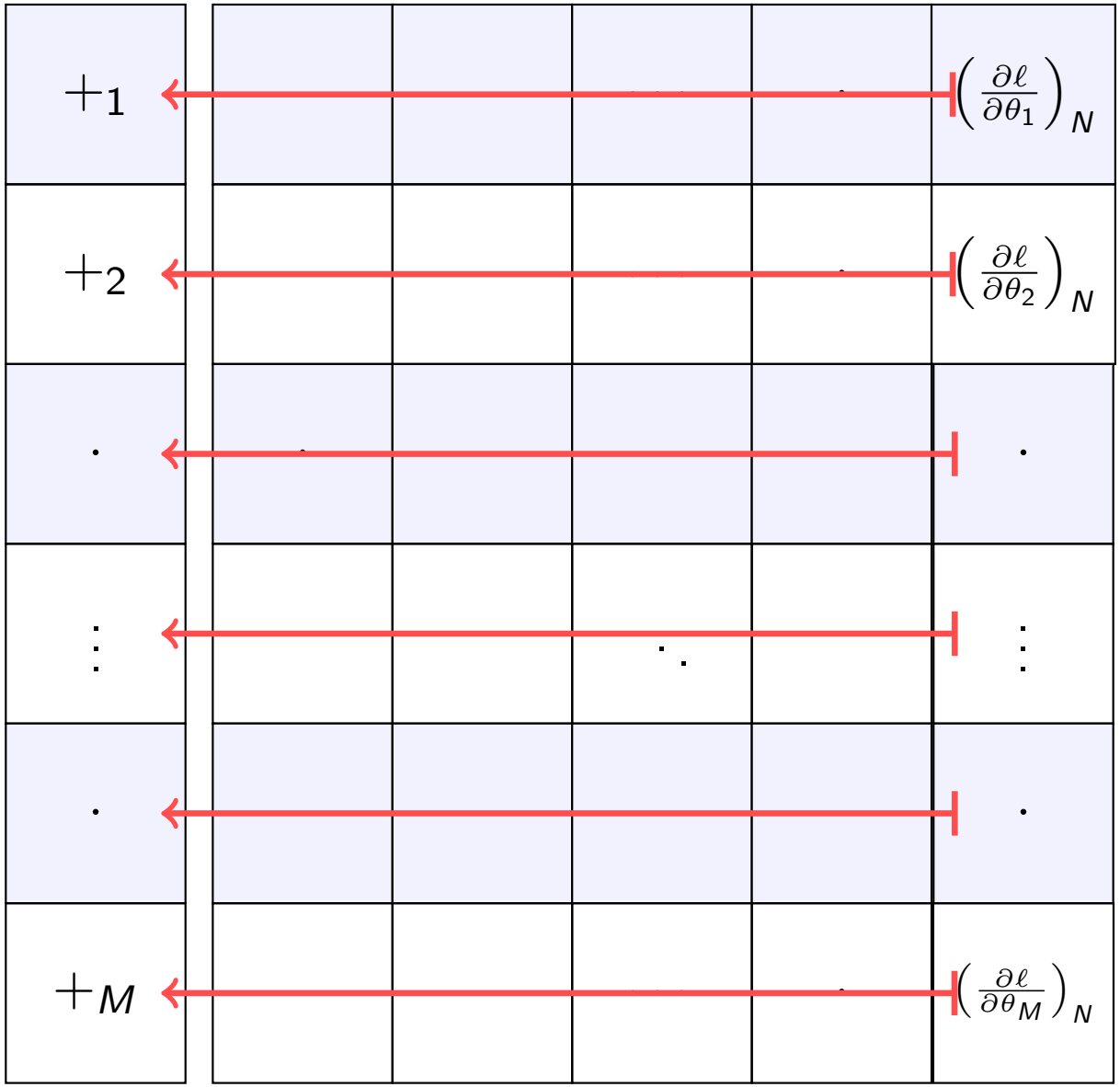
Parallel gradient calculations



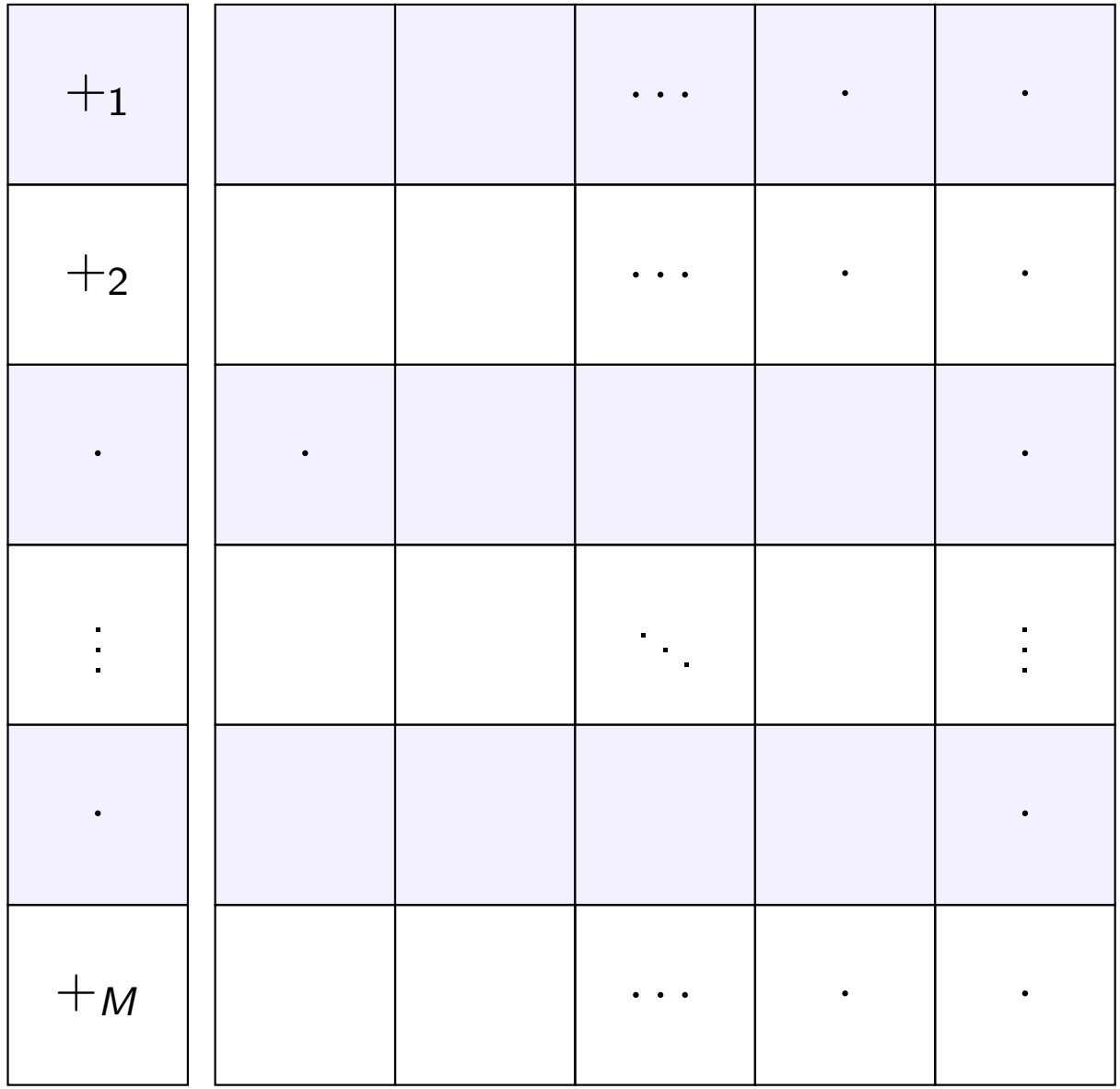
Parallel gradient calculations



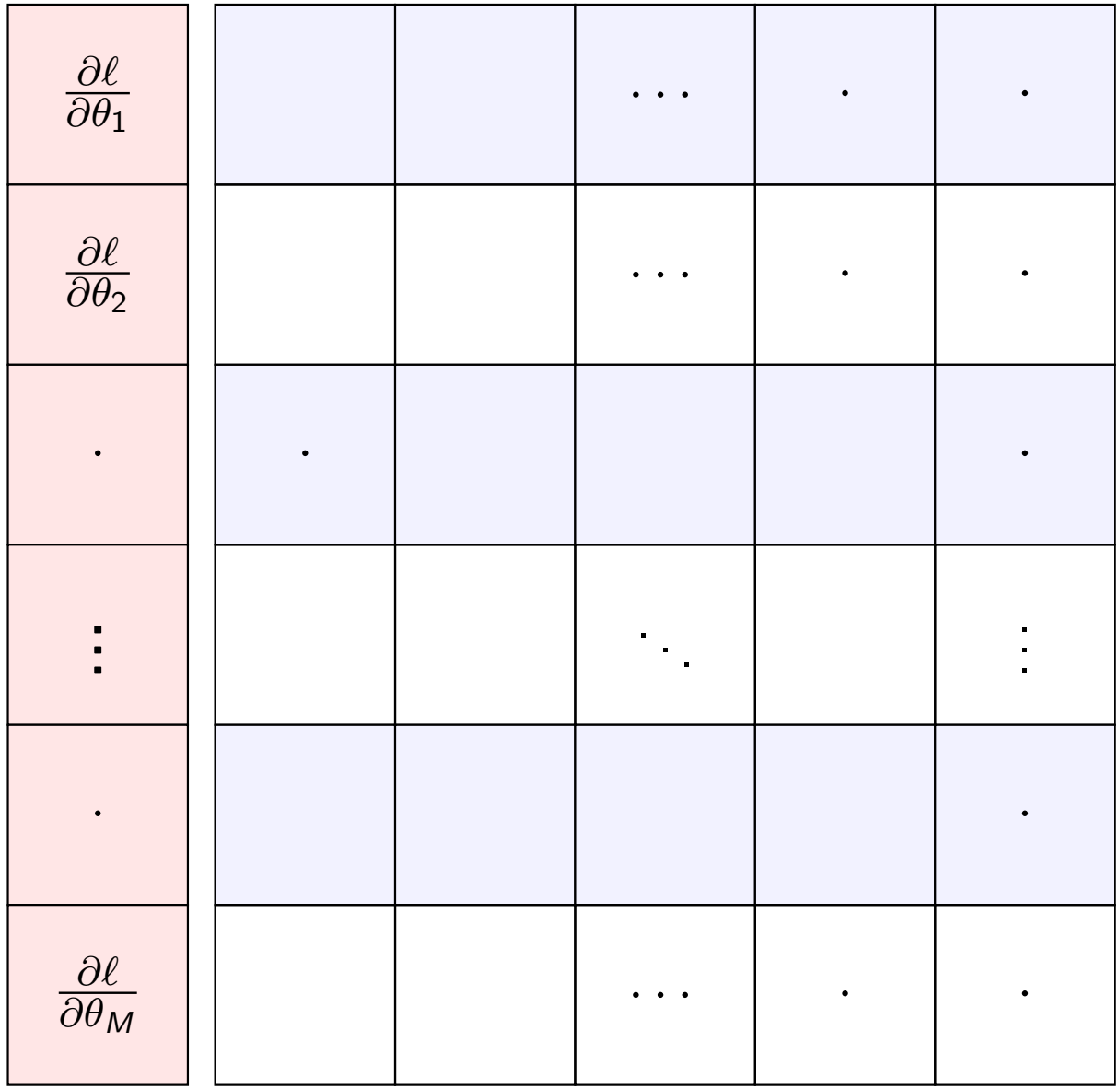
Parallel gradient calculations



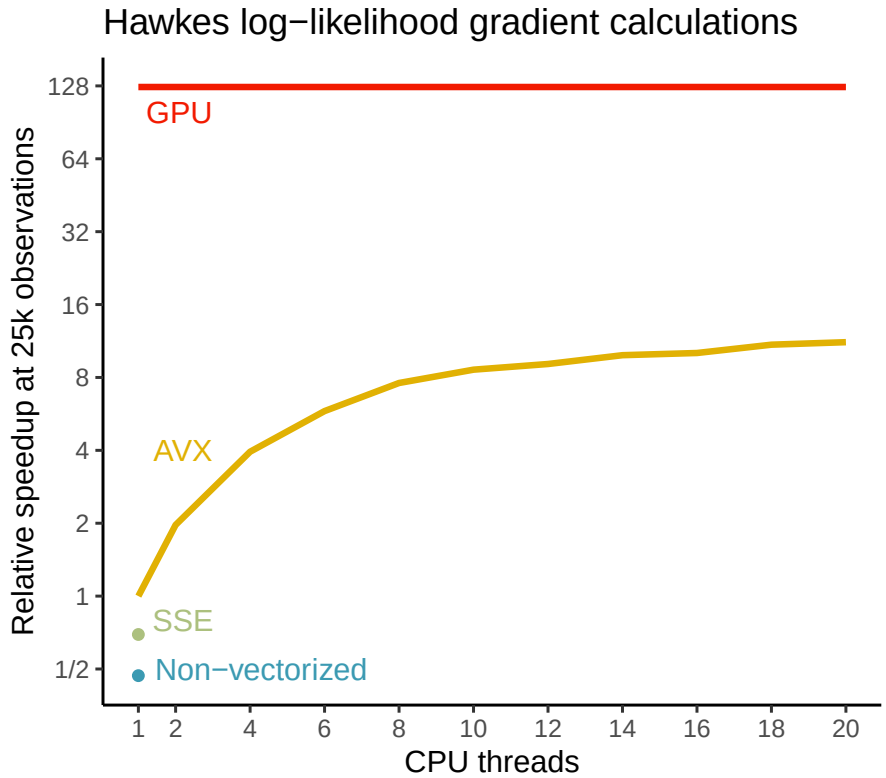
Parallel gradient calculations

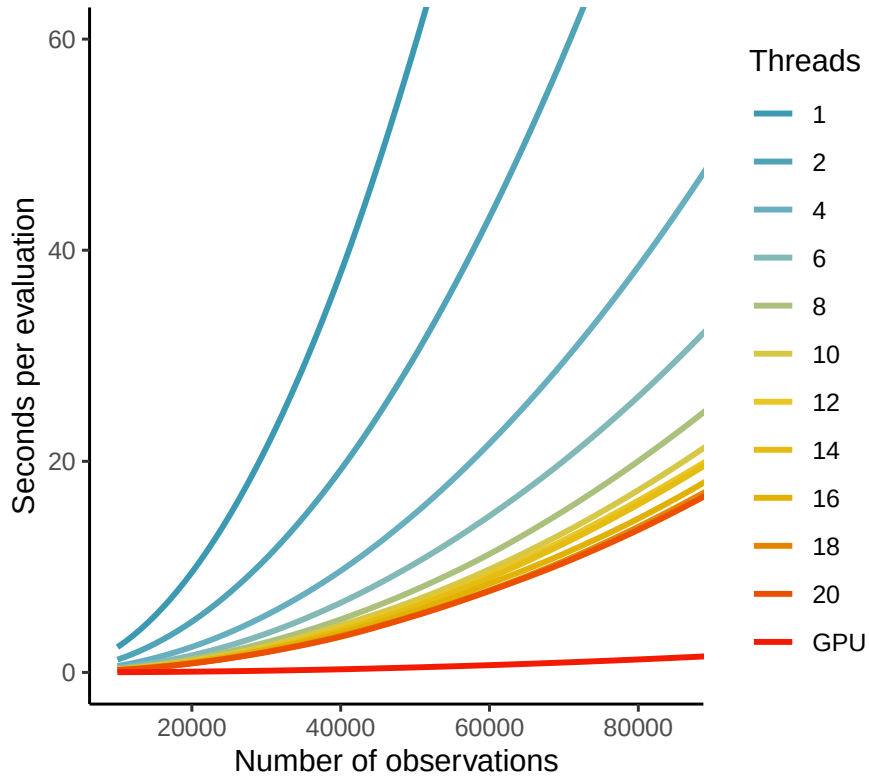


Parallel gradient calculations



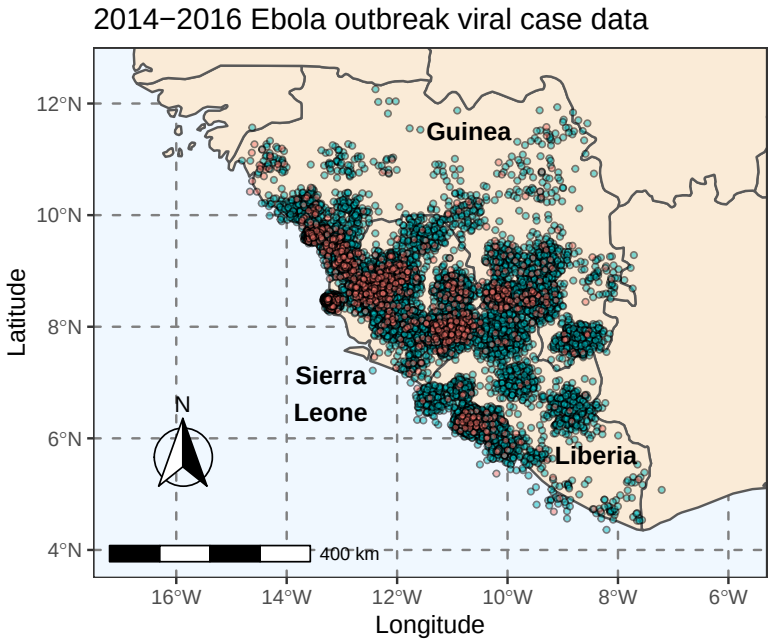
Significant speedups

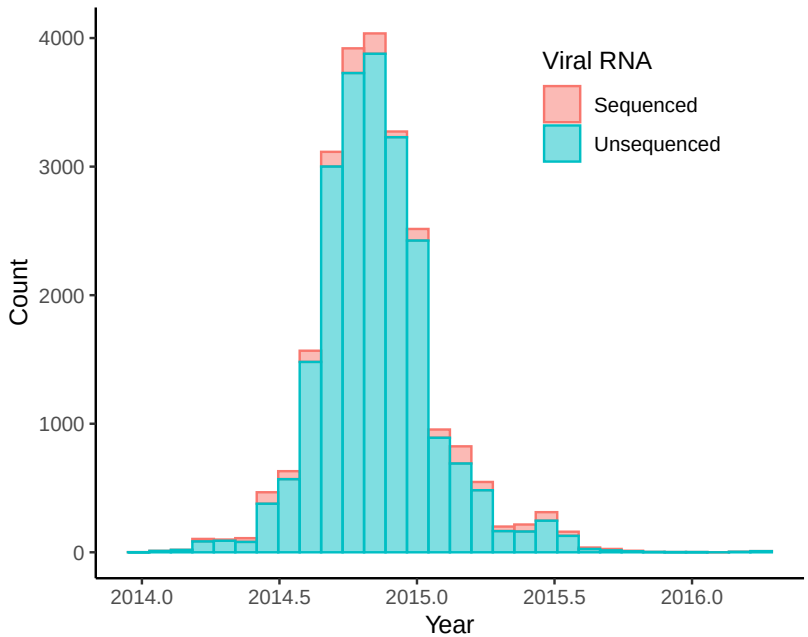




Part 3. Ebola outbreak of 2014-2016

2014-2016 Ebola virus outbreak in West Africa





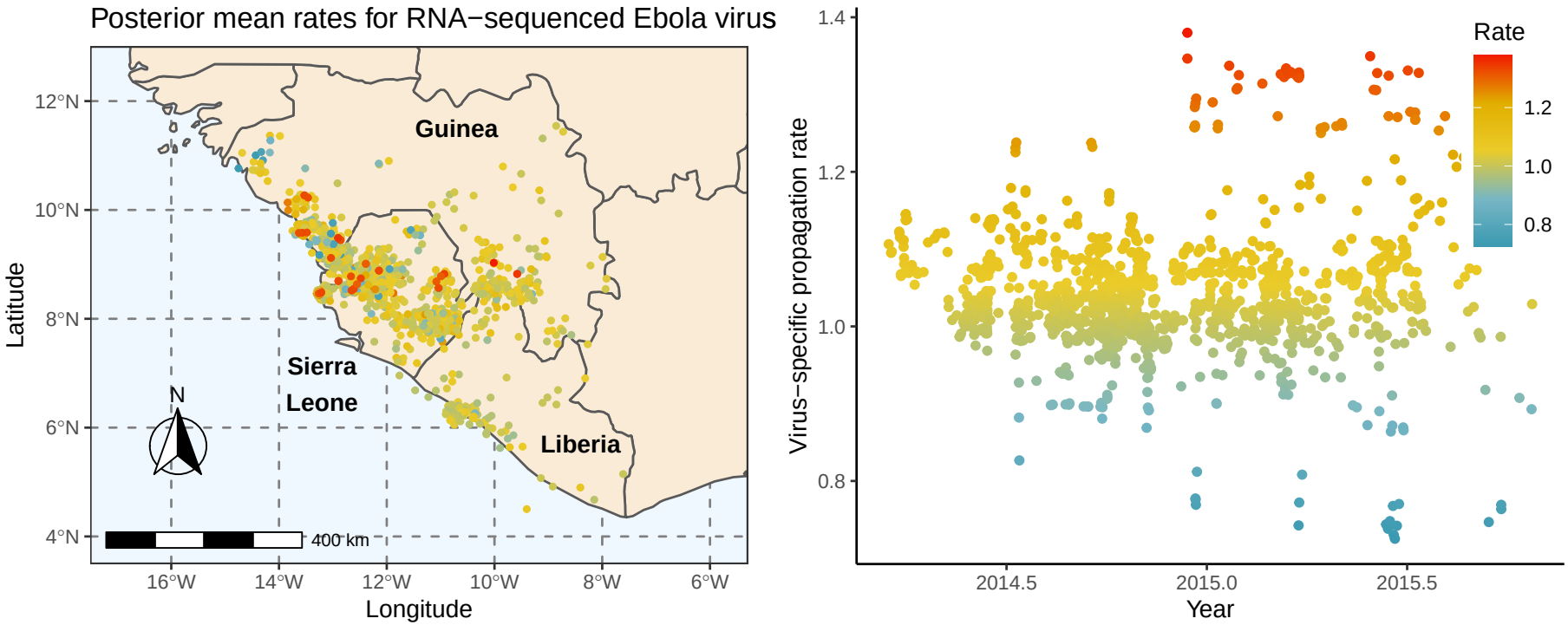
- 1,610 sequenced viruses (1,367 of which have locations data)
- 21,811 unsequenced cases

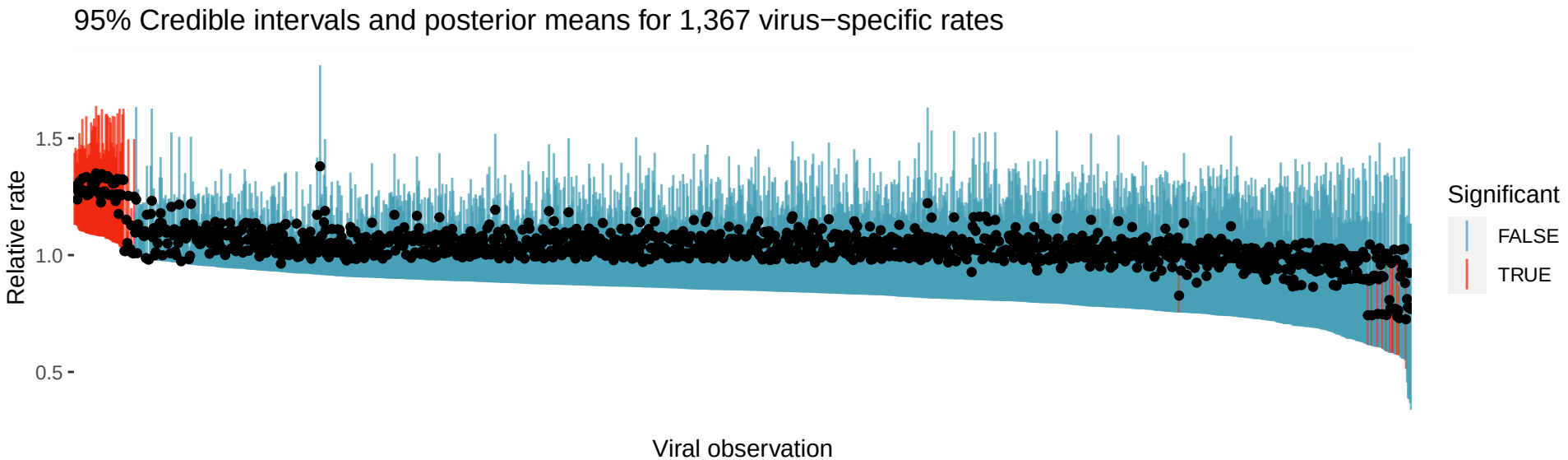
Posterior inference

Generate 20 million Markov chain states ($\sim$ 6 million samples/day on Nvidia GV100)

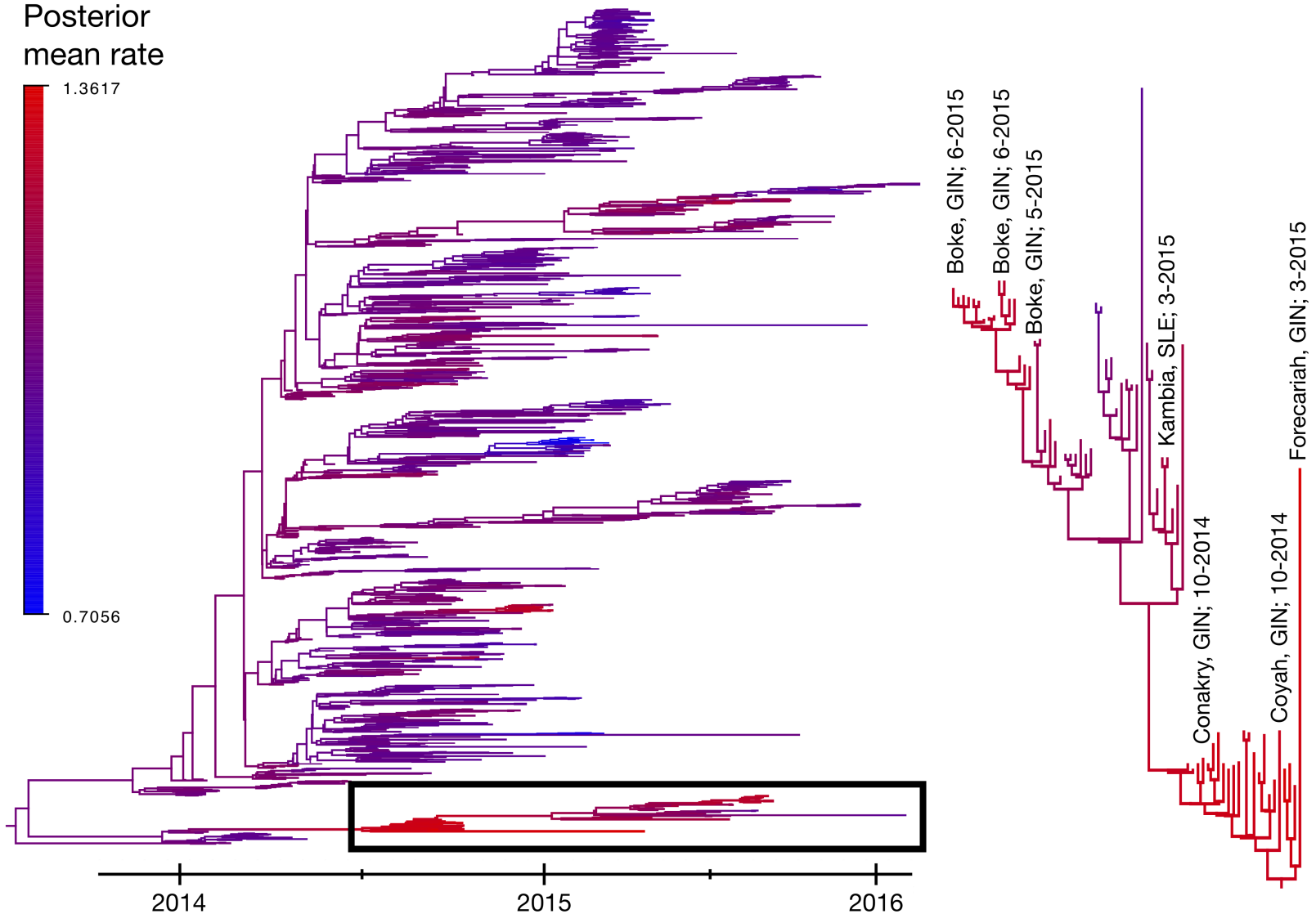
Hierarchical model module	Model parameter	Symbol	Posterior mean (95% HPD cred. int.)	Unit
Hawkes process	Background spatial lengthscale	τ_x	183 (151, 215)	km
	Self-excitatory temporal lengthscale	$1/\omega$	23.3 (22.9, 23.8)	days
	Self-excitatory spatial lengthscale	h	6.69 (6.59, 6.78)	km
	Normalized self-excitatory weight	$\theta_0/(\theta_0+\mu_0)$	0.69 (0.63, 0.74)	—
	Temporal trend coefficient	β	-0.449 (-0.450,-0.446)	—
Phylogenetic diffusion	Standard deviation	σ	3.26 (2.93, 3.62)	log rate

Inferred rates of contagion




95% Credible intervals and posterior means for 1,367 virus−specific rates
Relative rate
1.5
1.0
0.5
Viral observation
Significant
FALSE
TRUE

Biologically modulated rates



Part 4. Much work to do

Model development

1. Is the linear temporal downtrend sufficient?

2. Variable spatial bandwidths (Park et al., 2019)
   parallelizable after precomputing

3. (Everything else in the modern Hawkes toolbox)

4. Going global (probably) requires multivariate approach

Computational development

1. Overhaul structure of hpHawkes
2. Faster gradients by approximation (NNs, stochastic gradients, P^3M)
3. DNNs predict trees from RNA
4. Fast multivariate Hawkes inference

Acknowledgements

Joint work with

- Marc Suchard (UCLA)
- Xiang Ji (Tulane)

Funded by K-Award *Big Data Predictive Phylogenetics with Bayesian Learning*.

Thank you!