Numerical methods for multi-stage models of cancer incidence

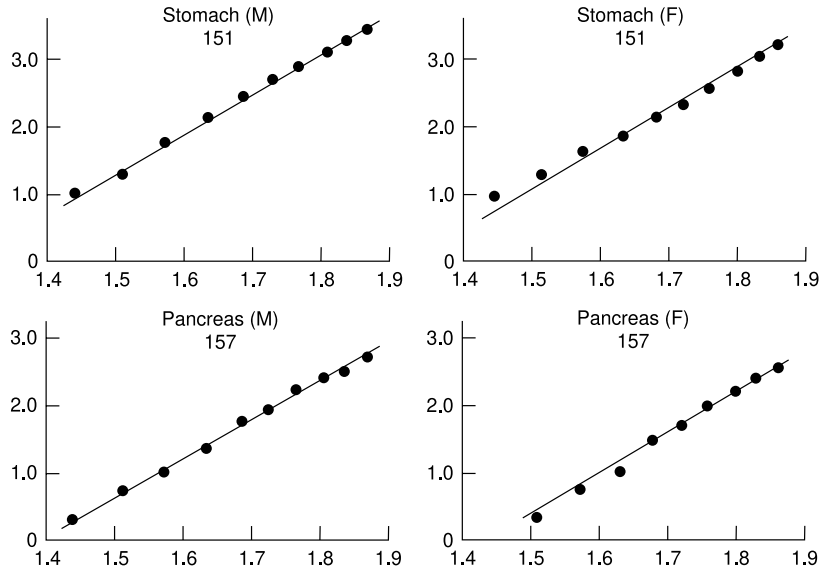
Chay Paterson

University of Manchester

May 9, 2023

Multi-stage models

P. Armitage and R. Doll[12]



[1] P. Armitage and R. Doll, British Journal of Cancer 1954; 8: 1–12

[2] P. Armitage and R. Doll, British Journal of Cancer 1957; 11(2): 161-169

Multi-stage models

A.G. Knudson[1][2]

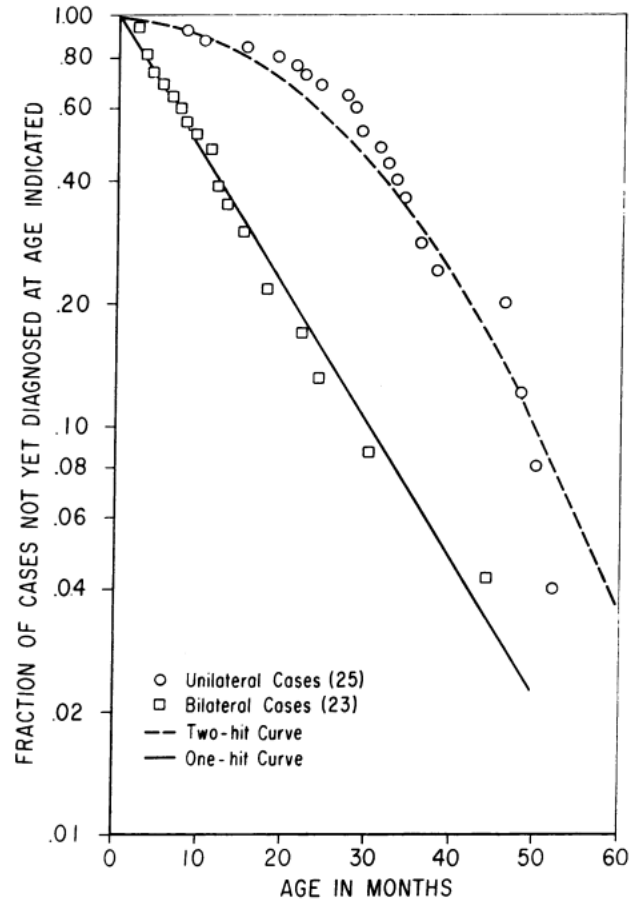


FIG. 1. Semilogarithmic plot of fraction of cases of retinoblastoma not yet diagnosed (S) vs. age in months (t). The one-hit curve was calculated from $\log S = -t/30$, the two-hit curve from $\log S = -4 \times 10^{-5} t^2$.

---

[1] AG. Knudson, PNAS 68.4 (1971): 820-823.

[2] F. Michor, Y. Iwasa and MA. Nowak, Nature Reviews Cancer 2004; 4: 197-205 doi:10.1038/nrc1295

How do these studies work?

What are the relevant observables in this type of longitudinal study & data analysis?
Follow a cohort that contains some of our patients in the study:

- Age-specific incidence $I(a)$: number of new cases are recorded in the cohort with ages between a and $a + da$
- Hazard function $h(a)$: *rate* with which individuals in cohort at age a are diagnosed.
- Survival function $S(a)$: proportion of individuals in cohort who have not yet been diagnosed by age a

How do these studies work?

How are these variables related?

Denote by $n(a)$ the number of people in the reference cohort that remain undiagnosed by age a. Then:

$$I(a) = n(a)h(a)da \tag{1}$$

and $n(a)$ is actually determined by $S(a)$ and the initial reference population of the cohort (num. of babies born at same time) $n(0)$:

$$n(a) = n(0)S(a) \tag{2}$$

so, logically:

$$I(a) = n(0)S(a)h(a)da \tag{3}$$

How do these studies work?

How are these variables related?

But it's also true that the number of people diagnosed during the period $[a, a + da)$ must be:

$$n(a) - n(a + da) = I(a) \tag{4}$$

hence:

$$I(a) = n(a) - n(a + da) = n(0)(S(a) - S(a + da))$$

$$\implies S(a) - S(a + da) = S(a)h(a)da$$

$$\implies h(a) = (S(a) - S(a + da))/(S(a)da) \to -\frac{d \ln S}{da}$$

as $da \to 0$.

So what?

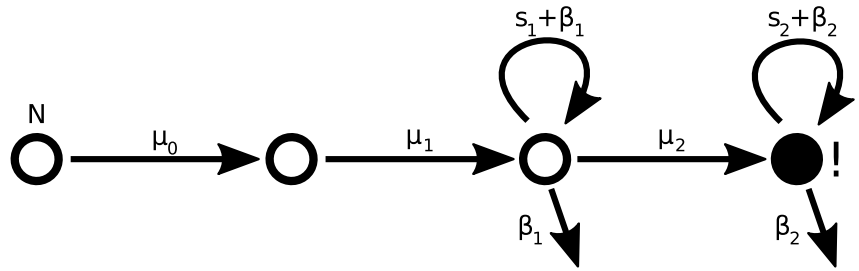
- The survival curve S(a) determines everything else of interest.
- The model determines the survival curve. The parameters determine the model. We want to know what model+parameters best agree with data from studies.
- To fit (or "train") the model to longitudinal data, we need to compute $S(a)$!

This is the central mathematical problem in cancer epidemiology. *How do we compute S?*

Multi-stage clonal expansion models

2-3 rate limiting steps[123]

Problem: how to compute $S(t)$ for a given model?

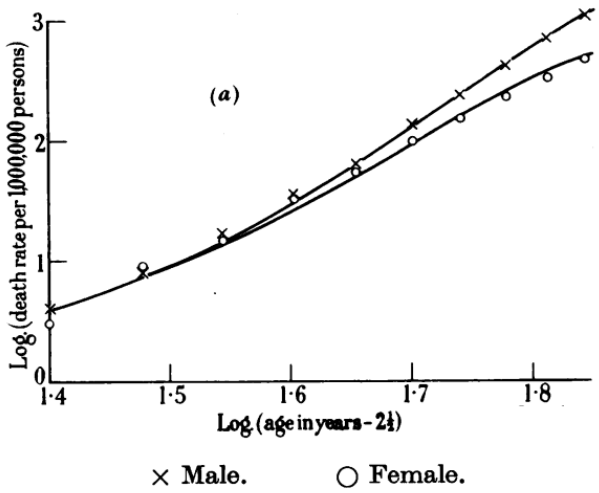


Different methods: Fast:

- Armitage + Doll's approximation[1]
- Moolgavkar + Venzon's quadrature[2]

Very slow:

- Gillespie algorithm + sampling [3]



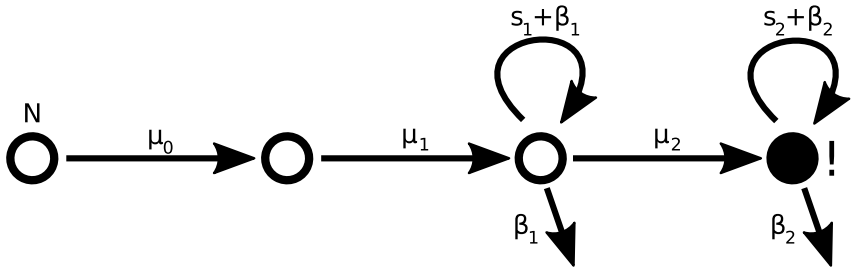
↓?

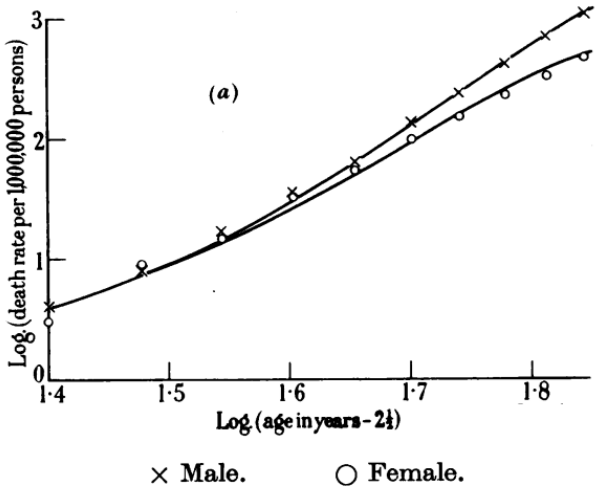
[1] P. Armitage and R. Doll, British Journal of Cancer 1957; 11(2): 161-169

[2] S. Moolgavkar and G. Luebeck, JNCI 1992; 84(8): 610-618

[3] C. Paterson, I. Bozic, H. Clevers, PNAS 2020; 117(34): 20681-20688 (supp. material)

Armitage and Doll's approximation





- Assume all the probabilities are small
- Then the relevant probability $S(a)$ is expressible in terms of expected values/population means, which implies

$$S(a) \propto a^k(e^{sa} - 1) \qquad (5)$$

with constants k and s.

- Don't use correlations, variances, or higher moments in stem cell populations – just ignore these.

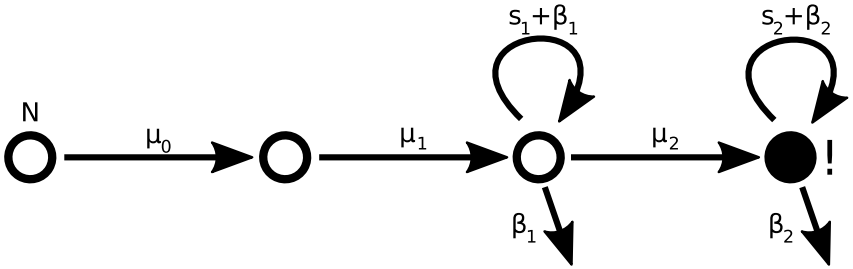
Armitage and Doll's approximation

What's wrong with this?

- In old age, the approximation will fail – it assumes the probabilities are small, we cannot use it when we know the probabilities will be high.
- Predicts that cancer risk should increase in an accelerating way with age
- (Which it doesn't – the hazard $h(a)$ levels off (R. Meza))

Which leads us to...

The "gold standard": Moolgavkar, Venzon, and Luebeck's approach



Same basic model as Armitage and Doll, but solved exactly (no approximations).

- Don't assume anything about probabilities or correlations.
- Transform the equations, and try to solve for the distribution directly.

Moolgavkar, Venzon, and Luebeck's approach

How it works

Define a set of generating functions Psi_k (one for each stem cell population k):

$$\Psi_k(t, \vec{x}) := \mathbb{E}\left[\prod_j x_j^{N_j} | N_k = 1, N_{j\neq k} = 0\right] \tag{6}$$

Derive Kolmogorov backward equation to evolve this backwards in time:

$$\frac{d\Psi_k}{dt} = (...\Psi_k \; and \; \Psi_{k+1}...) \tag{7}$$

and we can compute the survival curve $S(a)$ from

$$S(a) = \Psi_0(a, 1, 1, 1, ..., 1, 0) \tag{8}$$

Moolgavkar, Venzon, and Luebeck's approach

How it works

In fact, we get a recursive hierarchy of S curves for different models:

$$S_k(a) = \exp\left(\ldots \int_{z=0}^{a} S_{k+1}(z)\ldots dz\right) \tag{9}$$

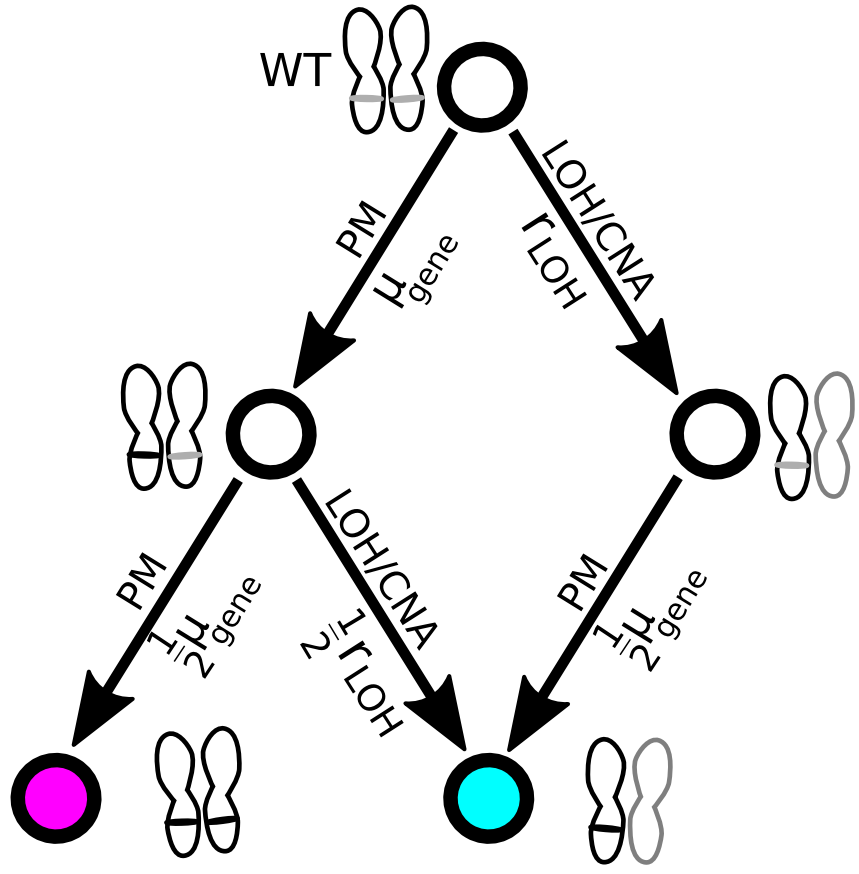
which can be evaluated very efficiently with numerical integration. This has been the best available method for evaluating $S(a)$ in multi-stage models since about 1992. (See: Bhat, Georg's package on R-CRAN).

So what's the problem with the "classical" approach?

Models on graphs

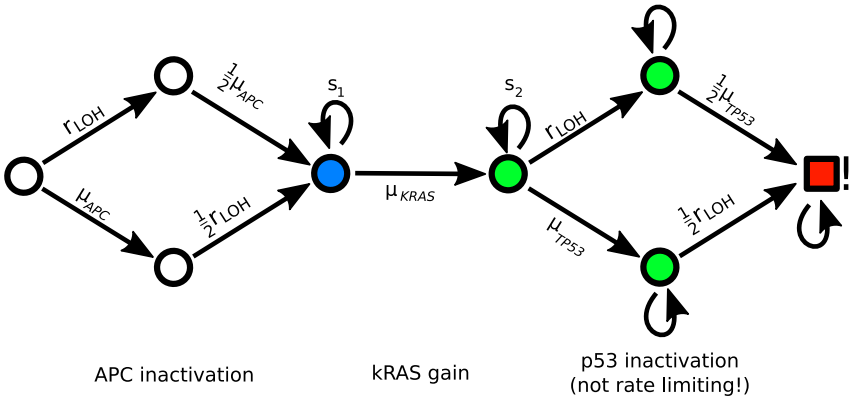
1. The classical approach is specialised for 2- and 3-hit models: it doesn't work on graphs
2. To study **specific genes** and mechanisms of interest (SNVs, LOH, CNA, etc.), we need to evaluate $S(a)$ for a model defined on a graph (right)
3. What methods do we actually have?

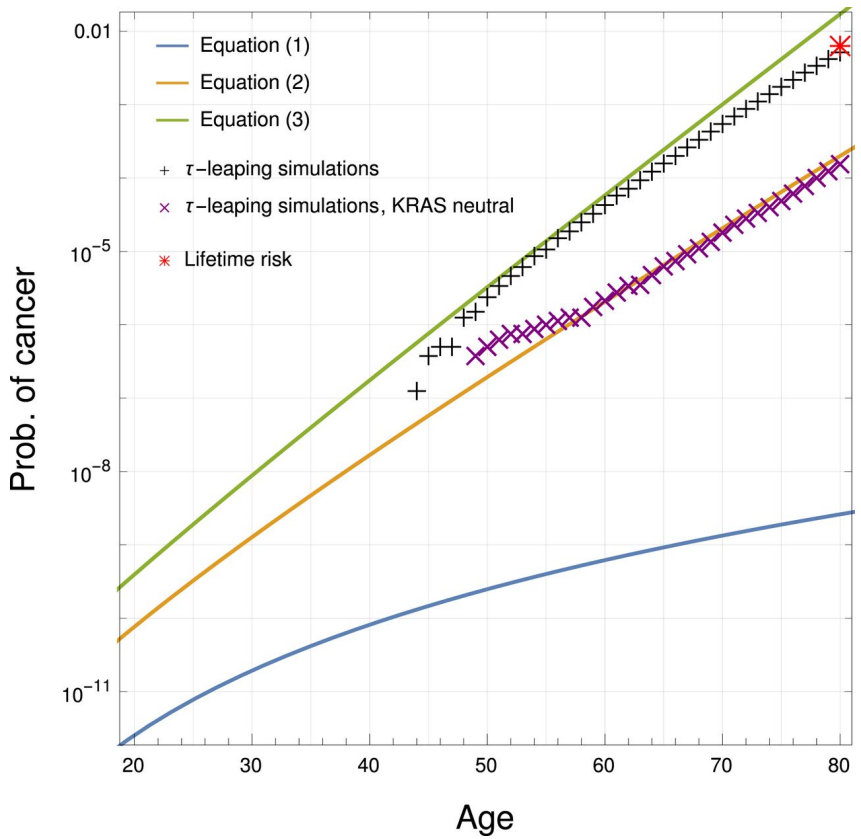
example model



This gets us the incidence of *specific karyotypes*

Colorectal adenocarcinoma model





- Can use A-D type approximation, or stochastic simulations

but:

- Mean-field breaks down at old ages / large probabilities
- Stochastic simulations are *extremely slow*

[1] C. Paterson, I. Bozic, H. Clevers, PNAS 2020; 117(34): 20681-20688

New idea!

Define a more general generating function Ψ:

$$\Psi(t, \vec{q}) = \mathbb{E}[\left[\prod_j q_j^{N_j}\right] \tag{10}$$

and derive Kolmogorov FORWARD equations instead. Then we can numerically integrate these, and get survival curves $S_i(a)$ for different types of cancer i. E.G. tumours with clonal LOH, or no clonal LOH.

$$S_i = \Psi(t, q_j = 1, \ldots, q_i = 0) \tag{11}$$

People knew about this approach for a long time (since 1988ish) but it was never considered as useful as backward equations.

[1] e.g. DW Quinn, S Moolgavkar both described the basic version at about this time

Easier said than done

In and out, twenty minute adventure...

Easier said than done

In and out, twenty minute adventure...

Five years later...

Fast forward method

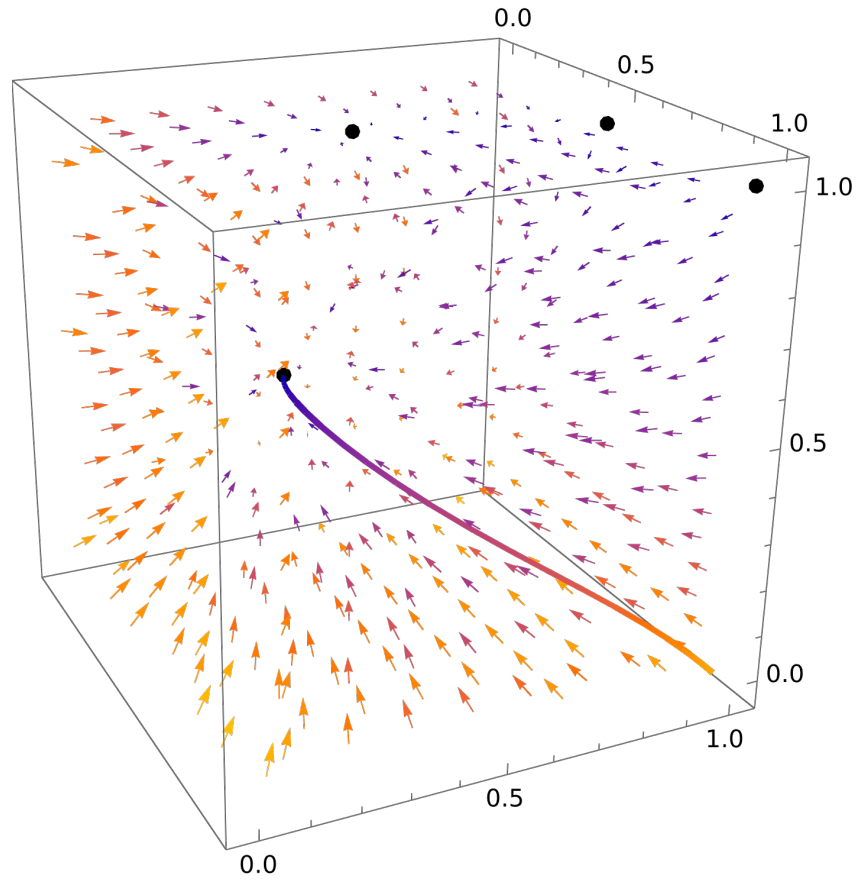
Using the big generating function Ψ, find the Kolmogorov forward equations:

$$\frac{d\Psi}{dt} = \vec{X} \cdot \nabla\Psi \qquad (12)$$

This can be solved using the method of characteristics, so we can instead integrate

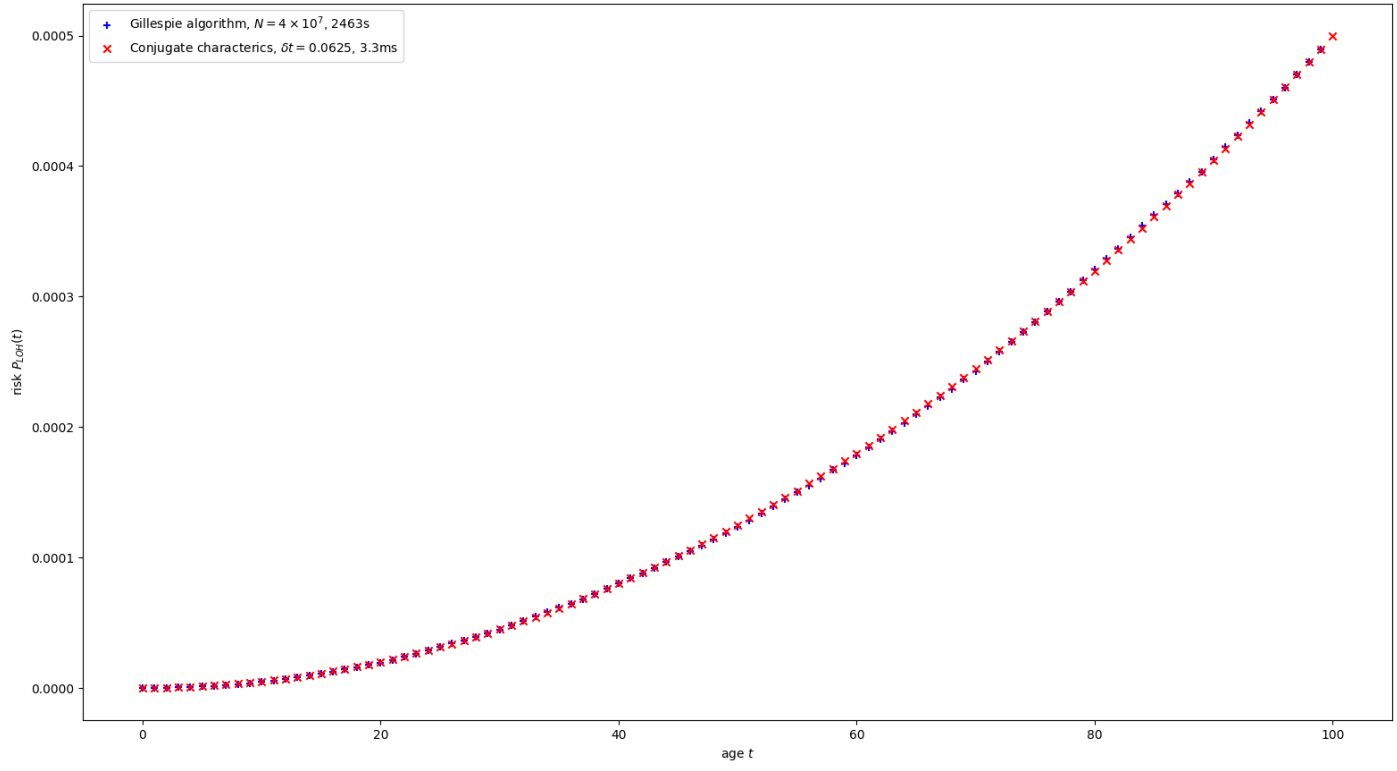
$$\frac{d\vec{\gamma}}{dt} = \vec{X} \qquad (13)$$

numerically, using a method like improved Euler integration.



The vector field $\vec{X}$ and a characteristic $\vec{\gamma}$

Fast forward method



Gillespie algorithm vs fast forward method:

Gillespie: 2500s

Fast forward: 3ms (1ms for same error)

To compare the two methods, ask under what conditions the errors are comparable. Stochastic algorithms have

$$\epsilon \sim N^{-1/2} \tag{14}$$

and will thus need

$$N \sim \mathcal{O}(\epsilon^{-2}) \tag{15}$$

runs, and overall runtime $T \propto N$.

Fast forward method

Error analysis

Fast forward method has

$$\epsilon \sim \Delta t^2 \tag{16}$$

and runs in

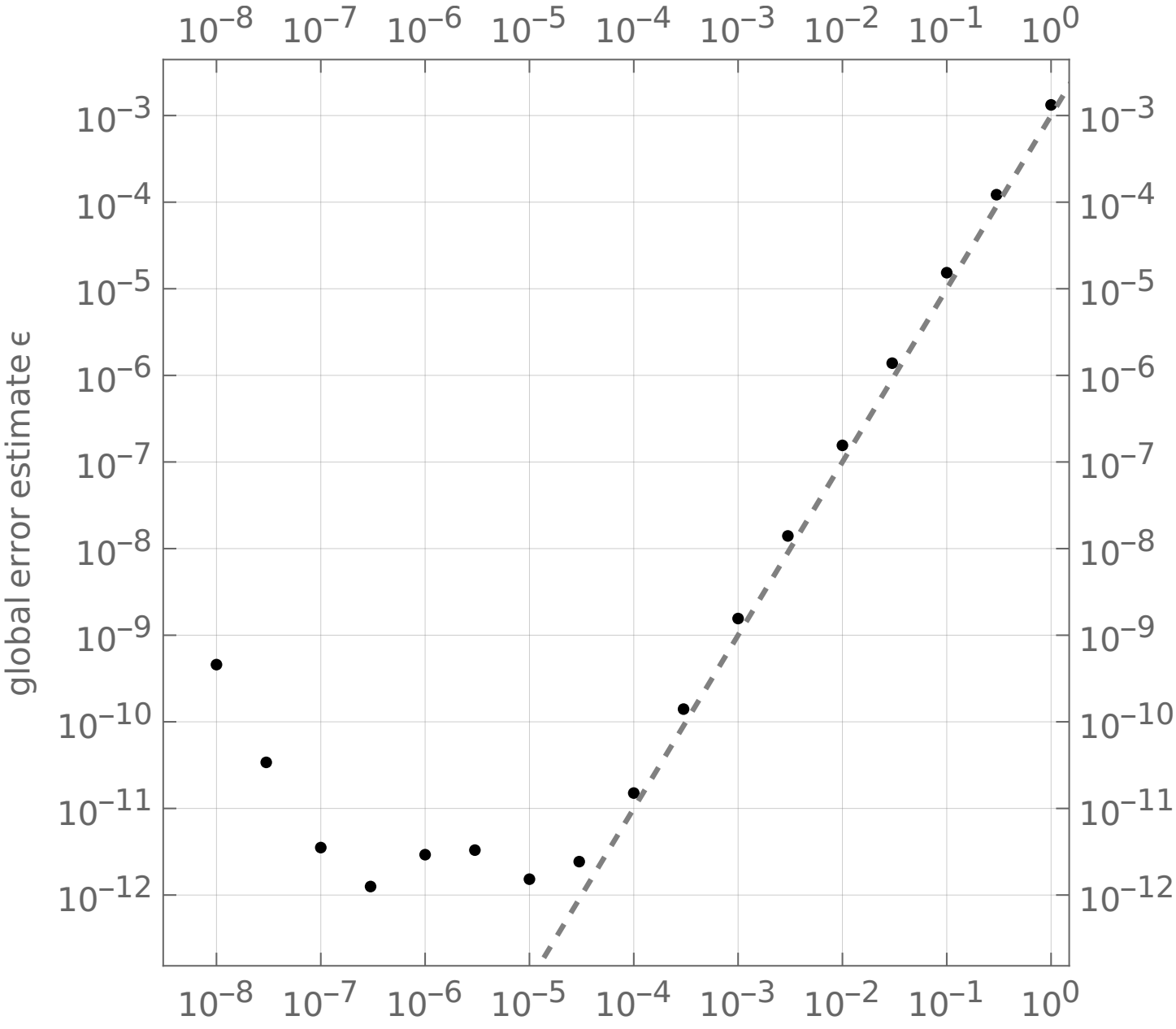
$$T \sim \Delta t^{-1} \sim \mathcal{O}(\epsilon^{-1/2}) \tag{17}$$

so new runtime $\sim \mathcal{O}\left(\text{ old runtime }^{1/4}\right)$.

Amazing!

Fast forward method

Error analysis



Fast forward method

Error analysis

Why wasn't this used before? DW Quinn and S Moolgavkar both studied forward equation+characteristic methods, but used Euler integration, which has

$$\epsilon \sim \Delta t \tag{18}$$

and required two passes, so it ran in

$$T \sim \Delta t^{-2} \sim \mathcal{O}(\epsilon^{-2}) \tag{19}$$

this is asymptotically just as bad as random sampling! You only get a constant speed-up.

Fast forward method

Parameter inference

- Can use fast forward method to quickly compute likelihoods
- Can then test different parameters in e.g. simulated annealing, and maximise the likelihood!
- This process can run in about a minute – this would be impossible with Gillespie's algorithm and e.g. ABC.

```
chay@atuin:~/Projects/Simulations/cancer-integ
Generating synthetic dataset...
Ground truth:
  mu = 5e-08
  rloh = 5e-07
  fitness1 = 0.05
  fitness2 = 0.03
  inipop = 1e+06
Done. Saving...
Target likelihood:
-log L = -nan
Starting annealing...
Initial likelihood:
-log L = 22390.6
System fully cooled after 4154 iterations
-log L = 178.46191
Best guesses:
  mu = 1.04331e-07
  rloh = 1.94891e-07
  fitness1 = 0.05
  fitness2 = 0.03
  initialpop = 1e+06

real    2m28.695s
user    2m28.641s
sys     0m0.036s
chay@atuin:~/Projects/Simulations/cancer-integ
```

Fast forward method

Parameter inference

Gets the right order of magnitude for some parameters, but:

- Very sensitive to choice of neighbours.
- Very sensitive to cooling schedule.
- Some parameters (initial stem cell population N_0) are not identifiable. This is a universal problem though.

Still a work in progress... But a great result!