Inference and data-driven modeling of population of cells

Stefan Engblom

Div of Scientific Computing, Dept of Information Technology, Uppsala University

SPARC, Uppsala, May 18th, 2016
Outline

Overview ("How to think")

1. Raw data, tracking of cells, & hypotheses

2. Models & inference

3. Results & visualization

Summary

Future
Long-term goal

“How to think”

From data to models...

- Overall aim is to put forward realistic and useful computational models of cells:
  - White blood cells being actively recruited to a transplant area
  - Raw microscopy data is available, but needs to be interpreted, e.g., by doing inference within the scope of some model
Long-term goal
“How to think”

From data to models...

- Overall aim is to put forward **realistic** and **useful** computational models of cells:
  - Realistic: flexible and understandable (= analyzable) numerical models, *that in the longer perspective can incorporate all conceivable relevant processes*
Long-term goal

“How to think”

From data to models...

- Overall aim is to put forward **realistic** and **useful** computational models of cells:
  - Realistic: flexible and understandable (= analyzable) numerical models, *that in the longer perspective can incorporate all conceivable relevant processes*
  - Useful: be able to *test hypotheses*, be of *predictive value* ("emergent behavior"), *help to build an argument in cases where many factors are unknown*
Long-term goal

“How to think”

From data to models...

- Overall aim is to put forward realistic and useful computational models of cells:
  - Realistic: flexible and understandable (= analyzable) numerical models, that in the longer perspective can incorporate all conceivable relevant processes
  - Useful: be able to test hypotheses, be of predictive value (“emergent behavior”), help to build an argument in cases where many factors are unknown

- Specifics here: white blood cells being actively recruited to a transplant area

- Raw microscopy data is available, but needs to be interpreted, eg. by doing inference within the scope of some model
Risk of over-modeling...

“...help to build an argument in cases where many factors are unknown...”

- really detailed, or,
- imaginary accuracy, or,
- just a plain overfit?
Long-term goal

From data to models...

Data equipped with hypotheses $\Rightarrow$

- Models **simple enough** to be able to express the hypotheses (useful, since data+models may then support the hypotheses)...
Long-term goal
From data to models...

Data equipped with hypotheses $\implies$

- Models **simple enough** to be able to express the hypotheses (useful, since data+models may then support the hypotheses)...  
- ...but **generalizable** such that they can later be understood in a more complete framework (so realism comes later, relies here on ‘flexible’ and ‘understandable’).
Long-term goal
From data to models...

Data equipped with hypotheses $\Rightarrow$

- Models **simple enough** to be able to express the hypotheses (useful, since data + models may then support the hypotheses)...
- ...but **generalizable** such that they can later be understood in a more complete framework (so realism comes later, relies here on ‘flexible’ and ‘understandable’)

Next:
1. Data, tracking, & hypotheses
2. Models & inference
3. Results & visualization
Data (>4GB)
Data, tracks
Data

Issues remaining

- Resolution in z much *much* worse
- Few accurate tracks ($n = 30$), but many inaccurate ones ($n = O(100)$)
- Remaining data artefacts: outliers, null observations, noise due to camera movements
Hypotheses and formulated questions

After looking at many many many movies...

1. When neutrophils enter the area, they seem to search the environment

2. Some of them appears to “get a mission” and transport themselves fairly quickly to some other point where they then remain

3. This transport seems to be initiated at a few ‘hot-spots’

So “Search” → “Transport” → “Recruited”...
Hypotheses and formulated questions

After looking at many many many movies...

1. When neutrophils enter the area, they seem to search the environment
2. Some of them appears to “get a mission” and transport themselves fairly quickly to some other point where they then remain
3. This transport seems to be initiated at a few ‘hot-spots’

So “Search” $\rightarrow$ “Transport” $\rightarrow$ “Recruited” ...

Some more questions:

- Are all missions obtained in a hot-spot?
- What characterizes the transport? Is it gravity-like, or...
- What characterizes the places where the neutrophils stay?
A simple enough 3-state model

State $X(t) \in \mathbb{R}^3$, noise $W(t) \in \mathbb{R}^3$, 

$$dX^{(S,R)}(t) = \sigma \, dW_t,$$

$$dX^{(T)}(t) = \nu(X^{(T)}) \, dt + \sigma \, dW_t,$$

$\sigma$ a noise scale, and $\nu$ a velocity.

For now, no particular mechanism to transfer between states $S \rightarrow T \rightarrow R$. 
Generalizable?
Embed in a larger model

Imaginable but unobserved internal state $S(t) \in \mathbb{R}^{\text{a lot}}$,

$$dS(t) = -\partial_S E(S, \mathcal{E}) \, dt + \xi \, dW_t,$$

$$dX(t) = f(S) \, dt + \sigma \, dW_t,$$

with $\mathcal{E}$ a representation of the environment ("cellular input"), and with $E$ the "cellular energy target".
Simple enough?

Inference

Because transporting neutrophils have different velocity, the speed $V := \|v\|$ is easier to infer over:

$$
E[\|X^{(S,R)}(t)\|^2] = 3\sigma^2 t, \\
E[\|X^{(T)}(t)\|^2] = V^2 t^2 + 3\sigma^2 t.
$$

In fact,

$$
P(\|X^{(S,R)}_t\|^2 = z) = \frac{z^{1/2} e^{-z/(2\sigma^2 t)}}{\sqrt{2\pi}(\sigma^2 t)^{3/2}}, \\
P(\|X^{(T)}_t\|^2 = z) = \frac{z^{1/4}}{2\sigma^2 t(V t)^{1/2}} e^{-\left(z+V^2 t^2\right)/(2\sigma^2 t)} I_{1/2}\left(\frac{V}{{\sigma^2} \cdot \sqrt{z}}\right).
$$
Expectation-Maximization iteration

1. E-step: assume that we know the probability (or responsibility, or mixture weight) that each one of the observations are, say, in mode $T$. Then we can compute estimates to the parameters $\sigma$ and $V$ by a weighted average.

2. M-step: assume that we know the parameters. Then we can compute the responsibility that each one of the observations are, say, in mode $T$. 

![Graph showing $\chi^2$ and Noncentral $\chi^2$ distributions]
Inference

Issues remaining

More development driven by data:

- In practise, the E-step is not closed and a fix-point iteration must be devised
- No particular mechanism to transfer between states $S \rightarrow T \rightarrow R$, hence consecutive estimates may flip back and forth (idea: penalty formulation)
3. Results & visualization

Convergence of EM

$\Rightarrow$ $(S, R)$-mode $\sim 65\%$, $T$-mode $\sim 35\% \ (n = 30)$. 
Convergence of EM (cont)
Back to data... (...then back again and repeat)
Hot-spots
Visualization

Issues remaining

- Selective visualization conditioned on estimated $(S, R, T)$-states — “inference-driven visualization”
- Use of “good” data (selected and controlled) and “bad” data (massive but with potentially large errors)
Summary

- “How to think”: realistic & useful models, via flexible/understandable/generalizable & simple enough
- 1. Data, tracking, & hypotheses (issues: artefacts)
- 2. Models & inference (issues: increased realism in inference)
- 3. Results & visualization (issues: “inference-driven visualization”)
Future

- Fall 2015: recruited a PhD student to project “Numerical modeling and simulation of living cells”
- Currently: remaining issues with Data, Inference, and Visualization...
- Next: finish ‘the’ manuscript, G. Christoffersson, S. Engblom, M. Phillipson: “Vascular sprouts independently induce local attraction of proangiogenic neutrophils”
- 2017: Joint VR-application (2nd attempt)
- 2017?: Joint KAW-application (2nd attempt)
Data is improving...
Data is improving...