Inference in biological physics
OMG, not **statistics**...

... Wake me up when it’s over.

Well... Statistical inference may sound “too theoretical,” but it is often needed to extract **information** from **data**:

[*] Sometimes suggests a new kind of measurement that tests a model more stringently, or distinguishes two different models more completely, than previous measurements.

[*] Sometimes our model is not obviously connected with what we can actually measure experimentally, and we need to **make a connection**.

[*] Sometimes the model that interests us involves the behavior of actors that we can only see indirectly in our data; theory may be needed to separate them out from each other, and from noise.

[*] If you like your science unified, you may feel that there’s got to be a **common thread** among that disconnected bag of tricks that people call “statistics.” That common thread is what I’ll call “inference.”
Part 1

1. Inference
2. Superresolution
3. Changepoint
4. Spike sorting

Conditional probability tells us what we can conclude from data, and we live in a world with boatloads of data, but conditional probability is not hardwired into our intuition, so we need to systematize it via the Bayes formula.
Suppose I stood here and said “all men are mortal; Socrates is mortal; therefore Socrates is a man.”

In classical logic it’s fairly easy to spot errors of inference. But what if I said “92.7% of all men are mortal...” Suddenly we find such questions tricky.
An everyday question in clinical practice

To diagnose colorectal cancer, the hemoccult test—among others—is conducted to detect occult blood in the stool. This test is used from a particular age on, but also in routine screening for early detection of colorectal cancer. Imagine you conduct a screening using the hemoccult test in a certain region. For symptom-free people over 50 years old who participate in screening using the hemoccult test, the following information is available for this region:

The probability that one of these people has colorectal cancer is 0.3 percent. If a person has colorectal cancer, the probability is 50 percent that he will have a positive hemoccult test. If a person does not have colorectal cancer, the probability is 3 percent that he will still have a positive hemoccult test. Imagine a person (over age 50, no symptoms) who has a positive hemoccult test in your screening. What is the probability that this person actually has colorectal cancer? ______ percent

Here are the replies of 24 practicing physicians, who had an average of 14 years of professional experience:
We are asked for $P(\text{sick} \mid +) = B/(B+D)$.
But what we were given was $P(+) \mid \text{sick}) = B/(A+B)$.

*These are not the same thing:* they have different denominators. To get one from the other we need some more information:

$$\frac{B}{B+D} = \frac{B}{A+B} \times \frac{A+B}{B+D}$$

$$P(\text{sick} \mid +) = P(+) \mid \text{sick}) \times \frac{P(\text{sick})}{P(+)}$$

Still need this

Posterior estimate (desired)
Likelihood (given, 50%)
Prior estimate (given, 0.3%)
In words

\[ P(X|\text{observed data}) = P(\text{data}|X) \frac{P(X)}{P(\text{data})} \]

“The probability that \(X\) is true given the data” is “The probability that the data you did observe would have been observed in a world where \(X\) is true” times “The prior probability of \(X\)” and “A normalization factor.”
Bayes Formula:

\[ P(\text{sick}|+) = P(+|\text{sick}) \times \frac{P(\text{sick})}{P(+)} \]

Is that last factor really important?

\( P(\text{sick}) \) was given, but we also need:

\[ P(+) = B + D \]

\[ = \frac{B}{A + B} (A + B) + \frac{D}{C + D} (C + D) \]

\[ = P(+|\text{sick}) P(\text{sick}) + P(+|\text{healthy}) P(\text{healthy}) \]

\[ = (0.5)(0.003) + (0.03)(0.997) \approx 0.03 \]

\[ \frac{P(\text{sick})}{P(+)} \approx \frac{0.003}{0.03} \approx 0.1 \]

Yes, it’s important: a positive test result means only a 5% chance you’re sick. Not 97%.
You can specifically label molecules of interest, and you can infer the center of an otherwise known distribution to high accuracy, but not if multiple light sources have overlapping diffraction spots, so you must switch on a subset of fluorescent tags, then another subset, etc.
Part 2: Superresolution microscopy

How does one measure myosin steps to within a few nm accuracy using visible light? The diffraction-limited spot is at least 200 nm wide! The key point is to realize that although we cannot resolve two spots closer than this, sometimes all we want is to detect motion of one spot.

Fluorescence Imaging at One Nanometer Accuracy...


F.I.O.N.A.
\[
\mathcal{P}(X|\text{observed data}) = \mathcal{P}(\text{data}|X) \frac{\mathcal{P}(X)}{\mathcal{P}(\text{data})}
\]

The posterior probability is

\[
\mathcal{P}(x_*|x_1, \ldots, x_M) = \text{const.} \times \frac{1}{\sqrt{2\pi \sigma^2}} e^{-(x_1-x_*)^2/(2\sigma^2)} \times \ldots \times \frac{1}{\sqrt{2\pi \sigma^2}} e^{-(x_M-x_*)^2/(2\sigma^2)}
\]

We wish to maximize this function over \( x_* \), holding \( \sigma \) and all the data \( \{x_1, \ldots, x_M\} \) fixed. The beauty of this approach is that it can be generalized to include more a accurate point-spread function, background, etc.

Its log is simple:

\[
\ln \mathcal{P}(x_* | x_1, \ldots, x_M) = \sum_{i=1}^{M} \left[ -\frac{1}{2} \ln(2\pi \sigma^2) - \frac{(x_i - x_*)^2}{2\sigma^2} \right].
\]
But usually we want an *image*, something a lot more structured than one point of light.

P. Nelson, *Physical models of living systems* (WH Freeman and Co 2015)

https://www.youtube.com/watch?v=RE70GuMCzww
Part 3

We’d like to know the spatial orientation of a molecule in real time, and polarized TIRF microscopy can deliver that information, but a cruel tradeoff must be made between orientation accuracy and time resolution, so we need to find the changepoints in order to optimize that tradeoff.

1. Inference
2. Superresolution
3. Changepoint
4. Spike sorting
Part 3: Changepoint analysis in single-molecule TIRF

JF Beausang, Yale Goldman, PN

✳ Sometimes our model is not obviously connected with what we can actually measure experimentally, and we need to make a connection.

✳ Sometimes the model that interests us involves the behavior of actors that we can only see indirectly in our data; theory may be needed to separate them out from each other, and from noise.

Many thanks to Haw Yang. See also Lucas P. Watkins and Haw Yang J. Phys. Chem. B 2005
Defects in myosin V are associated with human immunological and neurological disorders.

We’d like to know things like: How does it walk? What are the steps in the kinetic pathway? What is the geometry of each state? One classic approach is to monitor the position in space of a marker (e.g. a bead) attached to the motor. But this does not address the geometry of each state.

The approach I’ll discuss involves attaching a bifunctional fluorescent label to one lever arm. The label has a dipole moment whose orientation in space reflects that of the arm.
Myosin V stepping
Myosin V stepping
Myosin V stepping
To read out the orientation, we send in polarized light and see how many fluorescence photons, in each polarization, emerge.
Fluorescence illumination by the evanescent wave eliminates a lot of noise, and importantly, maintains the polarization of the incident light.

To tickle the fluorophore with every possible polarization, we need the incoming light to have at least two different beam directions.
pol-TIRF setup
pol-TIRF setup

Up to 8 incoming polarizations; 2 outgoing polarizations
Previous state of the art

For our purposes, the upshot is that: We need to know the arrival rates of photons in each of several channels. Once we’ve got that, then we can fit a physical model to determine the orientation of the molecule in space. Unfortunately, existing analyses gave noisy rate determinations. That in turn led to poor determinations of orientation—garbage in/garbage out.

Left: binned photon counts in 8 channels.
Right: Polar and azimuthal angles of the fluorescent label, inferred from data on the left. Noisy rate estimates lead to noisy position estimates. You could easily miss a short-lived state -- e.g. the elusive diffusive-search step (if it exists). Can we do better?

JN Forkey et al. Nature 2003
We know in general terms what is happening: Two Poisson processes, representing two of the photon channels, abruptly change rates at an unknown time. But only a few thousand photons were observed–makes it hard to estimate the “rate”:

So we have a tough choice:
- Shorter time bins give worse relative standard deviation for our rate estimate.
- Longer time bins degrade our ability to detect changepoints and to pinpoint their durations.

*Can we evade the cruel logic of photon statistics?* If only we could find the changepoints *first*, then use the *entire durations* between consecutive changepoints as our windows–the biggest choice possible! That would lead to the best possible estimate of photon rates, and hence the best possible estimate of orientation.

It’s a promising strategy because we know in general terms what is happening: A set of Poisson processes (blips in each photon channel), abruptly change rates at an unknown time. That is, we have some *prior knowledge* to exploit.
Here is some real data. For simplicity, we look at only two channels. Only a few thousand photons were observed.

When we classify the photons by polarization and bin them (here 20 bins were used), that reveals a definite changepoint. But when exactly did it occur? Probably not at the dashed line shown, but how can we be more precise?

Key point: *binning the data destroyed some information.* Something magical happens if instead of binning, we just plot photon arrival time versus photon sequence number. Despite some ripples from Poisson statistics, it’s obvious that each trace has a sharp changepoint, and moreover that the two changepoints found independently in this way are simultaneous. (A similar approach in the context of FRET was pioneered by Haw Yang.)

**JF Beausang, YE Goldman, and PCN, Meth. Enzymol. 487:431 (2011).**
Now that I have your attention

- **Why did that trick work?** How did we get such great time resolution from such cruddy data?
- **How well does it work?** If we have even fewer photons, for example because a state is short-lived, how can we quantify our confidence that any changepoint occurred at all?
- **Could we generalize and automate this trick?** Ultimately we’ll want to handle data with multiple polarizations, and find lots of changepoints.

Focus on just one “flavor” of photons (e.g. one polarization).

Suppose that in total time $T$ we catch $N$ photons at times $t_1, \ldots, t_N$.

We wish to explore the hypothesis that photons are arriving in a Poisson process with rate $R$ from time $0$ to time $t_*$, and thereafter arrive in another Poisson process with rate $R'$.

We want to find our best estimates of the three parameters $t_*$, $R$, and $R'$, and compare the null hypothesis that there was no change.

To do this, we again need the Bayes formula, which involves the “likelihood,” the probability that the data we actually observed would have been observed in a world described by our model with particular values of the unknown fit parameters:

$$
\log \mathcal{P}(t_1, \ldots, t_N|R, R', t_*) = \sum_{k=1}^{t_*/\Delta t} \log \left\{ \begin{array}{ll} R \Delta t & \text{if a photon in this slice} \\ (1 - R \Delta t) & \text{otherwise} \end{array} \right. + \sum_{k' = t_*/\Delta t + 1}^{T/\Delta t} \log \left\{ \begin{array}{ll} R' \Delta t & \text{if a photon in this slice} \\ (1 - R' \Delta t) & \text{otherwise} \end{array} \right.$$

From previous slide: In total time $T$ we catch $N$ photons at times $t_1,..., t_N$. Hypothesis is that photons are arriving in a Poisson process with rate $R$ from time 0 to time $t_*$, and thereafter arrive in another Poisson process with rate $R'$. 

$$\log \mathcal{P}(t_1, \ldots, t_N | R, R', t_*) = \sum_{k=1}^{t_*/\Delta t} \log \left\{ \frac{R \Delta t}{1 - R \Delta t} \right\} \text{ if a photon in this slice}$$
$$+ \sum_{k' = t_*/\Delta t + 1}^{T/\Delta t} \log \left\{ \frac{R' \Delta t}{1 - R' \Delta t} \right\} \text{ otherwise}$$

Now: Divide the $N$ photons into $n$ that arrived before the putative changepoint, and $n' = N - n$ that arrived after. Take the limit $\Delta t \to 0$:

$$\mathcal{P} \approx N \log(\Delta t) + n \log R + n' \log R' - \left( \frac{t_*}{\Delta t} - n \right) (R \Delta t) - \left( \frac{T - t_*}{\Delta t} - 1 - (N - n) \right) (R' \Delta t)$$

$$\approx \text{const} + n \log R + n' \log R' - Rt_* - R'(T - t_*)$$

Maximize this first over $R$ and $R'$:

$$R = \frac{n}{t_*}, \quad R' = \frac{n'}{(T - t_*)}$$

OK, duh, that was no surprise! But it does explain why we can just lay a ruler along the cumulative plot to get our best estimate of the before and after rates.

More interestingly, we can substitute these optimal rates into the formula for $\mathcal{P}$ to find the likelihood as a function of putative changepoint:
Here’s some very fake data; the photons arrive uniformly, not at random.

Here are two lines corresponding to non-optimal choices of the changepoint. We’d like to see the likelihood function and how it selects the “right” changepoint, which for fake data is known.

Our log-likelihood function has a huge peak as a function of putative changepoint time.
Left: More realistic, but still fake. data, shown in traditional binned form and in the improved version.

Below: Likelihood function for placement of the changepoint. Dashed line, maximum-likelihood point. Black triangle: Actual changepoint used to generate the simulated data. The analysis found a robust changepoint, even though there were a total of just 200 photons in the entire dataset.

Oh, yes--it also works on real experimental data, multiple-channel data, data with many different changepoints...

Previously, people would take data from multiple polarizations, bin it, and pipe the inferred intensities into a maximum-likelihood estimator of the orientation of the fluorophore.

That procedure leads to the rather noisy dots shown here. One problem is that if a transition happens in the middle of a time bin, then the inferred orientation in that time bin can be crazy.

Our approach first finds changepoints, shown as dashed lines.

Then the solid lines shown are the inferred orientations of the probe molecule during successive states defined by changepoint analysis. We see a nice alternating stride in $\phi$.

We got a 50-fold improvement in time resolution for finding changepoints, compared to the binning method, without changing the apparatus.

*JF Beausang, YE Goldman, and PCN, Meth. Enzymol. (2011); JF Beausang, DY Shroder, PCN, and YE Goldman, Biophys J (2013).*
When you only get a million photons, you’d better make every photon count.

A simple maximum-likelihood analysis accomplishes this.

In the context of TIRF it can dramatically improve the tradeoff between time resolution and accuracy.

That can help you find substeps, like the diffusive-search step in myosin-V’s kinetic scheme.

Part 4

1. Inference
2. Superresolution
3. Changepoint
4. Spike sorting

We’d like to know what neuron fired when, and retina has a 2D organization, ideal for making such recordings, but extracellular recording only gives a scrambled version of what we need, so we need to unscramble it before we can answer biological questions.
Part 4: Parallel recordings from dozens of individual neurons

✴ Sometimes our model is not obviously connected with what we can actually measure experimentally, and we need to make a connection.

✴ Sometimes the model that interests us involves the behavior of actors that we can only see indirectly in our data; theory may be needed to separate them out from each other, and from noise.
Retina is also an approachable, yet still complex, part of the brain. It’s a 2D carpet consisting of “only” three layers of neurons.

Visual scene in Retinal illumination pattern

Retina

Retinal ganglion cell spike trains

Brain

Behavior

Medium big picture
1. Experiment
2. Clustering
3. Fitting
4. Performance

Get data → Cluster → Fit → Interpret
Model organism

*Cavia porcellus.*

OK, mammals are harder than amphibians. But not that much harder.
Cf Meister, Pine, and Baylor 1994. Incredibly, one can keep a mammalian retina alive in a dish for over 6 hours while presenting it stimuli and recording its activity.
What's in the dish

67 ms of data, viewed as a movie.
[data taken at 10kHz have been smoothed. Biggest spikes about 400μV.]

Some spikes move across the array:

 Mostly we are hearing retinal ganglion cells, as desired, because they’re the ones that spike.

The spike-sorting problem is: Given raw data like these, convert to a list of discrete events (which cells fired at what times).

Big Data means Big Headaches. Here’s a tiny fragment of what comes out of a multielectrode array every few milliseconds.

Unfortunately many events are complex, with multiple overlapping spikes in many locations. And of course these may be the most interesting ones!

Our algorithm assessed “which neuron fired when” by using a **Bayesian inference** approach.

Many authors say *bursts* are a big problem, but here is a nice fit that we obtained with no special effort.

We even handle *overlapping spikes*, which some algorithms do not attempt.

---

*JS Prentice, J Homann, KD Simmons, G Tkacik, V Balasubramanian, PCN, PLoS ONE 6(7): e19884 (2011).*
Typical cluster

Superposing 50 traces chosen from 284 in this cluster shows that they really do all resemble each other.

Occasional events in which this event collides with another don’t affect the “archetype waveform” (template) (next slide).

Although the shape of each instance of the template is quite constant, still its amplitude has significant variation.

We scaled each instance of each template to get best agreement with the others, then took the median at each time point to find our best estimate of the consensus waveform (blue). As a check, the pointwise mean waveform looks the same (red). Resulting template waveform 3 ms
Suppose we measure some experimental data, and wish to make an inference about some situation that we cannot directly observe. That is, we imagine a variety of worlds with different values of $X$, and ask which is most probable given the observed data.

If we know the probability that those data would have arisen in a world with a particular value of $X$, then the Bayes formula gives us what we actually want:

$$P(X|\text{observed data}) = P(\text{data}|X) \frac{P(X)}{P(\text{data})}$$

We can ignore the denominator, if all we want is to compare two hypotheses (e.g. maximize over $X$).

For our application, we’d like $P(\text{spikes} | \text{data})$, where “data” is an observed waveform and “spikes” refers to a collection of spike templates $\mu_1, \ldots$ occurring at times $t_1, \ldots$ with amplitudes $A_1, \ldots$ relative to the amplitude of the corresponding template (neuron). Bayes’s formula gives what we want as

$$K \times (\text{likelihood}) \times (\text{prior}) = K P(\text{data} | \text{spikes}) P(\text{spikes})$$
Bayesian idea

Previous slide expressed \( P(\text{spikes} \mid \text{data}) \) as:

\[
K \times (\text{likelihood}) \times (\text{prior}) = K \ P(\text{data} \mid \text{spikes}) \ P(\text{spikes})
\]

Here “spikes” refers to a collection of spike templates \( \mu_1, \ldots \) occurring at times \( t_1, \ldots \) with amplitudes \( A_1, \ldots \) relative to the amplitude of the corresponding template.

To get the prior, \( P(\text{spikes}) \), assume that for a single spike it has the form

\[
P_{\text{cell}}(\mu) \ P_{\text{time}}(t) \ P_{\text{ampl}}(A \mid \mu)
\]

The three factors are respectively the popularity of this neuron, uniform in time, and a Gaussian reflecting its typical amplitude and amplitude variability. We get these priors from the data subset used in clustering.

To get the likelihood function \( P(\text{data} \mid \text{spikes}) \), suppose that the data consist of one template, plus noise. We measured the noise and found it to be Gaussian in character, and independent of which spikes fired.

Then the likelihood is that distribution, evaluated at the difference between the actual waveform and the idealized one. [Pouzat et. al. 2002]
Successfully fit overlaps

Closeup of four channels, showing four fit templates found by the algorithm.

Sum of those fits (color) versus actual data (black).
Successfully fit bursts

Even though successive spikes in a burst have different amplitudes, the algorithm fit them.
Theory can cut across apparently different kinds of experiment, offering useful methods to one domain from another without having to reinvent everything. Physicists are pretty good at this--especially as a part of a team involving life scientists. You do have to meet them halfway, but it’s worth it.
P(X|observed data) = P(data|X) \frac{P(X)}{P(data)}
Wait, there’s more

There is something weirdly -- unreasonably -- effective about approaching biological systems with a physical model. I don’t understand why. I don’t need to understand why.
Go long

“Overfitting: The biggest scientific problem you’ve never heard of.”
Thanks

University of Pennsylvania

NSF DMR, BIO, IBN

NSF NSEC

National Eye Institute Training grant

Computational Neuroscience Training grant
(Things I wish I’d read earlier)

M. Denny and S. Gaines
*Chance in biology*
Princeton Univ. Press 2000

D. S. Sivia and J. Skilling
*Data analysis: A Bayesian tutorial*
Oxford Univ. Press 2006

H. J. C. Berendsen
*Student’s guide to data and error analysis*
Cambridge Univ. Press 2011

Louis Lyons
*Discovery or fluke: statistics in particle physics*
Physics Today, 2012

P. Nelson
*From photon to neuron*
Princeton Univ Press

P. Nelson
*Physical models of living systems*
WH Freeman and Co.

P. Nelson and Tom Dodson
*Student’s guide to MATLAB for physical modeling*
Free at [www.physics.upenn.edu/biphys/PMLS](http://www.physics.upenn.edu/biphys/PMLS)

Jesse Kinder and P. Nelson
*Student’s guide to Python for physical modeling*
Princeton Univ Press
150x180µm recording spot
= 5x6 array of electrodes spaced 30µm (similar to RGC spacing).

[Data taken at 10kHz. Noise ~30µV. Big spikes ~400µV. Others go all the way down to the noise floor. Prior to analysis, filter out slow baseline drift. Also apply a spatial decorrelating filter, deduced from statistics of noise, to sharpen the “image” spatially.]