Hyperribbons and emergent models

Humans can comprehend low-dimensional systems. Once tens or hundreds of things are separately important at the same time, our intuition fails us. The fact that so many systems have comprehensible behavior is the foundation of science. Physics is founded on low-dimensional emergent laws that arise from small parameters — small frequencies and wavenumbers, temperatures low compared to bond strengths and ionization energies, … Science in other fields may be possible because they form hyperribbons.

Tradition: Start from experiment

Hyperribbons tell us a small number of measurements can characterize the system (span the long axes of the hyperribbon).
Basis for human intuition, wisdom and experience
Basis for superstition? Human pattern recognition?
Basis for ‘spherical cow’ physics modeling.

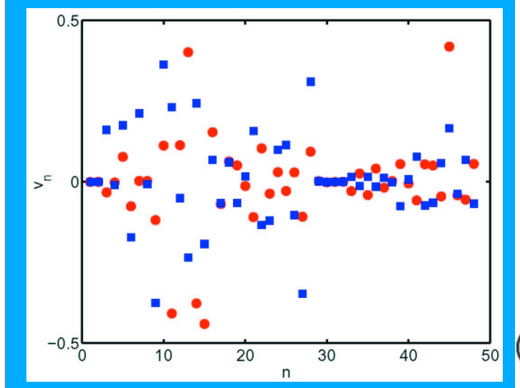
- Vary a few parameters (composition, annealing, deformation, radiation) [span hyperribbon long axes]
- Observe behavior, characterize with a few quantities (yield stress, fracture toughness, porosity)
- Derive intuitive explanation for link of behavior with parameters, parameter history); fit.

Starting from complex microscopic model

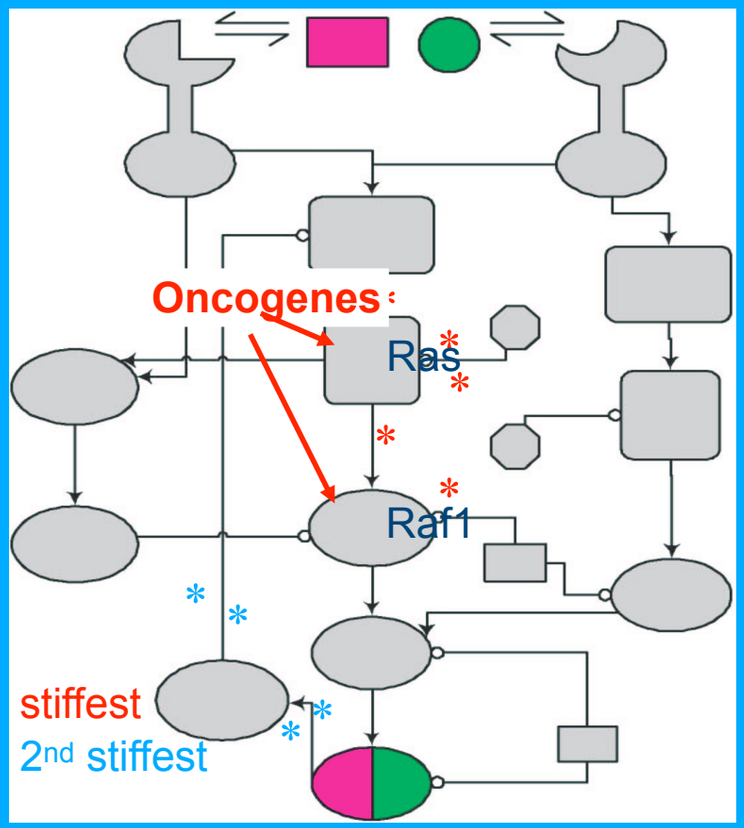
If a complex model has only five stiff directions, why not a simpler model with fewer parameters?

If most of the 48 dimensions of parameter space are sloppy, why not develop simpler models with fewer parameters?

Which Rate Constants are in the Stiffest Eigenvector?

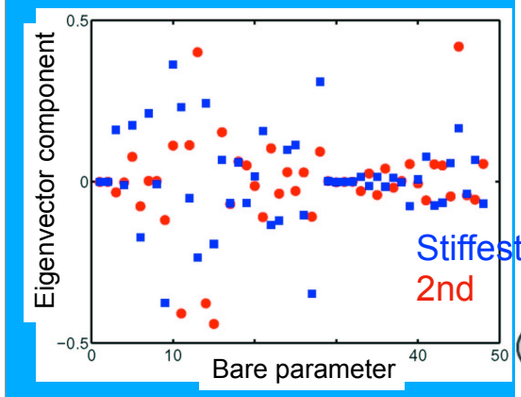


Eigenvector components along the bare parameters reveal which ones are most important for a given eigenvector.



We tried hard at the beginning. Our systems biology model was important biologically because two of the proteins (Ras and Raf) were products of known ‘oncogenes’ — genes which were often found mutated in cancer tumors. Our stiffest direction in parameter space (in red) did seem to correlate with reactions near the oncogenes. But it wasn’t compelling: things looked pretty random.

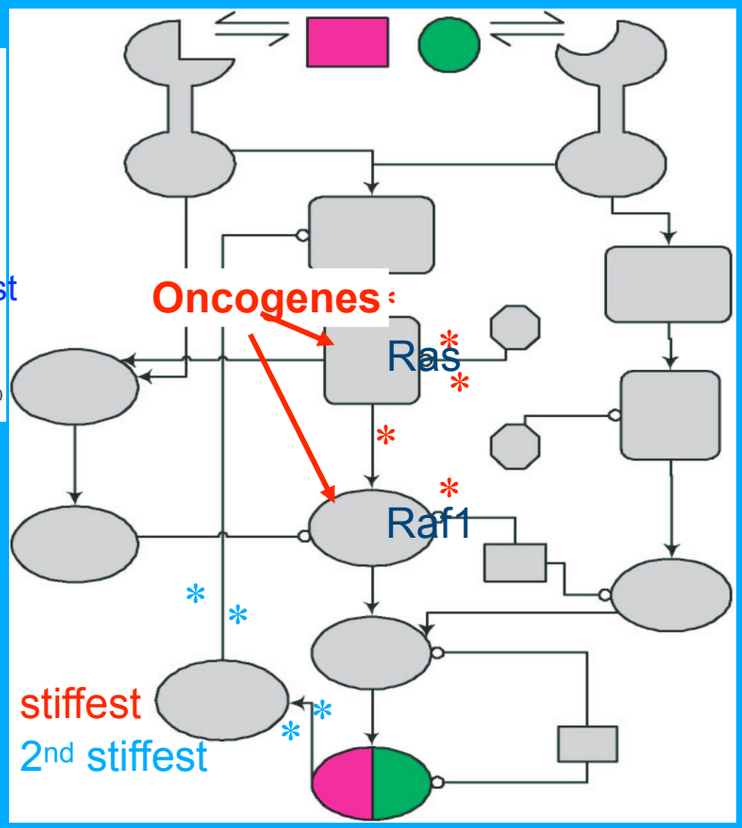
Simplify: Just keep the stiff ones?



Stiff eigendirections are weird, random-looking, & unintuitive.

Stiff + sloppy = Stiff.

No ‘perpendicular’ in parameter space.



We understand why the stiff eigendirections are usually incomprehensible. Moving along a direction given by a sum of the stiffest plus a sum of sloppy directions would naturally give the same change in behavior — the stiffest is special only because it is ‘shortest’ in parameter space: it minimizes the sum of squares of percentage changes in our reaction constants. This criterion is not physically sensible. Indeed, the only natural metric in parameter space g_{mu,nu}, that given by the change in behavior.

Removing redundant parameters?

Challenges in removing parameters.

- All parameters are important individually; only combinations might be removed.
- Even the stiff combinations are complicated.
- Experimentalists want to know what happens when we change each of the parameters.

Common shortcuts

- Search using a few ‘most sensitive’ parameters, according to $\mathscr{H}_{\alpha\alpha}$
- Assume shared values for similar parameters
- Take estimates for most from the literature, vary others

These likely work because the models are sloppy.

If most of the 48 dimensions of parameter space are sloppy, why not develop simpler models with fewer parameters?

Model boundaries!

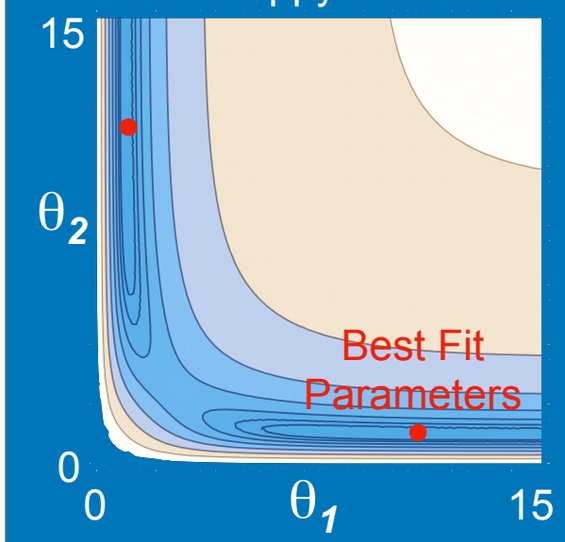
Transtrum's MBAM algorithm for deducing emergent, simpler models.

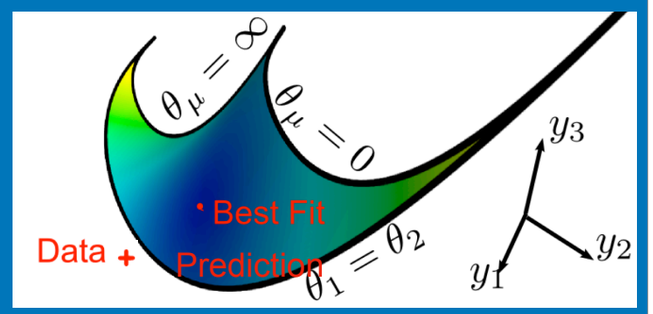
The Model Manifold: *Predictions*

Mark Transtrum
Ben Machta

Two exponentials θ_1,θ_2
$y_n = \exp(-\theta_1 t_n) + \exp(-\theta_2 t_n)$
fit to three data points at t_1,t_2,t_3

Parameter space
Stiff and sloppy directions





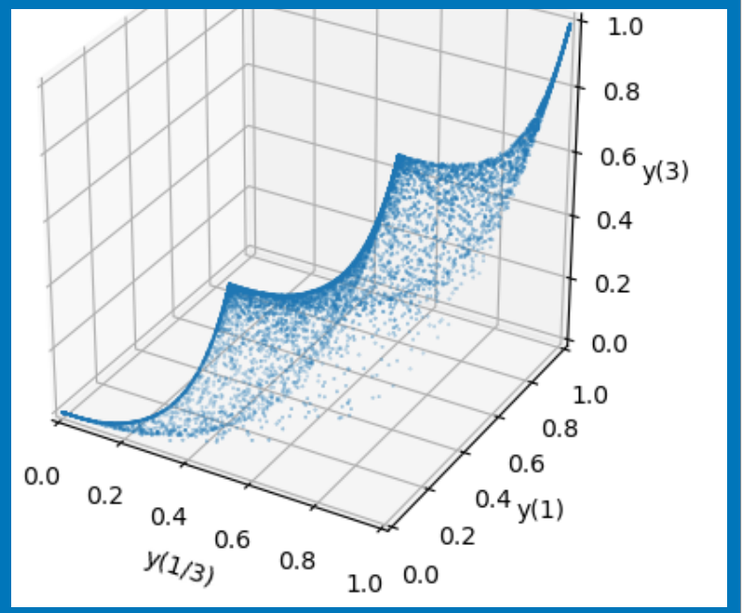
Behavior space
Manifold of model predictions
Parameters as coordinates
Model **boundaries** $\theta_\nu = 0,\infty,\theta_\mu$
are simpler models
Metric $g_{\mu\nu}$ from distance to data
Experimental datum **slices** manifold

Remember our model manifold for two exponentials. The edges and corners of the model manifold are simpler models, where there are fewer parameters. The corners have decay rates that are fixed to either zero or infinity (no free parameters), and the edges have one decay rate (the other zero, infinity, or equal to the first).

The Model Manifold

Three exponentials
fit to three data points at t_1, t_2, t_3

This is the model manifold for three exponentials, sampled at a number of values of $\theta_1, \theta_2, \& \theta_3$ (Exercise S3.1). Here the manifold fills a volume in the 3D space. Our sampling is not uniform in this volume (Jeffreys' prior), but is uniform in $\log \theta$ (as recommended by Jeffreys).

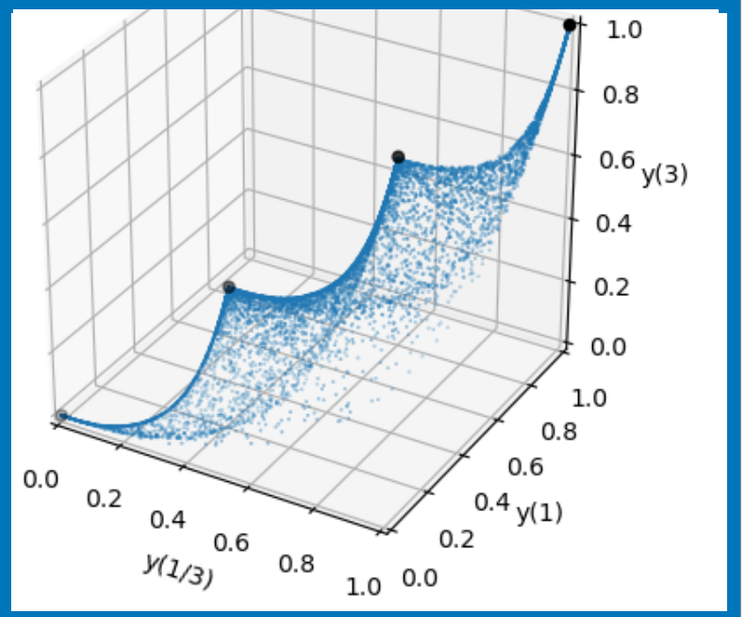


In the exercise "Plotting the model manifold", you will explore model manifolds for more exponentials. Here is a model manifold for the prediction space of the sum of three exponential decays, as it predicts the values at the same three times (t = 1/3, 1, 3). Since the model manifold fills a volume, we're filling it with random points. Shortly, we will discuss in depth different methods for doing this — the problem of priors.

The Model Manifold: *Corners*

Three exponentials
fit to three data points at t_1, t_2, t_3

The four corners of the model manifold are models with no free parameters (and thus zero-dimensional sets in behavior space.) Can you predict where the edges of the model manifold will be, from looking at the figure?

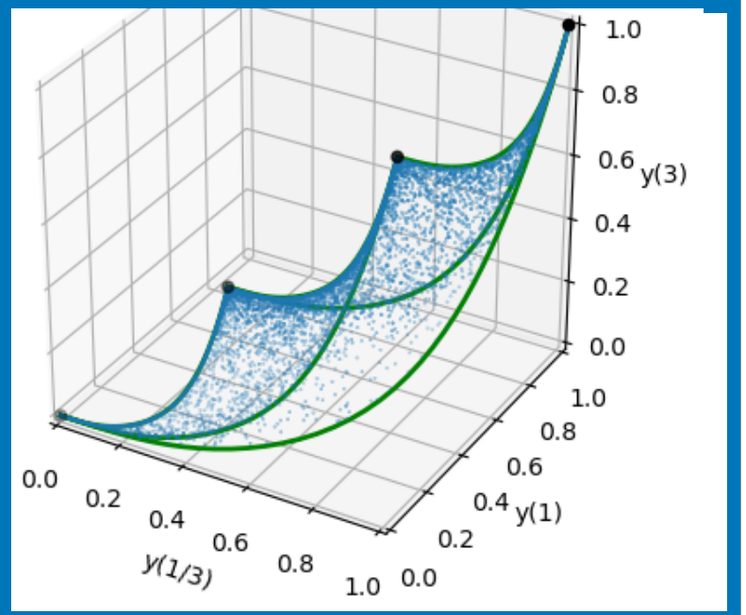


For three exponentials, there are four different predictions that are constant in time, with zero, one, two, or three of the decay rates equal to zero and the others infinite. These form the simplest models on the model manifold. You can see hints in the density of points about where the next-simplest models will lie. The lines where the density of points jumps are reminiscent of caustics, which could be places where the model manifold 'creases' as it is squashed into a 2D plot. Here they depict edges — cusps in the shape of the model manifold boundary.

The Model Manifold: *Edges*

Three exponentials
fit to three data points at t_1, t_2, t_3

The edges of the model manifold are models where there is only one exponential decay. Either two of the three decay rates are at the extremes (zero or infinity), or two or all three of the rates are the same. Can you tell which is which? Can you predict what the faces of the model manifold will look like?



These cusps correspond to curves along which only one decay rate describes the behavior. For a three-parameter model, there model boundary also has two-dimensional faces. For a 48 parameter model, there will be simpler models with 47, 46, … parameters, corresponding to hyper-edges — likely each with its cusp.

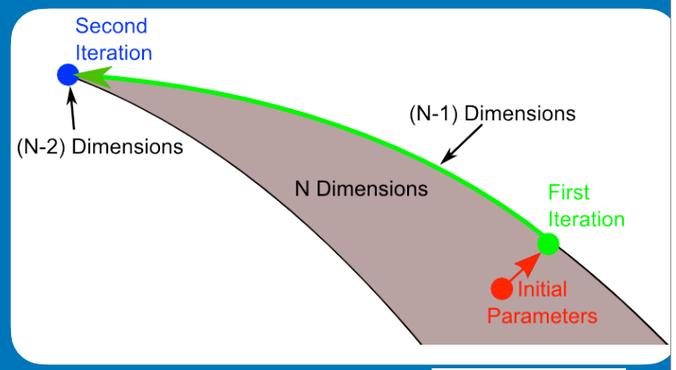
MBAM Generation of Reduced Models

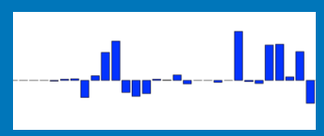
Mark Transtrum (not me)

Can we coarse-grain sloppy models? If most parameter directions are useless, why not remove some?
Transtrum has *systematic* "Model Boundary Approximation Method":

(1) Geodesic along sloppiest direction to nearby point on manifold boundary
(2) Eigendirection simplifies at model boundary to chemically reasonable simplified model

Coarse-graining = boundaries of model manifold.





Sloppiest Eigendirection

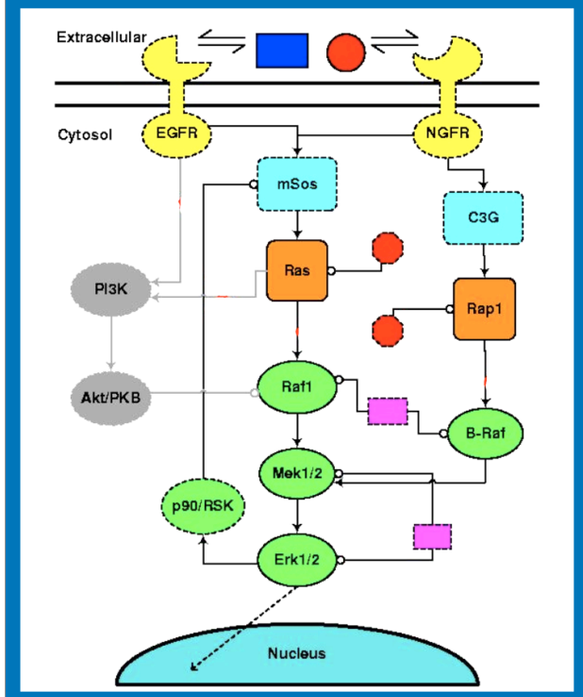
Simplified at Boundary
(Unsaturable reaction)

Mark Transtrum (I was not involved) developed what he calls the 'Model Boundary Approximation Method', which simplifies multiparameter models by geodesically following the sloppiest directions to the edge of the model manifold. At the edges, the

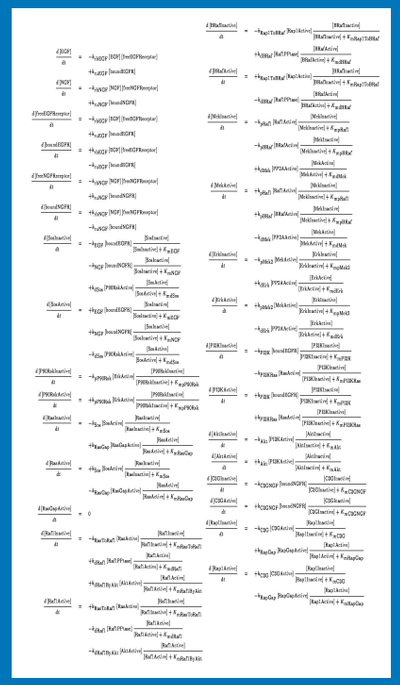
Let us examine Mark's application to our systems biology model.

MBAM Generation of Reduced Models

Mark Transtrum (not me)



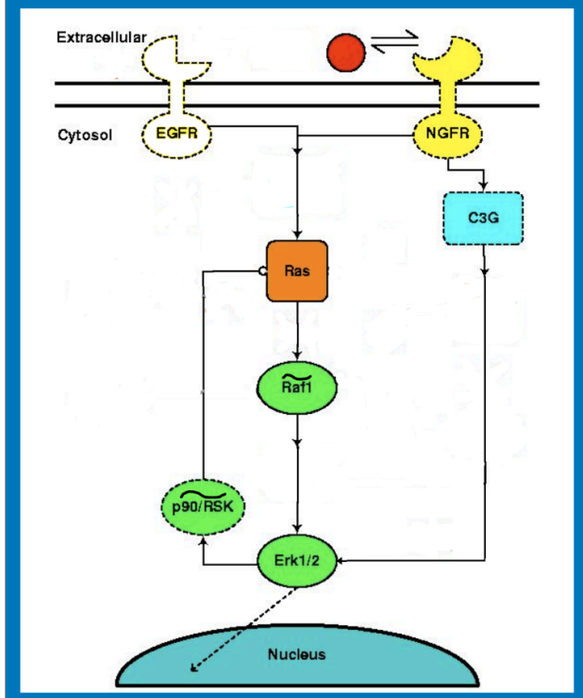
48 params
29 ODEs



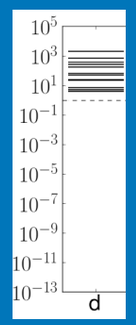
Mark applied this method to our original systems biology model. Remember that it began as a bewildering network with 48 parameters and a sloppy spectrum (spanning 15 orders of magnitude in eigenvalue, or 7 orders of magnitude in ‘length’).

MBAM Generation of Reduced Models

Mark Transtrum (not me)



12 params
6 ODEs



$$[\text{bEGFR}] = \begin{cases} 1 & \text{EGF Present} \\ 0 & \text{Otherwise} \end{cases}$$

$$\frac{d}{dt}[\text{bNGFR}] = \theta_1[\text{NGF}][\text{fNGFR}]$$

$$\frac{d}{dt}[\text{NGF}] = -\theta_1[\text{NGF}][\text{fNGFR}]$$

$$\frac{d}{dt}[\text{RasA}] = -[\text{RasA}][\widetilde{\text{P90RskA}}] + \theta_2[\text{bEGFR}] + \theta_3[\text{bNGFR}]$$

$$\frac{d}{dt}[\widetilde{\text{Raf1A}}] = \theta_4[\text{RasA}] - \theta_5[\widetilde{\text{Raf1A}}]/([\widetilde{\text{Raf1A}}] + \theta_6)$$

$$\frac{d}{dt}[\text{C3GA}] = \theta_7[\text{bNGFR}][\text{C3GI}]$$

$$[\text{Rap1A}] = \theta_8[\text{C3GA}]$$

$$[\text{MekA}] = [\widetilde{\text{Raf1A}}][\text{MekI}] + \theta_9[\text{Rap1A}]$$

$$\frac{d}{dt}[\text{Erk}] = -\theta_{10}[\text{ErkA}] + \theta_{11}[\text{MekA}][\text{ErkI}]$$

$$\frac{d}{dt}[\widetilde{\text{P90RskA}}] = \theta_{12}[\text{ErkA}]$$

Reduced model fits all experimental data

$$\theta_9 = \frac{[\text{BRafI}]\,\text{kRap1toBRaf}\,\text{KmdBRAF}\,\text{kpBRaf}\,\text{KmdMek}}{[\text{PP2AA}]\,[\text{Raf1PPtase}]\,\text{kdBRaf}\,\text{KmRap1toBRaf}\,\text{kdMek}}$$

Effective ‘renormalized’ params

After 36 parameters were removed by subsequent MBAM operations, we are left with a comprehensible network with one side loop and a negative feedback loop. The spectrum is no longer sloppy. Most important, MBAM preserves the dependence on all of the original parameters — writing the new effective ‘coarse-grained’ parameters as explicit formulas of the original ones.

We all value capturing the essential physics of a problem with the simplest possible model. Mark has applied this (among other things) to complex models used in drug design which the FDA wants to understand simply before they approve them, and models of the US electrical power system.

Priors for sloppy models

Mattingly, Transtrum, Abbott, Machta (not me)

Bayesian analysis relies on priors — an informed judgement about the relative likelihood of different model parameter sets. Can we generate a prior that satisfies Occam's razor — preferring the simpler models at the hyperedges and hypercorners of the model manifold? Can such a prior work be reparameterization invariant, and perform better than Jeffrey's prior, proportional to the square-root of the determinant of the metric?

Jeffreys prior is evil

Experiment observing x gives information about parameters θ. Want $p(\theta\,|\,x)$. Our model tells us $p(x\,|\,\theta)$. But $p(A\ \&\ B)/p(B) = p(A\,|\,B)$. Hence

Bayes' theorem: $p(\theta\,|\,x) = p(x\,|\,\theta)p(\theta)/p(x)$

Here $p(\theta)$ is the *prior* - the information you have from previous experiments.
But how do you start, without data? With an *uninformative* prior.

Letting $p(\theta) = 1$ is bad: if θ is a rate then high rates are favored, while if θ is a timescale then long times are favored.
Jeffreys suggested using the volume in behavior space near θ as the weight, $p_J(\theta) \propto \sqrt{\det g_{\alpha\beta}}$.

A typical Monte-Carlo move in our systems biology sloppy model changes $p_J(\Theta)$ by many orders of magnitude, so ignores $p(x\,|\,\theta)$ almost completely.

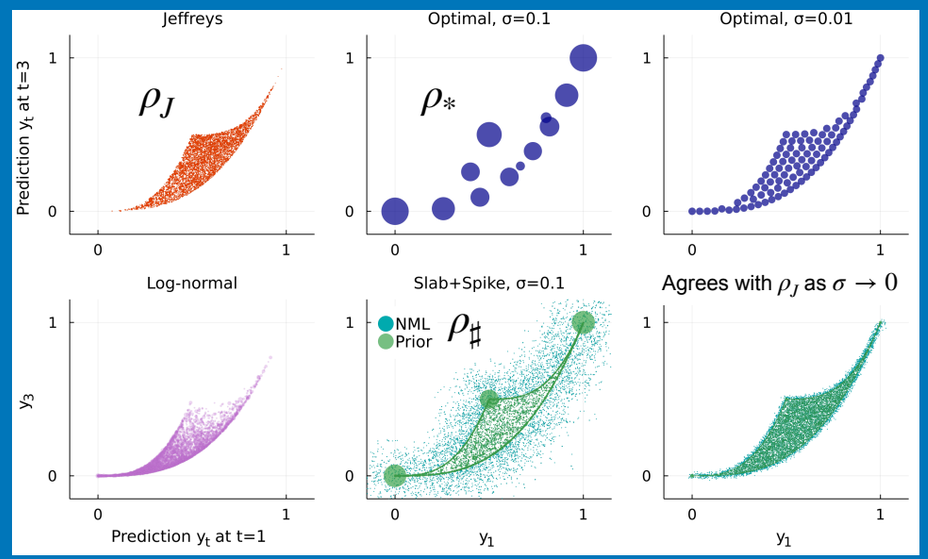
What is a prior? What is a good choice for an 'uninformative' prior if you don't have previous data or information? Why is Jeffrey's elegant choice doomed for our problems?

Occam's razor and model boundaries

Entia non sunt multiplicanda praeter necessitatem

Prefer the model that requires the fewest assumptions

Our model boundaries are simpler models. Want priors that concentrate at the boundary. Incorporate errors σ, imprecise data.

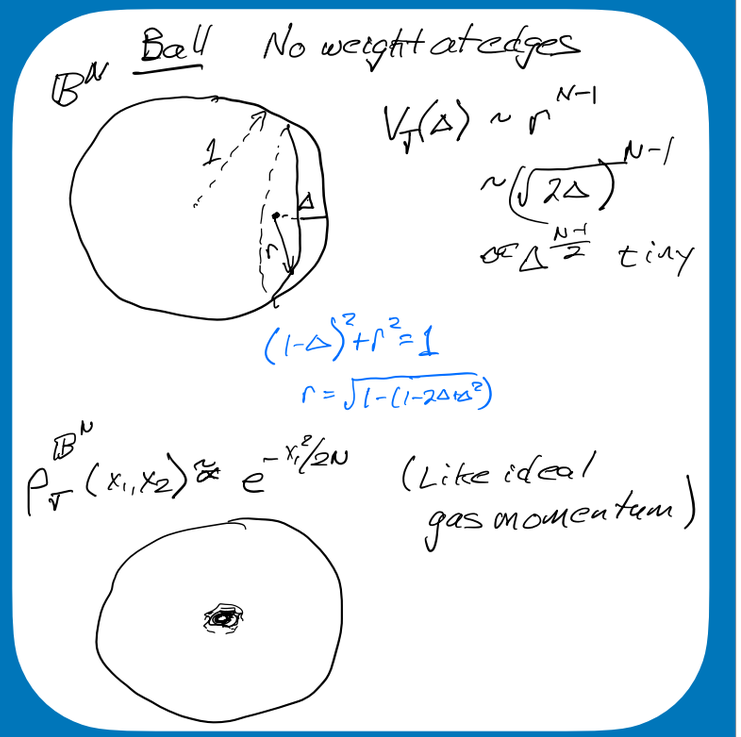


We've found that the edges of the model manifold are simpler models. If the simpler model fits the data just as well, wouldn't we rather use it? Occam in the 14th century is credited with this idea.

On the left are two standard priors, decorating the model manifold for a sum of two exponentials. On the right are their two new priors — the 'slab and spike' model that adds extra weight to the boundary points which are the best fit for noisy data, and one that optimizes the mutual information between the parameters and the data. Both have additional weight near the simpler models at the edges and corners.

Weirdness in high dimensions

Almost none of the Jeffrey's weight is near the (hyper-sharp) corners of the hyperribbon. Even for a ball $\mathbb{B}^N$, the volume at a distance Δ from its boundary goes as $\Delta^{(N-1)/2}$. A low-dimensional projection has all the weight of Jeffrey's prior at the center - where the model is thickest.

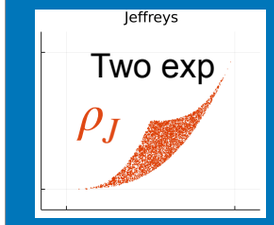


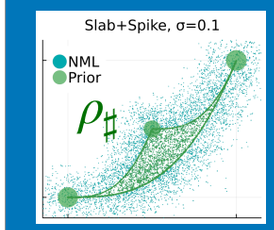
Jeffrey's prior not only doesn't add extra weight near the edges. For multiparameter models, it has essentially no weight near the edges!

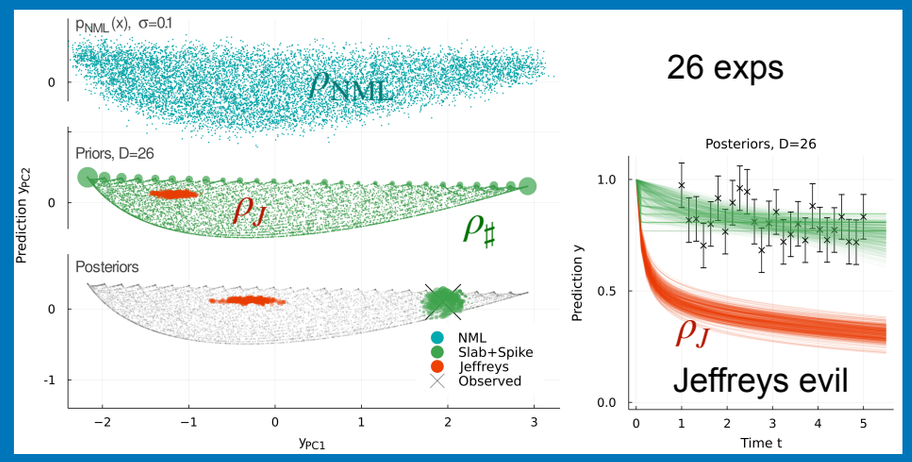
In high dimensions, there is no volume near the edges of the model manifold. Jeffrey's prior is concentrated in regions where the 'thin' directions of the model manifold are unusually thick. Here we see that the ball in N dimensions suffers from invisible edges. This gets much worse for a hyperribbon manifold with many thin directions.

Transtrum's slab-and-spike prior

Best fits of noisy data naturally concentrate at the edges







$p_{\mathrm{NML}}(\mathbf{x})$: Monte-Carlo in behavior space $\mathbf{x}$, sampled with max probability $p(\mathbf{x}\,|\,\Theta_{\max}, \sigma)$ best fit would give $\mathbf{x}$ with noise σ.

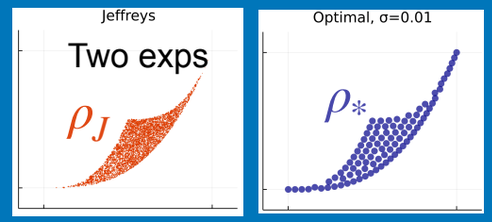
$p_\sharp(\Theta)$: Give all the probability to $\Theta_{\max}(\mathbf{x})$. Spikes at edges and corners.

The slab-and-spike prior starts by sampling points in behavior space, with a probability given by the distance to the best fit inside the manifold. Then every parameter Theta gets a prior weight given by the cloud of points for which it is the best fit!

See section 5.2 of
Quinn, Katherine N, Abbott, Michael C, Transtrum, Mark K, Machta, Benjamin B, and Sethna, James P (2022, dec). Information geometry for multiparameter models: new perspectives on the origin of simplicity. Reports on Progress in Physics, 86(3), 035901.

Mutual-information prior

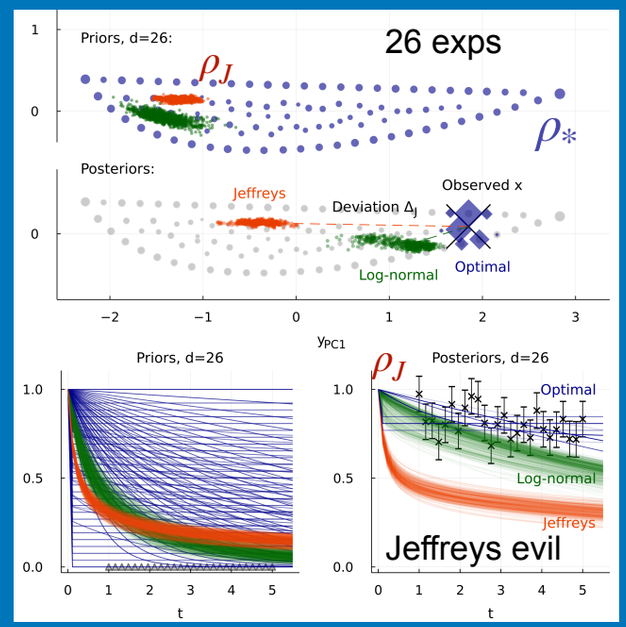
Maximizing reduction of parameter entropy after seeing data



Maximize mutual information

$$I(X;\Theta) = S(\Theta) - S(\Theta\,|\,X)$$
$$= \sum_x \int d\theta\, p(x\,|\,\theta)\, p(\theta)\, \log \frac{p(x\,|\,\theta)}{p(x)}\,.$$

Optimal but odd
Concentrated on points



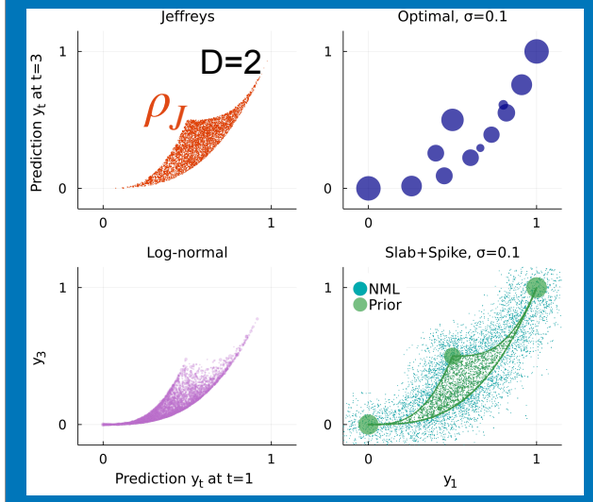
Mattingly et al. explored the prior that maximizes the amount of information gleaned by further measurements, adding the effects of finite data. As often happens in this kind of optimization, the best prior is on a finite set of points. This confuses the use of Bayes' theorem; one must redo the optimization to incorporate new information gleaned from further experiments.

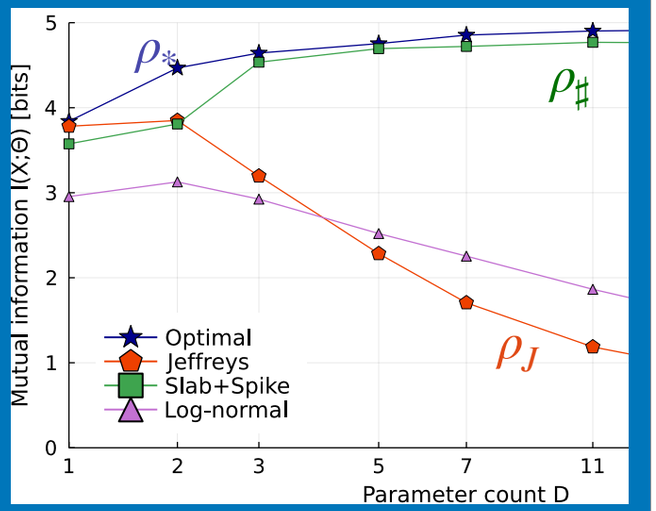
On the other hand, evaluating the probability for the prior and finding the 'best fit' is now easy — just compute the probability at a discrete set of points.

H. H. Mattingly, M. K. Transtrum, M. C. Abbott and B. B. Machta, Maximizing the information learned from finite data selects a simple model, PNAS 115 (2018) 1760–1765 [arXiv:1705.01166].

How optimal are they?

Using mutual information to evaluate priors





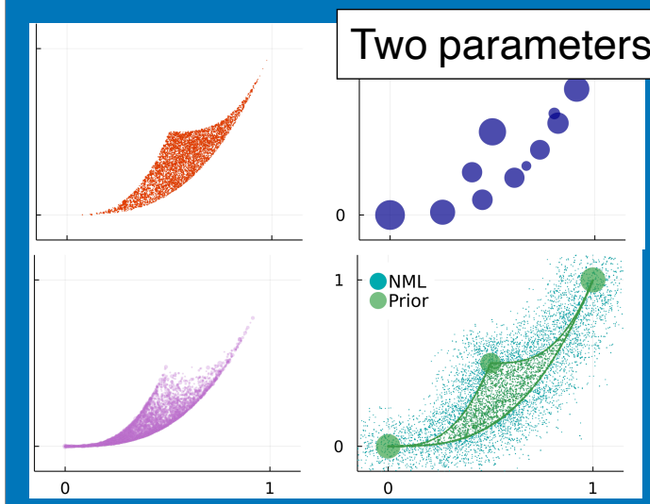
Transtrum's slab-and-spike model does amazingly well in high dimensions for our fitting-exponentials problem!

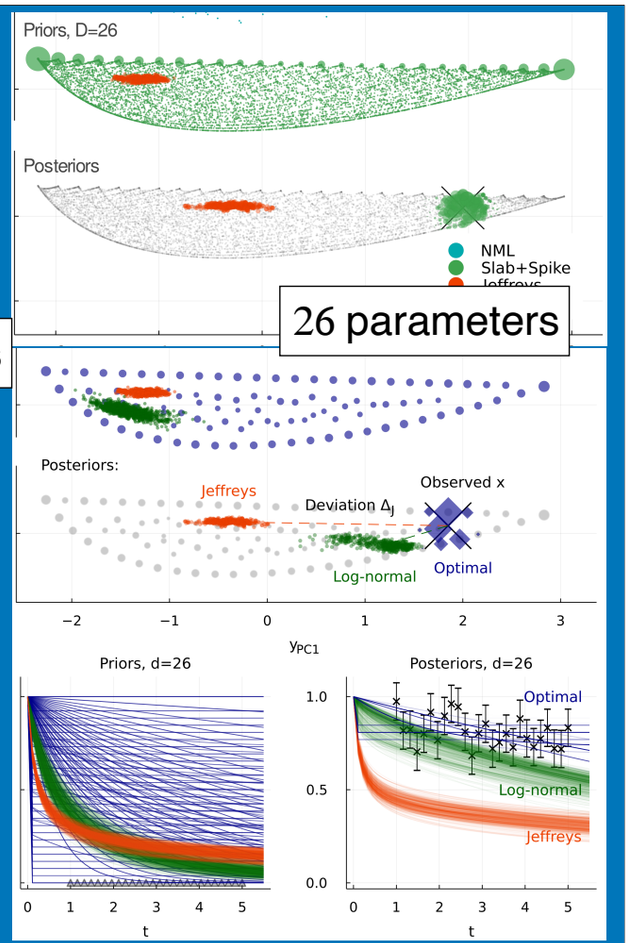
We can use the mutual information to measure the performance of different priors in comparison to the optimal prior of Mattingly et al. Transtrum's prior appears to do an excellent job, while the other priors perform poorly in high dimensions.

H. H. Mattingly, M. K. Transtrum, M. C. Abbott and B. B. Machta, Maximizing the information learned from finite data selects a simple model, PNAS 115 (2018) 1760–1765 [arXiv:1705.01166].

Jeffreys' Prior is Evil for sloppy models

Mattingly, Transtrum, Abbott, Machta (not me)





Jeffreys	Optimal
Lognormal	Slab-and-spike

Indeed, Jeffrey's prior suffers lethally of the curse of dimensionality. A uniform weight for all points on the model manifold works well in two dimensions, but all the weight of a hyperobject is near its center — massively distorting the Bayesian prediction for fits to the data. My colleagues have developed and explored two reparameterization-invariant priors — one which maximizes the mutual information (at the cost of being concentrated on a set of measure zero), and one they call a slab-and-spike prior which is nearly optimal and more conventional. Both give a weight to the simpler models on the boundary that is determined by the errors in the measurements being fit.