PLACATING PUGILISTIC PACHYDERMS PROPER PRIORS PREVENT POOR PERFORMANCE

Daniel Simpson
University of Toronoto
with Håvard Rue, Thiago Martins, Andrea Riebler, Sigrunn Sørbye

LONG AGO AND SO FAR AWAY

- Through the latter half of the 20th century Bayesian methods became a dominant force in applied and applicable statistics.
- Bayesian statistics provides a coherent way to update probabilities (or "belief statements") in the light of new data
- For a number of classical problems, Bayesian methods are eventually equivalent (with enough data) to the corresponding non-Bayesian/frequentist method
- The basic intuition is that If you have enough information about a parameter of inference, any sensible statistical method will work

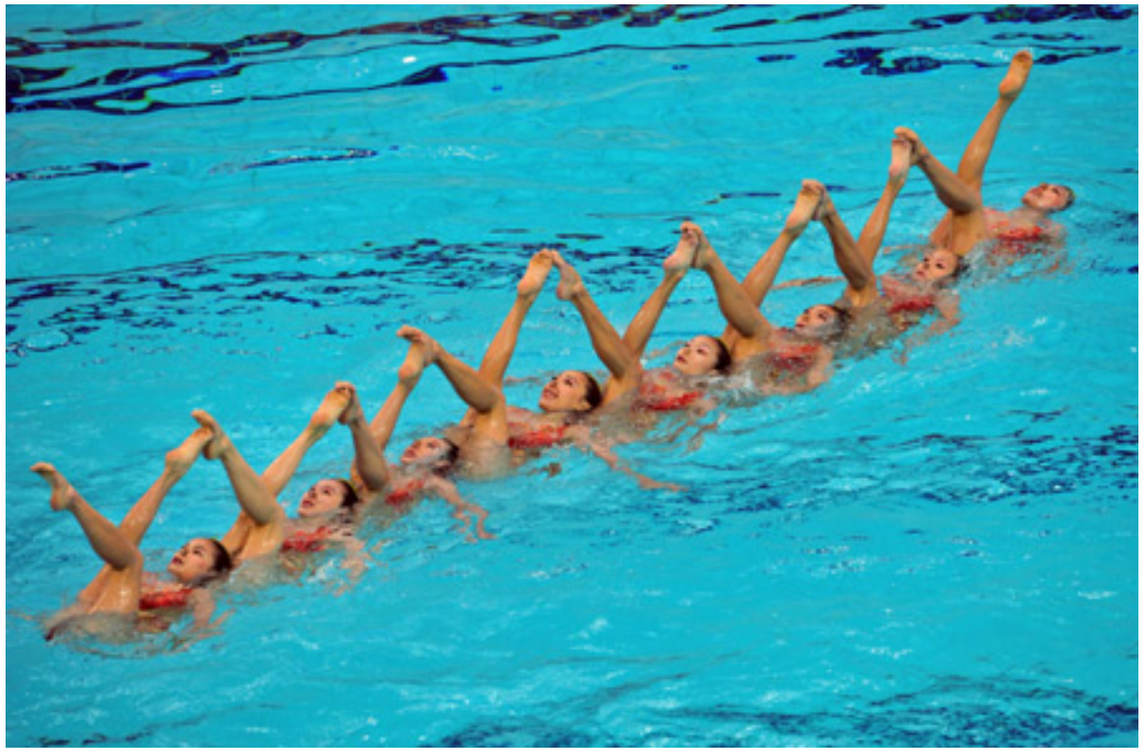
BEAST OF BURDEN

You should build your model as big as an elephant – Jimmie Savage

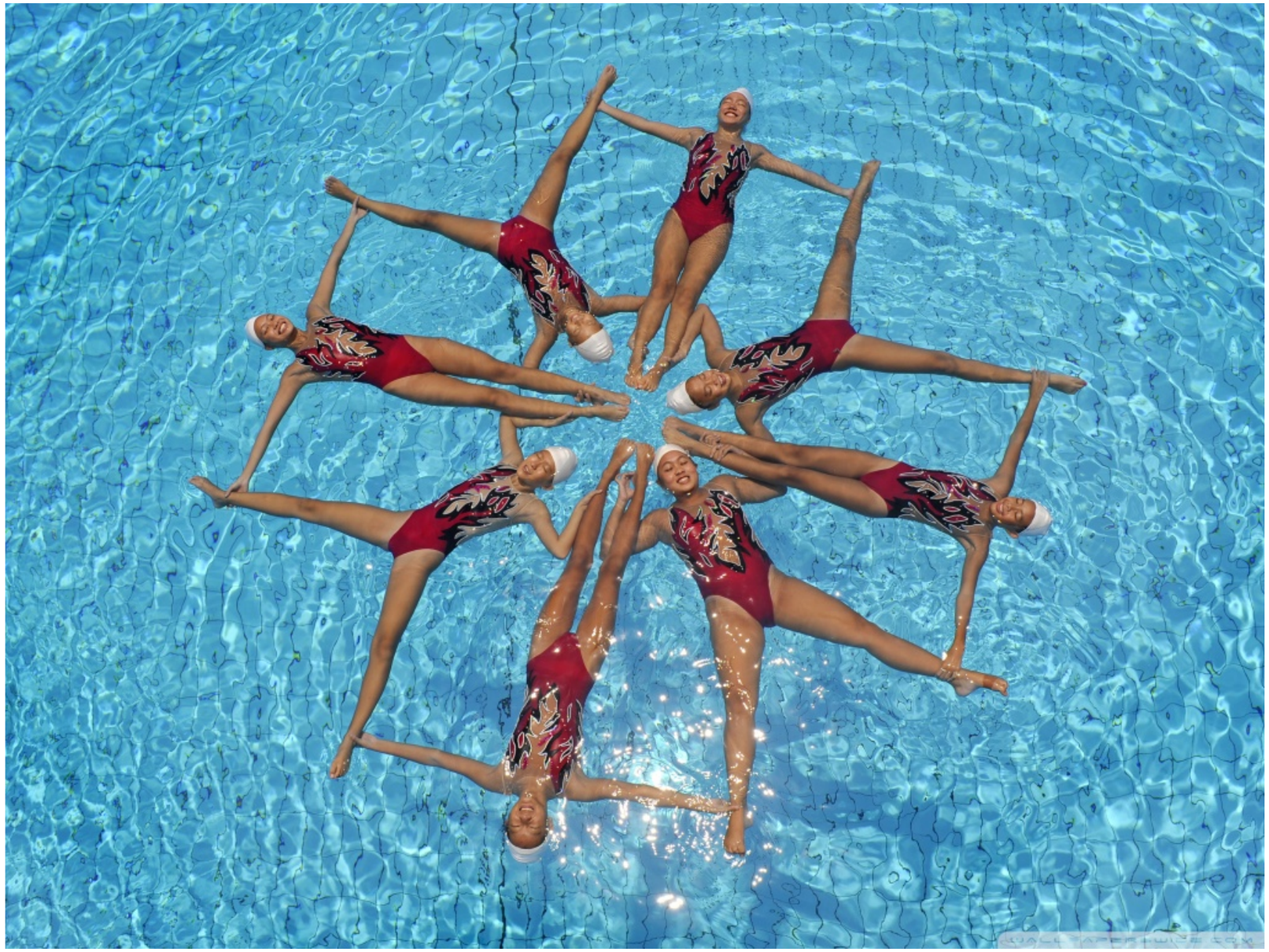
With four parameters I can fit an elephant, and with five I can make him wiggle his trunk. – John von Neuman

WE USED TO JUST ESTIMATE MEANS OF GAUSSIANS

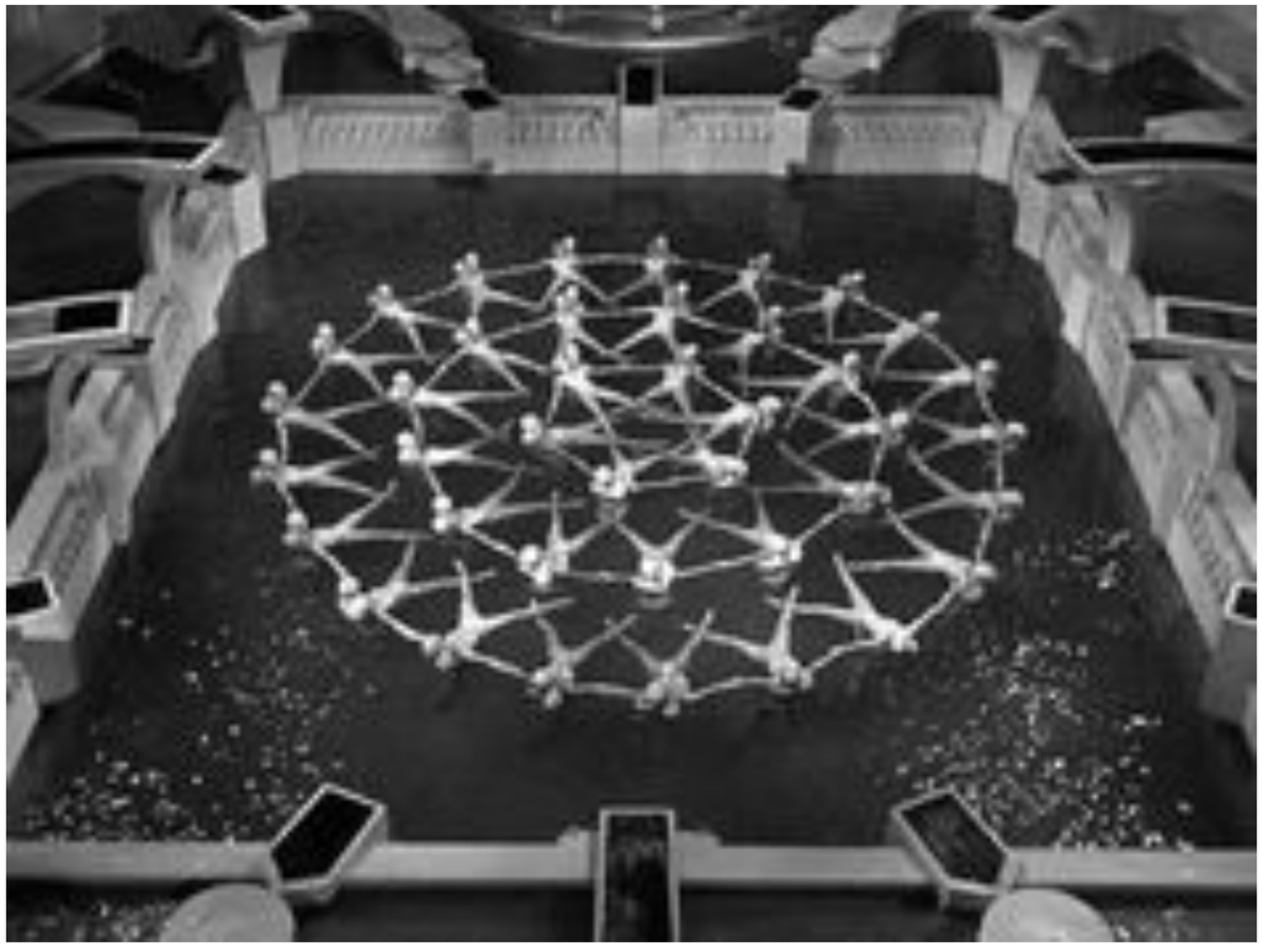
THEN THE MCMC REVOLUTION CHANGED EVERYTHING



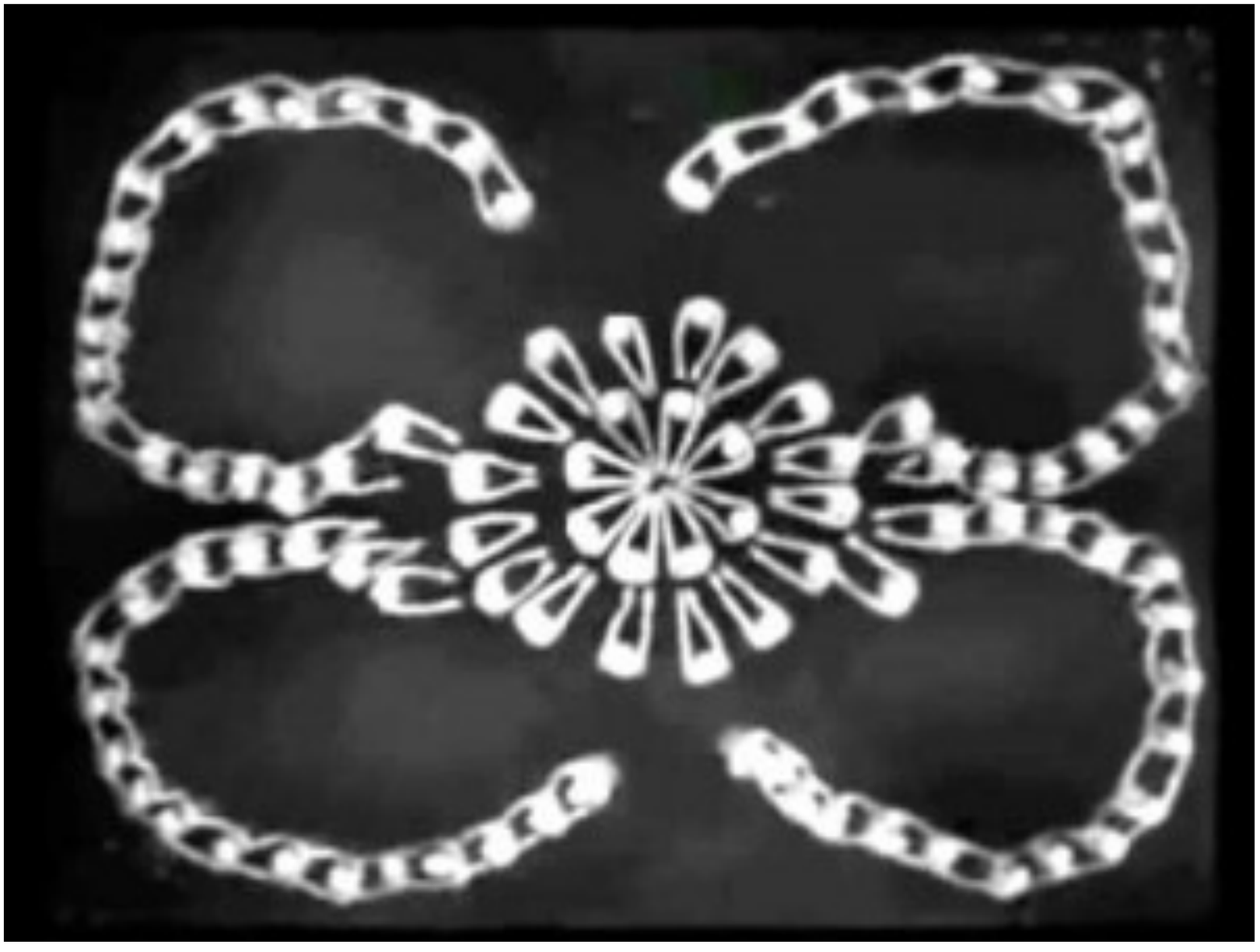
BUGS CAME ALONG AND REDEFINED THE POSSIBLE



METHODS LIKE INLA HELPED US SCALE UP



BUT THEN STAN CAME ALONG



IT'S A SMORGASBORD

- There's a whole smorgasbord of features of modern Bayesian models. Notably:
  - An overabundance of random effects
  - Multilevel models that borrow strength across different subpopulations to improve estimates
  - Correlated random effects, such as spatial or spatiotemporal random effects
  - Nonlinear effects of covariates (splines, splines, and more splines)
- With all these effects, it is not uncommon to have more parameters than data.

IF YOU STARE INTO THE PRIOR, THE PRIOR STARES BACK AT YOU

THE GANZFELD EFFECT

- Consider a mixed effects model with a random effect

$$u_j \sim N(\mu, \sigma_u^2)$$

- Why is it in the model?
- We add random effects to account for potential differences between groups.
- But what if there isn't a difference?
- We need to make sure we don't accidentally thrust a difference upon the data

I'M JUST A GIRL WHO CAIN'T SAY NO

- So how do we set a prior on the variance?
- Lots of choices in the literature.
- Breaks down basically into
  - "Uniformative priors" (try to put in no substantive information)
  - "Weakly informative priors" (Keep it sensible, stupid)
  - "Substantive priors" (Oh hi experts!)
  - "**shrug emoji** priors" (Well, someone else used it!)

DO WE EVER REALLY HAVE NO INFORMATION?

- Maybe? But often we have soft bounds.
- Estimating the number of people who committed tax fraud in Australia? Well the **maximum number** is less than 25 million.
- Estimating rainfall? You won't get a km of rain.
- But how does this filter into priors?

A SIMPLE MODEL FOR COUNT DATA

- Let's say we're modelling count data with the (cartoon!) model

$$y_{ij} \sim \text{Poisson}(\text{e}^{u_j})$$
$$u_j \sim N(\mu, \sigma_u^2)$$

- What should the prior for σ_u^2 be?
- Well, usually we'd want to conveniently store the observations on a computer, so a very very loose bound on the largest value y_{ij} can have is 2,147,483,648.
- If we want the **mean** of the Poisson to be less than this number, we need $\sigma_u < 7.2$.
- This suggests a prior like $\sigma_u \sim U(0,7.2)$ might work.

THIS ARGUMENT CAN BE EXTENDED TO MANY SITUATIONS

- Essentially, you can simulate from your model to see if your model is producing data that is wildly out of touch with reality.
- This can be used as a way to choose weakly informative priors (Wang, Nott, Drovndi, Mengersen, Evans (2018) refer to this as “history matching”)
- But it’s also a way to assess the priors that you have already chosen.

BUT WE KNOW MORE THAN JUST AN UPPER BOUND

BASIC INSTINCT

- I've already mentioned our other piece of information: it comes from the reason that the random effect is there in the first place.
- Mixed effects models have built into them the idea of a **base model**.
- In general, if we have a model component x with a distribution that is controlled by a **flexibility parameter** ξ, then the base model is the **simplest** model of the form $p(x \mid \xi)$
- We will always parameterize so that the base model occurs when $\xi = 0$

THE BIG CONCEPT

- We should only infer that $\xi > 0$ if the data **really needs it** to be bigger than zero.
- This is a version of Occam's razor: prefer simplicity over complexity.
- We can operationalize this idea by saying that the prior should have more mass **near** $\xi = 0$ than it has away from $\xi = 0$ and this decay should be, in some appropriate sense, monotone.
- A prior that doesn't put much mass near $\xi = 0$ will **overfit** the data

DENSITIES ARE ANNOYING

- But that's a statement about probability mass, but we really only tend to work with densities.
- So how can we actually apply this principle?
- **Big idea:** Choose a new parameterization $d = d(\xi)$ such that $d(0) = 0$ and Occam's razor (loosely) holds if $p(d)$ has a mode at zero and decays monotonically as d increases.
- We can then say loosely say that a prior **overfits** the data if $p(d(\xi) = 0) = 0$ (and, if we want to be more mathematical, it goes to zero rapidly near zero)

HOW WILL I KNOW (IF HE REALLY LOVES ME?)

- So what should this parameterization be?
- Well we need it to be computable, but also to naturally measure the increasing complexity as $d(\xi)$ increases.
- The natural measure of the difference in complexity between two distributions is the Kullback-Leibler divergence between the **flexible** model f and the **base** model b

$$KLD(f||b) = \int f(t)\log\left(\frac{f(t)}{b(t)}\right)\,dt$$

- This measures the information lost when f is replaced by b

THE MATHS GETS IN YOUR EYES

- It turns out that using the KL divergence directly actually isn't the most sensible thing.
- Why? Because it looks more like the square of a distance than a distance itself (see either the small ball limit of a squared Fisher distance, or Pinsker's inequality).
- So an actually good re-parameterization is

$$d(\xi) = \sqrt{2KLD\left(p(x \mid \xi) || p(x \mid \xi = 0)\right)}$$

CONSTANT RATE PENALIZATION

- Lacking any other knowledge of this parameter, it makes sense for the prior to decay at a constant rate.
- This means we want to use an exponential prior

$$p(d) = \lambda \exp(-\lambda d)$$

or

$$p(\xi) = \lambda \exp(-\lambda d(\xi)) \left| \frac{dd}{d\xi} \right|$$

- But how do we choose λ?

THIS IS WHERE WE NEED SOME EXPERT KNOWLEDGE

- We need two things:

  1. A substantively interpretable quantity $Q(\xi)$

  2. A value that is large (or small) for Q

- We can put these together to choose λ through the condition

$$\Pr(Q(\xi) > U) = \alpha$$

  for some small α.

- (Clear link to the history matching idea here)

PENALIZED COMPLEXITY PRIORS

- We call these priors **Penalized Complexity Priors** or **PC Priors**.
- They behave nicely in all of the practical situations we have used them in.
- Partly it is because they are built up from **four principles**:

1. Occam's Razor
2. Parameterize to measure complexity
3. Penalize complexity at a constant rate
4. Get the user to define the scaling

WHY ARE PRINCIPLES USEFUL (EVEN IF THEY AREN'T UNIQUE)

- Because they encode where the information in the prior comes from.
- Because they can be taken and examined individually to see if they make sense.
- Because they can help you to communicate the model assumptions with stakeholders
- Because it sounds fancy.

BACK TO OUR OLD FRIEND

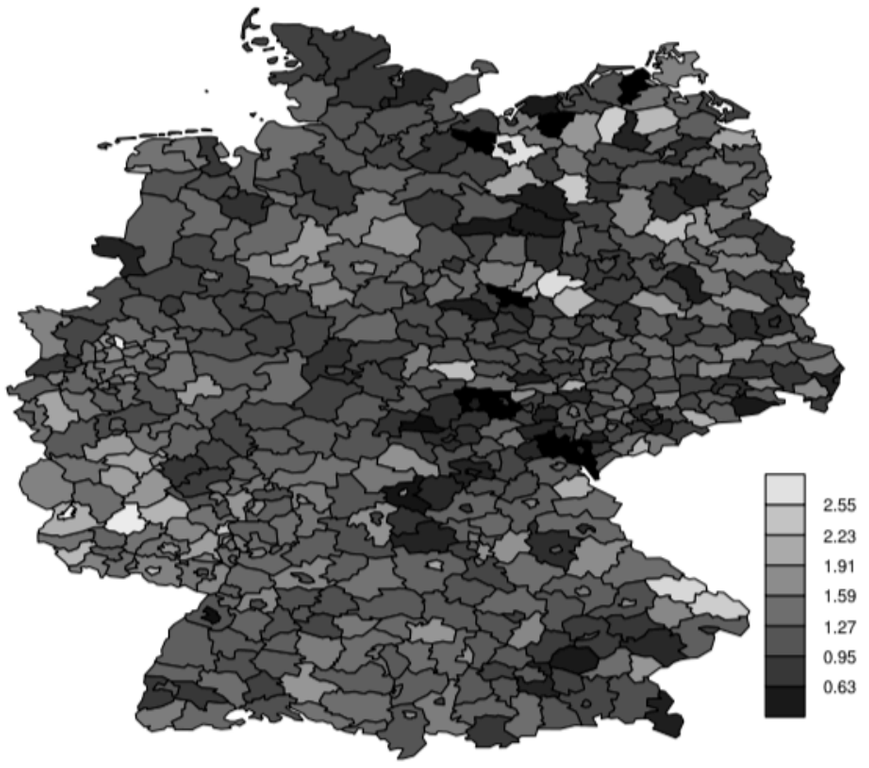
- So what is the appropriate prior for variance of a Gaussian random effect?
- Well, it's easy to show that

$$KLD\left[N(0,\sigma^2)||N(0,\epsilon^2)\right] = \frac{\sigma^2}{2} + O(\epsilon^2)$$

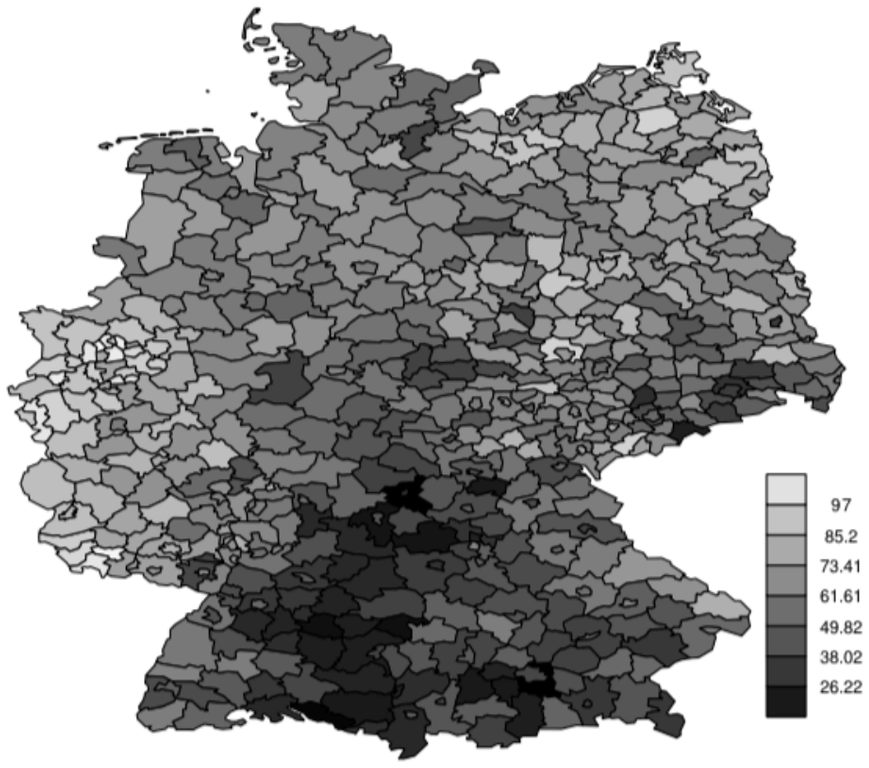
- This suggests that $d(\sigma^2) = \sigma$
- There is a lot of common sense here: we can reason about standard deviations easily.
- So the PC prior for a variance parameter in a Gaussian random effect is an exponential prior on the standard deviation

WHAT IF THERE'S MORE THAN ONE PARAMETER?

YES BUT WHY DO I CARE?



Incidence of larynx cancer



Smoking rates

How would we model risk?

A BASIC MODEL FOR ESTIMATING DISEASE COUNTS

$$Counts_i \sim Poisson(\lambda_i)$$

$$\log(\lambda_i) = \log(Expected\ Counts) + u_i + v_i$$

Relative risk!

Overdispersion
$u \sim N(0,\sigma_u^2)$

Spatial correlation
$v \sim N(0,\sigma_v^2)$

I don't know how to deal with all of this stuff together

WHAT'S THE PROBLEM?

- Well u and v are both adding variance to the model
- And, from what we already know, we want to control the **total variance** of the random effect.
- But right now this is controlled by two different parameters σ_u and σ_v
- We need to reparameterize!

LEVELS OF COMPLEXITY

- There's a natural hierarchy of complexity for this model
- No variability → iid random effect → spatial random effect
- We can reflect that in the parameterization:

$$u + v = \sigma_{re}\left(\sqrt{1-\gamma}v^* + \sqrt{\gamma}u^*\right)$$

- Here v^*, u^* are unit variance random effects
- We now have two parameters σ_{re} controls the overall scale of the random effect (base model 0)
- γ controls the proportion of the variance attributed to the spatial effect (base model 0)
- They do **different things** so independent priors make sense!

WHAT DOES THE PRIOR ON THE MIXING PARAMETER LOOK LIKE

What does the PC prior on γ look like?

- The covariance matrix is $\boldsymbol{\Sigma}(\gamma) = \gamma \boldsymbol{I} + (1-\gamma)\boldsymbol{R}^{-1}$
- The squared distance is then

$$d(\gamma)^2 = n\gamma\left(\frac{1}{n}\operatorname{tr}(\boldsymbol{R}^{-1}) - 1\right) - \log\left|(1-\gamma)\boldsymbol{I} + \gamma\boldsymbol{R}^{-1}\right|$$

- For sparse $\boldsymbol{R}$, the trace is easy to compute, and the evaluation costs one sparse Cholesky decomposition
- The PC prior is then

$$\pi(\gamma) = \frac{\lambda \exp(-\lambda d(\gamma))}{1-\exp(-\lambda d(1))}\left|\frac{\partial d(\gamma)}{\partial \gamma}\right|.$$

- (NB: $d(1)$ is finite, and so we use a truncated exponential!)

PC PRIORS ARE SOMETIMES JOINT

WHAT IF SPACE IS CONTINUOUS

- I'm a mathematician by training and disposition, so I'm going to ignore all of the interesting bits of that model and just focus on the maths bit!
- So let's just focus on GP regression
- Because I **know** somethings about GP regression
- Namely, I know that I can consistently estimate all of the parameters, so the MCMC should be fine.

$$y_i = f(x_i) + \epsilon_i$$

iid Gaussian

WHAT IS A GAUSSIAN PROCESS?

- A Gaussian process is a random function f with the property that if we evaluate it at a finite set of points, then the joint distribution of its values is always multivariate Gaussian.

$$[f(s_1), f(s_2), \ldots, f(s_m)]^T \sim N(0, \Sigma)$$

- This means we can easily work with point observations as long as we can specify a way to build the covariance matrix.
- Clever people realized that we can build it entrywise as $\Sigma_{ij} = c(s_i, s_j)$ for some **positive definite covariance function** c.
- There are well studied families of parameterized covariance functions we can use

EVERYTHING IS PROBABLY NOT GOING TO BE OK

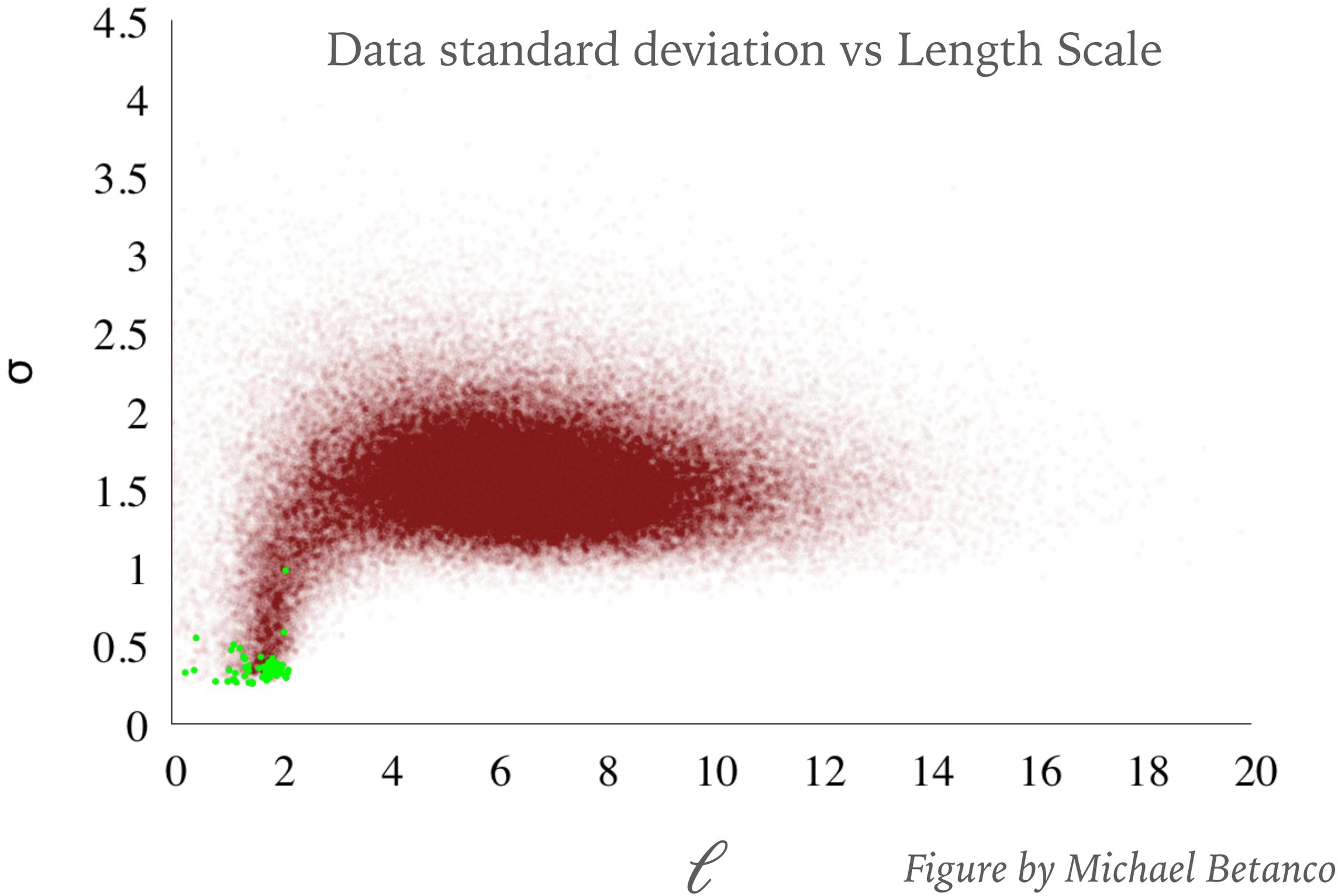


Figure by Michael Betancourt

THE MATÉRN COVARIANCE FUNCTION

- A common class of covariance function is the Matérn covariance function

$$c(s_1, s_2) \propto \sigma^2 \left(\kappa \|s_1 - s_2\|^2\right)^{\nu} K_{\nu} \left(\kappa \|s_1 - s_2\|^2\right)$$

- Here ν is a smoothness parameter that we will fix
- κ controls the bandwidth
- σ^2 is a scale parameter
- (K_{ν} is the modified Bessel function of the second kind)
- **So what prior should we put on κ and σ**

WHAT ARE WE OBSERVING

- Our problem is that we've only defined the GP through what happens when you observe it.
- So either our PC prior will be very design dependent or we will have to do maths.
- We opted for the latter, and examined what the distance should be if we saw an extremely dense set of observations.
- This turned out to be a good approximation to the design-based PC prior.
- But there were some problems

MATHS IS HARD

- It turns out the the KL divergence is frequently infinite when dealing with non-parametric models.
- We got around this by using the parameterization

$$(\kappa, \tau^2) = (\kappa, \kappa^\nu \sigma^2)$$

- With τ fixed, $d(\kappa) = \kappa^{D/2}$.
- With κ fixed, τ is just a scaling parameter and we can use our previous result to say that the PC prior will be an exponential.

BUT WHAT ABOUT THE INTERPRETABLE QUANTITY?

- For κ it turns out to be related to the range and a direct application of the PC prior principles says that in 2D is exponential
- For τ the interpretable quantity is $\sigma = \kappa^{\nu/2}\tau$, which depends on κ
- This means that we get a joint PC prior rather than the product of two independent priors.
- It turns out that if you transform this back to the orignal prameterization, you get independent priors (an inverse gamma on the range, and an exponential on the standard deviation).

SLOUCHING TOWARDS A CONCLUSION

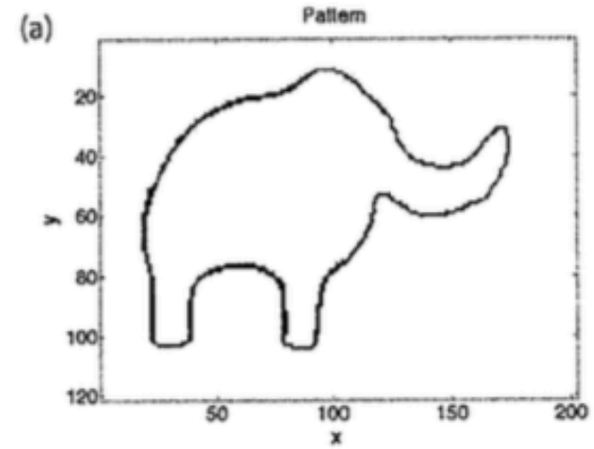
IF LOVE WERE ALL

This example shows just a corner of the power of PC priors

- Splines
- Skew-Gaussian distributions
- Correlation matrices
- AR(p)
- Over-dispersion in Negative Binomials
- Hurst Parameters for fractional Brownian motion
- Degrees of freedom in a Student-t
- Non-stationary GRFs
- Correlated random effects
- Variances in multilevel models
- +++

PLACATING PUGILISTIC PACHYDERMS

- Setting priors is hard
- Bayesian models should not be used / interpreted unless you can interpret all levels of your model (including your prior)
- We need to match the ambition and complexity of the applied modellers
- Otherwise, instead of giving them enough rope to hang themselves, we are cutting out the middle man



Mayer, Khairy, and Howard,
Am. J. Phys. 78, 648 (2010)

REFERENCES

- Daniel Simpson, Håvard Rue, Thiago Martins, Andrea Riebler, and Sigrunn Sørbye,. (2017). Penalising model component complexity: A principled, practical approach to constructing priors (with Discussion). Statistical Science. 32(1): 1-28.
- Geir-Arne Fuglstad, Daniel Simpson, Finn Lindgren, and Håvard Rue. (2018). Constructing Priors that Penalize the Complexity of Gaussian Random Fields. Journal of the American Statistical Association. Volume 114(525), pp. 445–452.