Placating pugilistic pachyderms

Proper priors prevent poor performance

Daniel Simpson
Håvard Rue, Thiago Martins, Andrea Riebler, Sigrunn Sørbye,
Charisse Farr, Kerrie Mengersen

Department of Mathematical Sciences
University of Bath

Outline

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**
- The interesting problems occur when there is *not* an over-abundance of information.

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**
- The interesting problems occur when there is *not* an over-abundance of information.

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**
- The interesting problems occur when there is *not* an over-abundance of information.

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**
- The interesting problems occur when there is *not* an over-abundance of information.

Beast of Burden

A Savage Quotation

You should build your model as big as an elephant

A von Neumann quote

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

Beast of Burden

A Savage Quotation

You should build your model as big as an elephant

A von Neumann quote

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

Come to the supermarket (in old Peking)

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.

(In fact, it's not uncommon to have **several infinite dimensional** parameters!)

Come to the supermarket (in old Peking)

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.

(In fact, it's not uncommon to have **several infinite dimensional** parameters!)

Come to the supermarket (in old Peking)

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.

(In fact, it's not uncommon to have **several infinite dimensional** parameters!)

Busby Berkleey Dreams

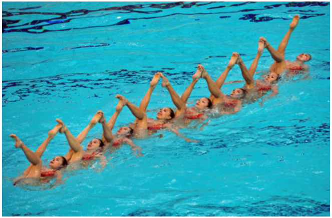
We have made things worse

- Estimating the mean of a Gaussian
- MCMC changed everything
- BUGS brought it to the masses
- I work on INLA, which does fast infrerence for latent Gaussian models
- Stan is even worse!

Busby Berkleey Dreams

We have made things worse

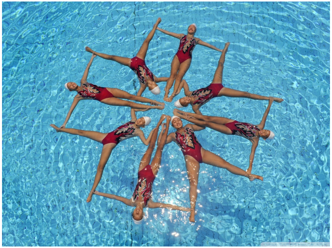
- Estimating the mean of a Gaussian
- MCMC changed everything
- BUGS brought it to the masses
- I work on INLA, which does fast infrerence for latent Gaussian models
- Stan is even worse!



Busby Berkleey Dreams

We have made things worse

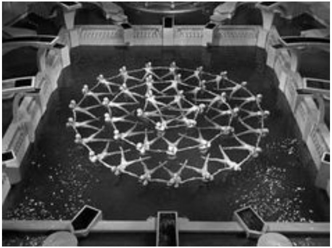
- Estimating the mean of a Gaussian
- MCMC changed everything
- BUGS brought it to the masses
- I work on INLA, which does fast infrerence for latent Gaussian models
- Stan is even worse!



Busby Berkleey Dreams

We have made things worse

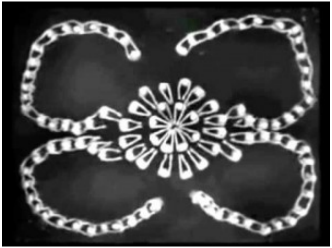
- Estimating the mean of a Gaussian
- MCMC changed everything
- BUGS brought it to the masses
- I work on INLA, which does fast infrerence for latent Gaussian models
- Stan is even worse!



Busby Berkleey Dreams

We have made things worse

- Estimating the mean of a Gaussian
- MCMC changed everything
- BUGS brought it to the masses
- I work on INLA, which does fast infrerence for latent Gaussian models
- Stan is even worse!



You cain't get a man with a gun

The real question is then **How do you set sensible priors for realistic models?**

- There is no universally applicable way to do this
- There are, however, lots of bad ways to do this
- Some of these bad ways may still work sometimes
- Our focus will be on hierarchical models (specifically Latent Gaussian Models)
- **Nothing** is going to infinity!

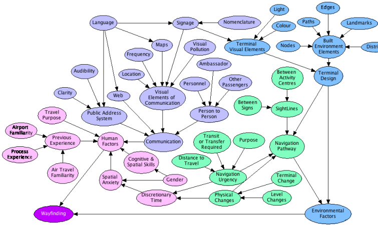
Today I will sketch our approach to this problem: Penalised Complexity (PC) priors.

Outline

Me. **I** am Mariah... the elusive chanteuse

One way to set priors is by **expert elicitation**

- Elicit probabilities for each quantity of interest
- Easiest when nodes are discrete
- A *Bayesian Network* is a useful tool for eliciting and combining information
- Turns elicited *conditional* probabilities in to a joint distribution



But what if there's more than one expert?

I've got 99 problems

Airports of the future project (ARC Linkage project LP0990135)

- From **99 "experts"** (airport users)
- probabilities for **49 binary nodes** were elicited
- All experts are equal
- Questions about how different elements of airport design affect the "wayfinding" experience

Treat the experts as measuring devices

We can consider this as a **measurement error** problem, in which each expert is providing a noisy measurement of the 49 nodes;

Expert elicitation as a measurement error model [Farr, S, Ruggeri, Mengersen, 2014]

Observed probability of node j for expert i:

$$\boldsymbol{p}_{ij} \sim \mathcal{B}(a_{ij}, b_{ij})$$

GLMM for the logit-mean

$$\text{logit}\left(\frac{a_{ij}}{a_{ij} + b_{ij}}\right) = \mu_j + w_i + \epsilon_j$$

Latent level:

- μ_j: Consensus (logit) probability for node j
- w_i: Systemic bias for expert i
- ϵ_j: The measurement bias for node j

Outline

The Ganzfeld Effect

Now it's time to attack the nuisance effects.

- u_i is the observer bias
- Standard random effect $w_i \sim N(0, \tau^{-1})$
- We need to put a prior on τ

Key point: We don't want this here!

I'm just a girl who cain't say no

So how do we set a prior on a precision?

- Lots of "expert guidance" from the literature
- Some of it is saying how to set priors on the precision
- Some of it is setting priors on the precision for a specific problem
- Conjugate priors, reference priors, weakly informative priors, ...
- When will it end?

We only want this effect to be in the model if it is required to fit the data.

We don't want a prior that the data has to drag towards "no effect"!

I'm just a girl who cain't say no

So how do we set a prior on a precision?

- Lots of "expert guidance" from the literature
- Some of it is saying how to set priors on the precision
- Some of it is setting priors on the precision for a specific problem
- Conjugate priors, reference priors, weakly informative priors, ...
- When will it end?

We only want this effect to be in the model if it is required to fit the data.

We don't want a prior that the data has to drag towards "no effect"!

I'm just a girl who cain't say no

So how do we set a prior on a precision?

- Lots of "expert guidance" from the literature
- Some of it is saying how to set priors on the precision
- Some of it is setting priors on the precision for a specific problem
- Conjugate priors, reference priors, weakly informative priors, ...
- When will it end?

We only want this effect to be in the model if it is required to fit the data.

We don't want a prior that the data has to drag towards "no effect"!

I'm just a girl who cain't say no

So how do we set a prior on a precision?

- Lots of "expert guidance" from the literature
- Some of it is saying how to set priors on the precision
- Some of it is setting priors on the precision for a specific problem
- Conjugate priors, reference priors, weakly informative priors, ...
- When will it end?

We only want this effect to be in the model if it is required to fit the data.

We don't want a prior that the data has to drag towards "no effect"!

I'm just a girl who cain't say no

So how do we set a prior on a precision?

- Lots of "expert guidance" from the literature
- Some of it is saying how to set priors on the precision
- Some of it is setting priors on the precision for a specific problem
- Conjugate priors, reference priors, weakly informative priors, ...
- When will it end?

We only want this effect to be in the model if it is required to fit the data.

We don't want a prior that the data has to drag towards "no effect"!

I'm just a girl who cain't say no

So how do we set a prior on a precision?

- Lots of "expert guidance" from the literature
- Some of it is saying how to set priors on the precision
- Some of it is setting priors on the precision for a specific problem
- Conjugate priors, reference priors, weakly informative priors, ...
- When will it end?

We only want this effect to be in the model if it is required to fit the data.

We don't want a prior that the data has to drag towards "no effect"!

I'm just a girl who cain't say no

So how do we set a prior on a precision?

- Lots of "expert guidance" from the literature
- Some of it is saying how to set priors on the precision
- Some of it is setting priors on the precision for a specific problem
- Conjugate priors, reference priors, weakly informative priors, ...
- When will it end?

We only want this effect to be in the model if it is required to fit the data.

We don't want a prior that the data has to drag towards "no effect"!

Basic instinct

A base model

- We have a model component with distribution $\pi(\boldsymbol{x} \mid \xi)$
- ξ is a **flexibility parameter**,
- $\xi = 0$ is indexes the **base model**
- The base model is the **simplest model**

Idea: Build a prior that has a mode at the base model. The posterior only concentrates on $\xi > 0$ if the data requires the more complex model.

Some examples

Case	Parameter	ξ	Base
Student-t	ν (dof)	$\xi = 1/\nu$	$\xi = 0$ (Gaussian)
IID	τ (precision)	$\xi = 1/\tau$	$\xi = 0$ (no random effect)
IGMRFs	τ (precision)	$\xi = 1/\tau$	$\xi = 0$ (const, linear, plane)
AR(1)	ρ (correlation)	$\xi = \rho$	$\xi = 0$ (no dep. in time)
		$\xi = \rho$	$\xi = 1$ (no changes in time)
FGN	H (Hurst param.)	$\xi = H$	$\xi = 0.5$ (White noise)
Correlation matrix	$\boldsymbol{R}$	$\boldsymbol{\xi} = \boldsymbol{R}$	$\boldsymbol{\xi} = \boldsymbol{I}$ (no correlation)

The pleasure principle

To build a prior that knows about the base model, I'm going to introduce the idea of **Penalised Complexity (PC) Priors**

- PC priors are our attempt to put together a set of principles that lead to a unique prior
- You can interrogate / criticise / modify the principles individually

Principle I: Occam's razor

Prefer simplicity over complexity

Consider the more complex model

$$\pi(x|\xi), \qquad \xi \geq 0$$

with base model $\pi(x|\xi = 0)$.

- The prior for $\xi \geq 0$ should penalise the complexity introduced by ξ
- The prior should be decaying with increasing measure by the complexity (the mode should be at the base model)

A prior will cause **overfitting/force complexity** if, loosely speaking,

$$\pi_\xi(\xi = 0) = 0$$

Principle I: Occam's razor

Prefer simplicity over complexity

Consider the more complex model

$$\pi(x|\xi), \qquad \xi \geq 0$$

with base model $\pi(x|\xi = 0)$.

- The prior for $\xi \geq 0$ should penalise the complexity introduced by ξ
- The prior should be decaying with increasing measure by the complexity (the mode should be at the base model)

A prior will cause **overfitting/force complexity** if, loosely speaking,

$$\pi_{\xi}(\xi = 0) = 0$$

Principle I: Occam's razor

Prefer simplicity over complexity

Consider the more complex model

$$\pi(x|\xi), \qquad \xi \geq 0$$

with base model $\pi(x|\xi = 0)$.

- The prior for $\xi \geq 0$ should penalise the complexity introduced by ξ
- The prior should be decaying with increasing measure by the complexity (the mode should be at the base model)

A prior will cause **overfitting/force complexity** if, loosely speaking,

$$\pi_\xi(\xi = 0) = 0$$

Principle I: Occam's razor

Prefer simplicity over complexity

Consider the more complex model

$$\pi(x|\xi), \qquad \xi \geq 0$$

with base model $\pi(x|\xi = 0)$.

- The prior for $\xi \geq 0$ should penalise the complexity introduced by ξ
- The prior should be decaying with increasing measure by the complexity (the mode should be at the base model)

A prior will cause **overfitting/force complexity** if, loosely speaking,

$$\pi_\xi(\xi = 0) = 0$$

Principle I: Occam's razor

Prefer simplicity over complexity

Consider the more complex model

$$\pi(x|\xi), \qquad \xi \geq 0$$

with base model $\pi(x|\xi = 0)$.

- The prior for $\xi \geq 0$ should penalise the complexity introduced by ξ
- The prior should be decaying with increasing measure by the complexity (the mode should be at the base model)

A prior will cause **overfitting/force complexity** if, loosely speaking,

$$\pi_\xi(\xi = 0) = 0$$

Principle II: Measure of complexity

Use Kullback-Leibler discrepancy to measure the increased complexity introduced by $\xi > 0$,

$$\mathrm{KLD}(f \| g) = \int f(x) \log\left(\frac{f(x)}{g(x)}\right) dx$$

for flexible model f and base model g.

Gives a measure of the information lost when the base model is used to approximate the more flexible models

Principle II: Measure of complexity

Use Kullback-Leibler discrepancy to measure the increased complexity introduced by $\xi > 0$,

$$\mathrm{KLD}(f \| g) = \int f(x) \log\left(\frac{f(x)}{g(x)}\right) dx$$

for flexible model f and base model g.

Gives a measure of the information lost when the base model is used to approximate the more flexible models

Principle III: Constant rate penalisation

Define

$$d(\xi) = \sqrt{2\ \mathrm{KLD}(\xi)}$$

as the (uni-directional) "distance" from flexible-model to the base model. Need the square-root to get the scale right.

Constant rate penalisation:

$$\pi(d) = \lambda \exp\left(-\lambda d\right), \qquad \lambda > 0$$

with mode at $d = 0$

Invariance: OK

Principle III: Constant rate penalisation

Define

$$d(\xi) = \sqrt{2\,\mathrm{KLD}(\xi)}$$

as the (uni-directional) "distance" from flexible-model to the base model. Need the square-root to get the scale right.

Constant rate penalisation:

$$\pi(d) = \lambda \exp\left(-\lambda d\right), \qquad \lambda > 0$$

with mode at $d = 0$

Invariance: OK

Principle III: Constant rate penalisation

Define

$$d(\xi) = \sqrt{2\,\mathrm{KLD}(\xi)}$$

as the (uni-directional) "distance" from flexible-model to the base model. Need the square-root to get the scale right.

Constant rate penalisation:

$$\pi(d) = \lambda \exp\left(-\lambda d\right), \qquad \lambda > 0$$

with mode at $d = 0$

Invariance: OK

Principle IV: User-defined scaling

The rate λ is determined from knowledge of the *scale* or some interpretable property or impact, $Q(\xi)$ of ξ:

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

- Problem dependent: must be!!!
- Can make the prior more informative or weakly informative this way

Principle IV: User-defined scaling

The rate λ is determined from knowledge of the *scale* or some interpretable property or impact, $Q(\xi)$ of ξ:

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

- Problem dependent: must be!!!
- Can make the prior more informative or weakly informative this way

The precision of a Gaussian

PC prior for the precision τ when $\tau = \infty$ defines the base model

- "random effects"/iid-model
- The smoothing parameter in spline models
- etc...

Result Let $\pi_\tau(\tau)$ be a prior for $\tau > 0$ where $E(\tau) < \infty$, then $\pi_d(0) = 0$ and the prior overfits.

The precision of a Gaussian

PC prior for the precision τ when $\tau = \infty$ defines the base model

- "random effects"/iid-model
- The smoothing parameter in spline models
- etc...

Result Let $\pi_\tau(\tau)$ be a prior for $\tau > 0$ where $E(\tau) < \infty$, then $\pi_d(0) = 0$ and the prior overfits.

The precision case (II)

The resulting prior is a type-2 Gumbel

$$\pi(\tau) = \frac{\lambda}{2}\tau^{-3/2}\exp\left(-\lambda/\sqrt{\tau}\right), \qquad \mathbb{E}(\tau) = \infty,$$

$\text{Prob}(\sigma > u) = \alpha$ gives

$$\lambda = -\frac{\ln(\alpha)}{u}$$

Alternative interpretation

$$\pi(\sigma) = \lambda \exp(-\lambda\sigma)$$

The precision case (II)

The resulting prior is a type-2 Gumbel

$$\pi(\tau) = \frac{\lambda}{2}\tau^{-3/2} \exp\left(-\lambda/\sqrt{\tau}\right), \qquad \mathbb{E}(\tau) = \infty,$$

$\text{Prob}(\sigma > u) = \alpha$ gives

$$\lambda = -\frac{\ln(\alpha)}{u}$$

Alternative interpretation

$$\pi(\sigma) = \lambda \exp(-\lambda\sigma)$$

Link with the tradition

Other (good) priors for the precision are

- A half-Gaussian on the standard deviation. (lighter tail than the PC prior)
- A half-Cauchy on the standard deviation. (heavier tail)
- A half-Student-t with more than 2 d.o.f. (heavier tail, similar risk properties)

The important thing here is that they all have a maximum at the base model. The tail behaviour is more "controversial"

Outline

Knowing me, knowing you

The final component of our model is nodal measurement error ϵ_j

- Big question: Is the measurement error independent across nodes?
- Maybe not?
- Nearby nodes measure "similar" things, so we would expect correlation
- We propose a BYM model

$$\epsilon_j = v_j + u_j$$

where $v_j \overset{\text{iid}}{\sim} N(0, \tau_v^{-1})$ and $\boldsymbol{u} \sim N(0, \tau_u^{-1}\boldsymbol{Q}^{+})$ is a Besag model.

Knowing me, knowing you

The final component of our model is nodal measurement error ϵ_j

- Big question: Is the measurement error independent across nodes?
- Maybe not?
- Nearby nodes measure "similar" things, so we would expect correlation
- We propose a BYM model

$$\epsilon_j = v_j + u_j$$

where $v_j \stackrel{\text{iid}}{\sim} N(0, \tau_v^{-1})$ and $\boldsymbol{u} \sim N(0, \tau_u^{-1}\boldsymbol{Q}^{+})$ is a Besag model.

Knowing me, knowing you

The final component of our model is nodal measurement error ϵ_j

- Big question: Is the measurement error independent across nodes?
- Maybe not?
- Nearby nodes measure "similar" things, so we would expect correlation
- We propose a BYM model

$$\epsilon_j = v_j + u_j$$

where $v_j \overset{\text{iid}}{\sim} N(0, \tau_v^{-1})$ and $\boldsymbol{u} \sim N(0, \tau_u^{-1}\boldsymbol{Q}^{+})$ is a Besag model.

Knowing me, knowing you

The final component of our model is nodal measurement error ϵ_j

- Big question: Is the measurement error independent across nodes?
- Maybe not?
- Nearby nodes measure "similar" things, so we would expect correlation
- We propose a BYM model

$$\epsilon_j = v_j + u_j$$

where $v_j \stackrel{\text{iid}}{\sim} N(0, \tau_v^{-1})$ and $\boldsymbol{u} \sim N(0, \tau_u^{-1}\boldsymbol{Q}^{+})$ is a Besag model.

The structured effect

The structured difference in u between neighbouring regions is $N(0, \tau_u^{-1})$.

$$\pi(\boldsymbol{u}) \propto \tau_u^{(n-1)/2} \exp\left(-\frac{\tau_u}{2} \sum_{i \sim j} (u_i - u_j)^2\right). \tag{1}$$

"$i \sim j$" denotes the set of all *unordered* pairs of neighbours.

- This is the Besag model.
- It is rank deficient.
- How do we put a prior on τ_u?
- Big thing: It will depend on the graph!

Building a better BYM

Base model $= 0 \rightarrow$ iid $\rightarrow$ dependence $=$ more flexible model

Rewrite the model as

$$\eta = \frac{1}{\sqrt{\tau}}\left(\sqrt{1-\gamma}v^* + \sqrt{\gamma}u^*\right)$$

where $\cdot^*$ is a unit-variance standardised model.

- Marginal precisions τ.
- γ gives it interpretation: independence ($\gamma = 0$), maximal dependence ($\gamma = 1$)]
- PC prior on γ (base model $\gamma = 0$) depends on the graph!
- Parameters control different features. Use the PC priors for τ and γ separately.

Building a better BYM

Base model $= 0 \to$ iid $\to$ dependence $=$ more flexible model

Rewrite the model as

$$\eta = \frac{1}{\sqrt{\tau}}\left(\sqrt{1-\gamma}v^* + \sqrt{\gamma}u^*\right)$$

where $\cdot^*$ is a unit-variance standardised model.

- Marginal precisions τ.
- γ gives it interpretation: independence ($\gamma = 0$), maximal dependence ($\gamma = 1$)]
- PC prior on γ (base model $\gamma = 0$) depends on the graph!
- Parameters control different features. Use the PC priors for τ and γ separately.

Building a better BYM

Base model $= 0 \rightarrow$ iid $\rightarrow$ dependence $=$ more flexible model

Rewrite the model as

$$\eta = \frac{1}{\sqrt{\tau}} \left(\sqrt{1-\gamma} v^* + \sqrt{\gamma} u^* \right)$$

where $\cdot^*$ is a unit-variance standardised model.

- Marginal precisions τ.
- γ gives it interpretation: independence ($\gamma = 0$), maximal dependence ($\gamma = 1$)]
- PC prior on γ (base model $\gamma = 0$) depends on the graph!
- Parameters control different features. Use the PC priors for τ and γ separately.

Building a better BYM

Base model $= 0 \to$ iid $\to$ dependence $=$ more flexible model

Rewrite the model as

$$\eta = \frac{1}{\sqrt{\tau}}\left(\sqrt{1-\gamma}v^* + \sqrt{\gamma}u^*\right)$$

where $\cdot^*$ is a unit-variance standardised model.

- Marginal precisions τ.
- γ gives it interpretation: independence ($\gamma = 0$), maximal dependence ($\gamma = 1$)]
- PC prior on γ (base model $\gamma = 0$) depends on the graph!
- Parameters control different features. Use the PC priors for τ and γ separately.

Building a better BYM

Base model $= 0 \to$ iid $\to$ dependence $=$ more flexible model

Rewrite the model as

$$\eta = \frac{1}{\sqrt{\tau}}\left(\sqrt{1-\gamma}v^* + \sqrt{\gamma}u^*\right)$$

where $\cdot^*$ is a unit-variance standardised model.

- Marginal precisions τ.
- γ gives it interpretation: independence ($\gamma = 0$), maximal dependence ($\gamma = 1$)]
- PC prior on γ (base model $\gamma = 0$) depends on the graph!
- Parameters control different features. Use the PC priors for τ and γ separately.

Get behind me, Esther Williams!

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

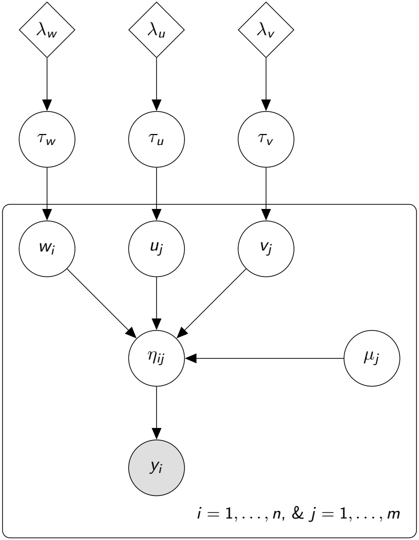
$$d(\gamma)^2 = n\gamma\left(\frac{1}{n}\operatorname{tr}(\boldsymbol{R}^{-1}) - 1\right) - \log\left|(1-\gamma)\boldsymbol{I} + \gamma\boldsymbol{R}^{-1}\right|$$

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

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

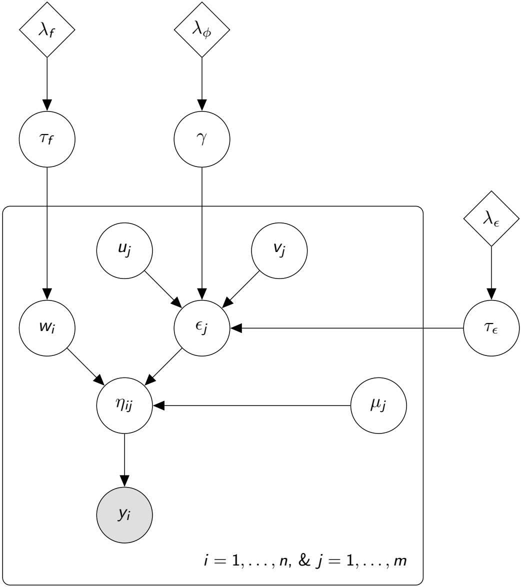
- (NB: $d(1)$ is finite, and so we use a truncated exponential!)

The BYM



Where is the variation coming from?

Building a better BYM

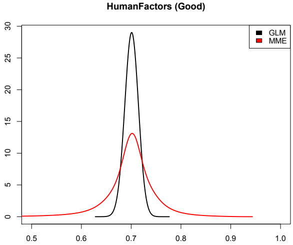


Outline

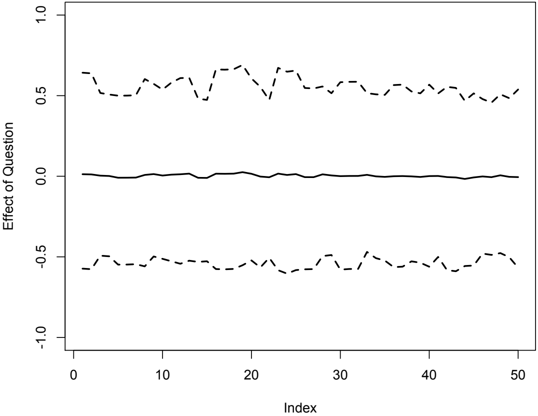
As long as you follow

So what was the outcome with the airports?

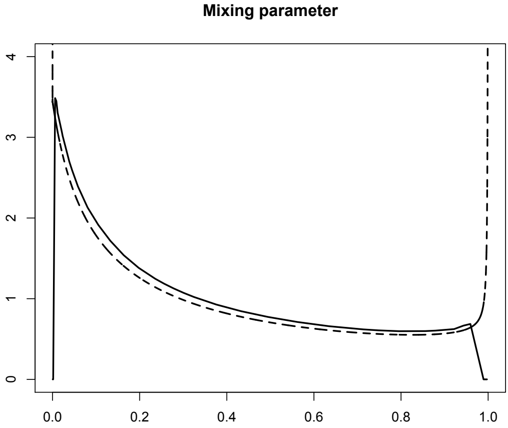
- The observer random effect was small with small credible regions
- The posterior estimates of the consensus probabilities have a larger IQR than those produced with a plain GLM
- But is this real?



Did anything happen?



Was the graphical structure useful?



Miss Otis Regrets

- In the end, the results for this particular problem were boring
- **This is good!**
- The aim of the PC prior project is to make priors that can find nothing when nothing is there
- The new BYM parameterisation gives a more interpretable way to look at the structure of the random effect
- The PC priors for this model satisfy a basic principle: **If something important in your model changes, the corresponding priors should also change**

Miss Otis Regrets

- In the end, the results for this particular problem were boring
- **This is good!**
- The aim of the PC prior project is to make priors that can find nothing when nothing is there
- The new BYM parameterisation gives a more interpretable way to look at the structure of the random effect
- The PC priors for this model satisfy a basic principle: **If something important in your model changes, the corresponding priors should also change**

Miss Otis Regrets

- In the end, the results for this particular problem were boring
- **This is good!**
- The aim of the PC prior project is to make priors that can find nothing when nothing is there
- The new BYM parameterisation gives a more interpretable way to look at the structure of the random effect
- The PC priors for this model satisfy a basic principle: **If something important in your model changes, the corresponding priors should also change**

Miss Otis Regrets

- In the end, the results for this particular problem were boring
- **This is good!**
- The aim of the PC prior project is to make priors that can find nothing when nothing is there
- The new BYM parameterisation gives a more interpretable way to look at the structure of the random effect
- The PC priors for this model satisfy a basic principle: **If something important in your model changes, the corresponding priors should also change**

Miss Otis Regrets

- In the end, the results for this particular problem were boring
- **This is good!**
- The aim of the PC prior project is to make priors that can find nothing when nothing is there
- The new BYM parameterisation gives a more interpretable way to look at the structure of the random effect
- The PC priors for this model satisfy a basic principle: **If something important in your model changes, the corresponding priors should also change**

Outline

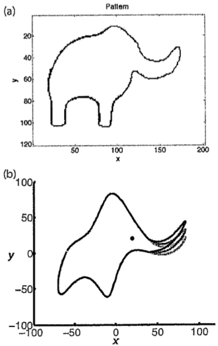
If love were all

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
- Parameters in Gaussian random fields (partially identifiable!)
- Non-stationary GRFs
- Correlated random effects
- Variances in multilevel models
- $+++$

Placating pugilistic pachyderms

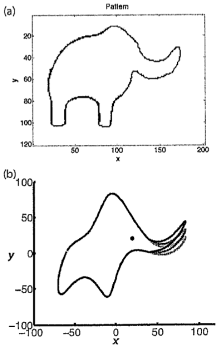
- Under everything, this was a talk about setting prior distributions
- This is hard.
- Bayesian models should not be used / interpreted unless you can interpret all levels of your model (including your prior)
- This doesn't fix the general problem of Bad Bayesian Analysis
- But it helps: we need to 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)

Placating pugilistic pachyderms

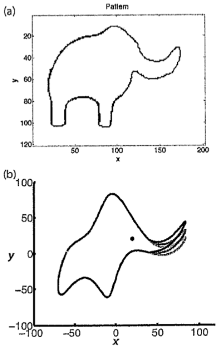
- Under everything, this was a talk about setting prior distributions
- This is hard.
- Bayesian models should not be used / interpreted unless you can interpret all levels of your model (including your prior)
- This doesn't fix the general problem of Bad Bayesian Analysis
- But it helps: we need to 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)

Placating pugilistic pachyderms

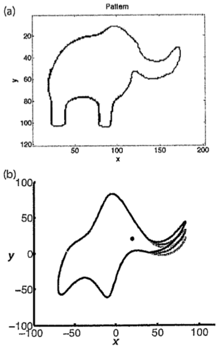
- Under everything, this was a talk about setting prior distributions
- This is hard.
- Bayesian models should not be used / interpreted unless you can interpret all levels of your model (including your prior)
- This doesn't fix the general problem of Bad Bayesian Analysis
- But it helps: we need to 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)

Placating pugilistic pachyderms

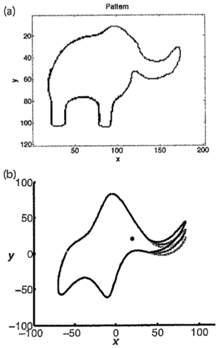
- Under everything, this was a talk about setting prior distributions
- This is hard.
- Bayesian models should not be used / interpreted unless you can interpret all levels of your model (including your prior)
- This doesn't fix the general problem of Bad Bayesian Analysis
- But it helps: we need to 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)

Placating pugilistic pachyderms

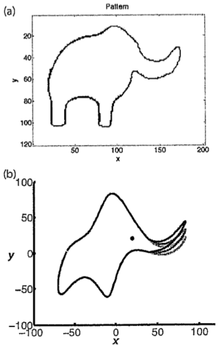
- Under everything, this was a talk about setting prior distributions
- This is hard.
- Bayesian models should not be used / interpreted unless you can interpret all levels of your model (including your prior)
- This doesn't fix the general problem of Bad Bayesian Analysis
- But it helps: we need to 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)

Placating pugilistic pachyderms

- Under everything, this was a talk about setting prior distributions
- This is hard.
- Bayesian models should not be used / interpreted unless you can interpret all levels of your model (including your prior)
- This doesn't fix the general problem of Bad Bayesian Analysis
- But it helps: 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

- D. P. Simpson, H. Rue, T. G. Martins, A. Riebler, and S. H. Sørbye (2014) *Penalising model component complexity: A principled, practical approach to constructing priors.* arxiv:1403.4630
- T. G. Martins and H. Rue. *Prior for flexibility parameters: the Student's t case.* Technical report S8-2013, Department of mathematical sciences, NTNU, Norway, 2013.
- S. H. Sørbye, and H. Rue (2014) *Scaling intrinsic Gaussian Markov random field priors in spatial modelling*, Spatial Statistics.
- A. C. Farr, D. P. Simpson, F. Ruggeri and K. Mengersen (2014). *Combining opinions for use in Bayesian Networks: a Measurement Error Approach*. QUT Eprints:79211.