A better prior for the precision parameter in Skyride/grid/track

Luiz Max F. de Carvalho
Institute of Evolutionary Biology, University of Edinburgh.

October 13, 2018

Abstract

I claim that the Gamma($\alpha = 0.001, \beta = 0.001$) prior we currently use leads to poor parameter estimates for small data sets and poor MCMC performance because it fails to regularise the posterior and lets the chain wander into unwelcome regions. I propose the penalised complexity (PC) prior from Simpson et al. (2017) as a solution that regularises inference in small data settings and does not lead to substantially different answers for bigger data regimes. I present a small simulation study and analyses of a few real-world data sets to support my claims.

Key-words: Skygrid; Gaussian process; PC prior; precision; performance.

The model

The first such model I will consider here is the Skyride model (Minin et al., 2008), which improves on previous semi-parametric models (Pybus et al., 2000) of piece-wise population size change by (i) assuming population size changes smoothly over time and (ii) places a smooth Gaussian process prior on the population sizes. Skyride operates on inter-coalescent intervals, i.e., intervals of time between coalescent events. For a phylogeny with n tips/leaves, let $\boldsymbol{s} = (\boldsymbol{s}_2, \ldots, \boldsymbol{s}_n)$ be the inter-coalescent intervals. If sampling is heterochronous, sampling times further divide inter-coalescent intervals in sub-intervals, i.e., $\boldsymbol{s}_k = (s_{k0}, \ldots, s_{kj_k})$. If we denote the population sizes by $\boldsymbol{\theta} = (\theta_2, \ldots, \theta_n)$, the likelihood becomes

$$Pr(\boldsymbol{s}|\boldsymbol{\theta}) = \prod_{k=2}^{n} Pr(s_k|\theta_k),$$

with

$$Pr(s_k|\theta_k) = \frac{n_{k0}(n_{k0}-1)}{2\theta_k} \exp\left(-\sum_{j=0}^{j_k} \frac{n_{kj}(n_{kj}-1)s_{kj}}{2\theta_k} \right).$$

If we make the convenient transformation $\gamma_k = \log(\theta_k), k = 2, \ldots, n$, we can then place the Gaussian Markov random field (GMRF) prior on $\boldsymbol{\gamma}$:

$$Pr(\boldsymbol{\gamma}|\tau) \propto \tau^{(n-2)/2} \exp\left(-\frac{\tau}{2} \sum_{k=2}^{n-1} \frac{(\gamma_{k+1} - \gamma_k)^2}{\delta_k} \right),$$

where δ_k is the (1d) distance between intervals and τ is the precision parameter associated with the smoothing. For details please see Minin et al. (2008).

The second model I will consider is the Skygrid model, an extension of the Skyride that allows for multiple loci. While in Skyride the estimated trajectory changes at coalescent times, in Skygrid changes occur at pre-specified fixed points in (calendar) time. This

allows population sizes to be estimated for multiple genealogies at once, e.g., when several genes are under analyses and have different genealogies. The researcher must select the number M of grid points to be used and a cut-off K in calendar time. The cut-off K is crucial to the Skygrid analysis, as it is the last point at which population sizes change and hence should be chosen commensurate with the age of the root. As with Skyride, the smoothness of the Skygrid prior is controlled by a precision parameter τ. The Skygrid model presents better statistical properties and is more general, which has led to it superseding Skyride in recent years. These models are parameter-rich and their use is preferable when the data are strongly informative about population history.

Hyperpriors

The precision parameter, τ, is not known, so we have to estimate it from the data. As good Bayesians that we are, we place a prior measure on it, $\pi_\tau(\tau)$. Currently we have a Gamma prior:

$$\pi_1(\tau \mid \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} \cdot \tau^{\alpha-1} \exp(-\tau/\beta)\,, \tau > 0,$$

such that $E_{\pi_1}[\tau] = \alpha\beta$ and $\mathrm{Var}_{\pi_1}(\tau) = \alpha\beta^2$. The choice we currently make is to set $\alpha = 0.001$ and $\beta = 1000$, so $E_{\pi_1}[\tau] = 1$ and $\mathrm{Var}_{\pi_1}(\tau) = 1000$, under the justification of creating a non-informative prior. I claim that this prior does not behave well. While I strongly dislike this "non-informative" argument in general, in what follows I shall argue my case on empirical rather than conceptual grounds.

Simpson et al. (2017) put forth a proposal for the construction of priors that penalises complexity by creating a prior that (a) includes the "base model", which of course varies in a model-by-model fashion and (b) penalises models that stray too far from said base model. An example is a hierarchical model where the base model is "no difference across groups" – complete shrinkage –, which translates to a zero standard deviation for the group effects. In page 14 they propose a PC prior on the precision τ from the Gumbel type II family – closely related to the Weibull. Let $a, b > 0$ be the shape and scale hyperparameters respectively; the probability density function is then

$$\pi_2(\tau \mid a, b) = ab \cdot \tau^{-a-1} \exp\left(-b\tau^{-a}\right),\ \tau > 0. \tag{1}$$

The authors recommend $a = 1/2$ and b to be chosen such that $Pr(1/\sqrt{\tau} > S) = p$, where the value S and the probability p are to be chosen on substantive grounds. This leads to $b = \ln(p)/S$. Here I will choose $S = 1$ and $p = 0.1$, which is to say that I will construct the prior such that there is a 10% probability that the standard deviation of the log-population sizes γ_i will be greater than 1. This gives $b = 2.302585$. For this prior, $E_{\pi_2}[\tau] = b^{\frac{1}{a}}\Gamma(1 - 1/a)$ and $\mathrm{Var}_{\pi_2}(\tau) = b^{\frac{2}{a}}\left[\Gamma(1 - 1/a) - \Gamma(1 - 1/a)^2\right]$, which means that for $0 < a \leq 1$ the first moment does not exist and for $0 < a \leq 2$ the variance is infinite. As weird as this prior may sound, Simpson et al. (2017) show that while **any** Gamma prior will overfit – in a particular sense discussed in detail in the paper –, the Gumbel prior will not. As we shall see, this prior does seem to have better empirical behaviour, specially for a small data set. Figure 1 shows a plot of the current Gamma prior and the proposed Gumbel prior.

In it is of interest to anyone, the induced distributions on the standard deviation σ,

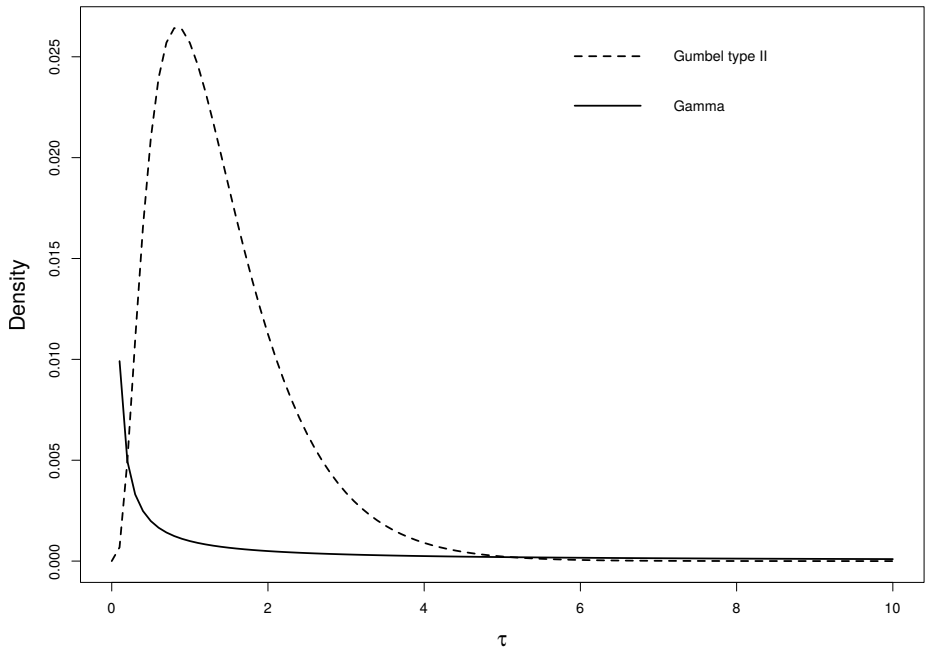

Figure 1: **Priors on the precision parameter,** τ. I show the current default Gamma prior ($\alpha = 0.001$, $\beta = 1000$, solid line) and the proposed Gumbel type II prior with $a = 1/2$ and $b = 2.302585$ (dashed line).

$\pi_\tau(1/\sigma^2)|J|$, are

$$\pi_1(\sigma|\alpha,\beta) = \frac{2}{\Gamma(\alpha)\beta^\alpha} \cdot \sigma^{-(2\alpha+1)} \exp\left(-\frac{1}{\beta\sigma^2}\right), \ \sigma > 0,$$

for the Gamma prior and

$$\pi_2(\sigma|a,b) = 2ab \cdot \sigma^{2a-1} \exp\left(-b\sigma^{2a}\right), \ \sigma > 0,$$

for the Gumbel prior, with absolute Jacobian $|J| = 2/\sigma^3$ (see Figure 2).

Experiments

Simulated data

I simulated five hundred 20- and 200-taxa time-calibrated phylogenies under a constant population coalescent model with $N_e = 10$. I then used the function `BNPR()` in the **phylodyn**[1] package (Lan et al., 2015) to get the population size trajectories, because I cannot be asked to go through all the scripting needed to do a simulation study in BEAST. If anybody takes issue – and they will be justified in doing so – they can do the simulation themselves and I bet £50 that the results will be scientifically the same.

[1]The PC prior stuff is implemented in my fork: https://github.com/maxbiostat/phylodyn/tree/pc_prior

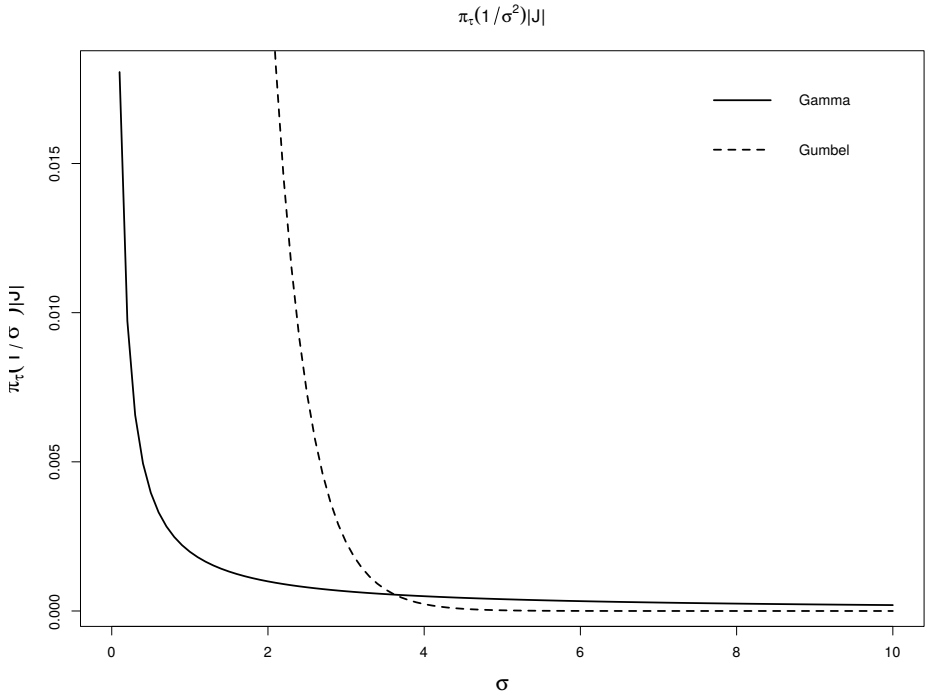

Figure 2: **Induced priors on the standard deviation,** σ. I show the transformation of the current default Gamma prior ($\alpha = 0.001$, $\beta = 1000$, solid line) and the proposed Gumbel type II prior with $a = 1/2$ and $b = 2.302585$ (dashed line).

I expect the BEAST reconstructions to a bit noisier if said saint person would simulate sequences down the trees and estimate the trees (rather than keeping the "true" tree fixed).

I follow Gill et al. (2012) and Hall et al. (2016) and compute:

- The percent bias:

$$b_{\%} = \frac{100}{R}\int_0^R \frac{\hat{N(t)} - N(t)}{N(t)}dt, \text{and}$$

- the size of the posterior BCI:

$$s = \frac{1}{R}\int_0^R \frac{\hat{N}_{0.975}(t) - \hat{N}_{0.025}(t)}{N(t)}dt$$

where R is the time of the most recent sample.

The results are presented in Figures 3, 4 and 5, and show that the proposed prior leads to slightly more accurate reconstructions with narrower 95% Bayesian credibility intervals (BCIs). In the next section I will show an example where the Gumbel prior actually leads to broader a HPD for a small data set [2] will offer an explanation of why I think the proposed prior is doing the right thing. It is nevertheless encouraging to see that *on average* it will in fact lead to less biased inferences with more certainty.

[2] "Small" meaning of the same order of magnitude as the 20 taxa ones here.

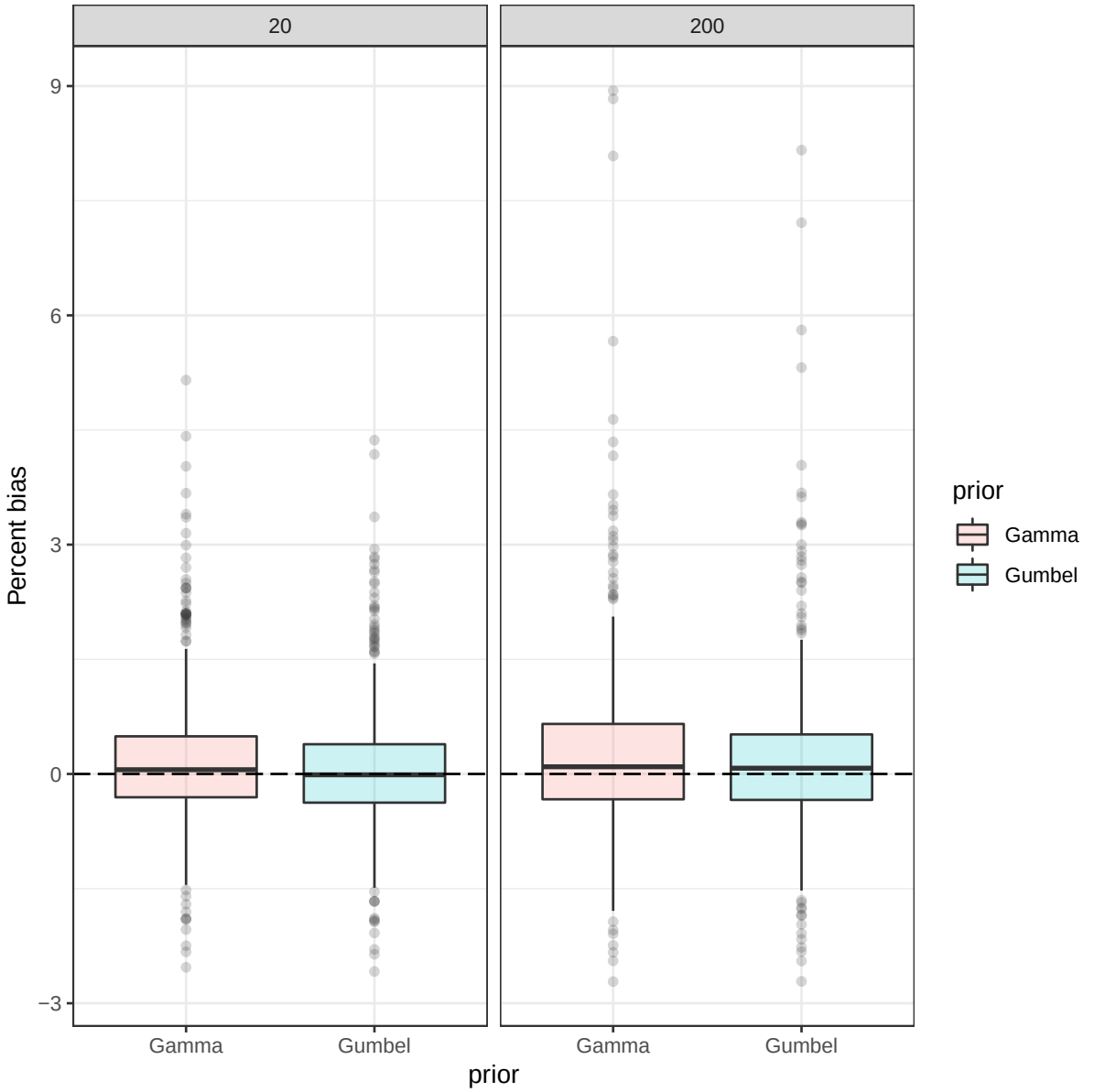

Figure 3: **Percent bias ($b_{\%}$) for 500 replicates of each number of taxa (20, 200)**. Vertical tiles show different numbers of taxa. Dashed line marks zero for convenience.

Real-world data sets

I used BEAST to estimate the Skygrid parameters under both priors for a few data sets. BEAST was run three independent times for a a sufficient amount of interations to be sure that the standard error (Monte Carlo error) was below 5% of the standard deviation for the parameters of interest.[3] Table 1 summarises my findings. Before I proceed I would like to sound a warning: estimates of quantiles via Monte Carlo (let alone MCMC) are VERY noisy. Take everything you see here about quantiles with a pinch of salt.

I also present the reconstructions for the Dengue 4 data set, in Figure 6A. Now notice how the blue line (Gumbel) wiggles more and also has a wider HPD interval. This is

[3]Also, all ESS were larger than 200. I actually don't quite agree with this "ESS bigger than 200" rule of thumb, but for my purposes here it will do.

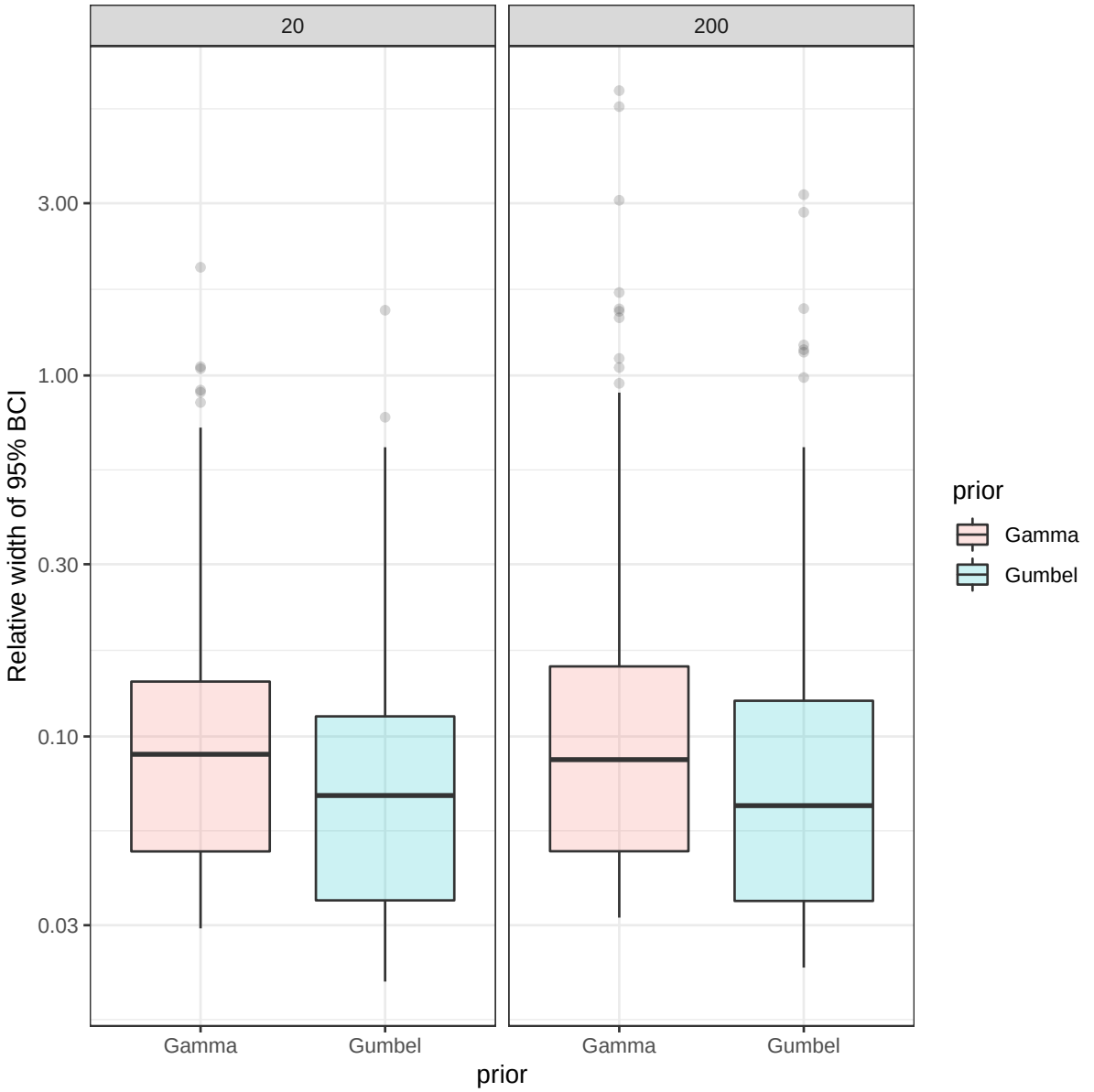

Figure 4: **Size of the posterior BCI (s) for 500 replicates of each number of taxa (20, 200)**. Vertical tiles show different numbers of taxa.

because of bigger uncertainty about $\boldsymbol{\gamma}$, which makes estimates of the median noisier. This could be regarded as a poorer reconstruction, since the data set clearly supports a constant population size through time. But I argue that this reasoning is wrong: if we have little data, it makes sense to be less certain about the reconstructions, which is exactly what the proposed prior induces. The current prior places way too much mass on small precisions (which would make things noisier) while at the same time allowing rather extreme values, which would artificially inflate confidence (see Figures 1 and 2). For a larger data set with a lot more information about the population dynamics, the bands are quite similar (Figure 6B).

Finally, to drive home my point about prior sensitivity, I have re-created the prior sensitivity analysis (PSA) in Table 4 of Gill et al. (2012) but using the Dengue 4 data set

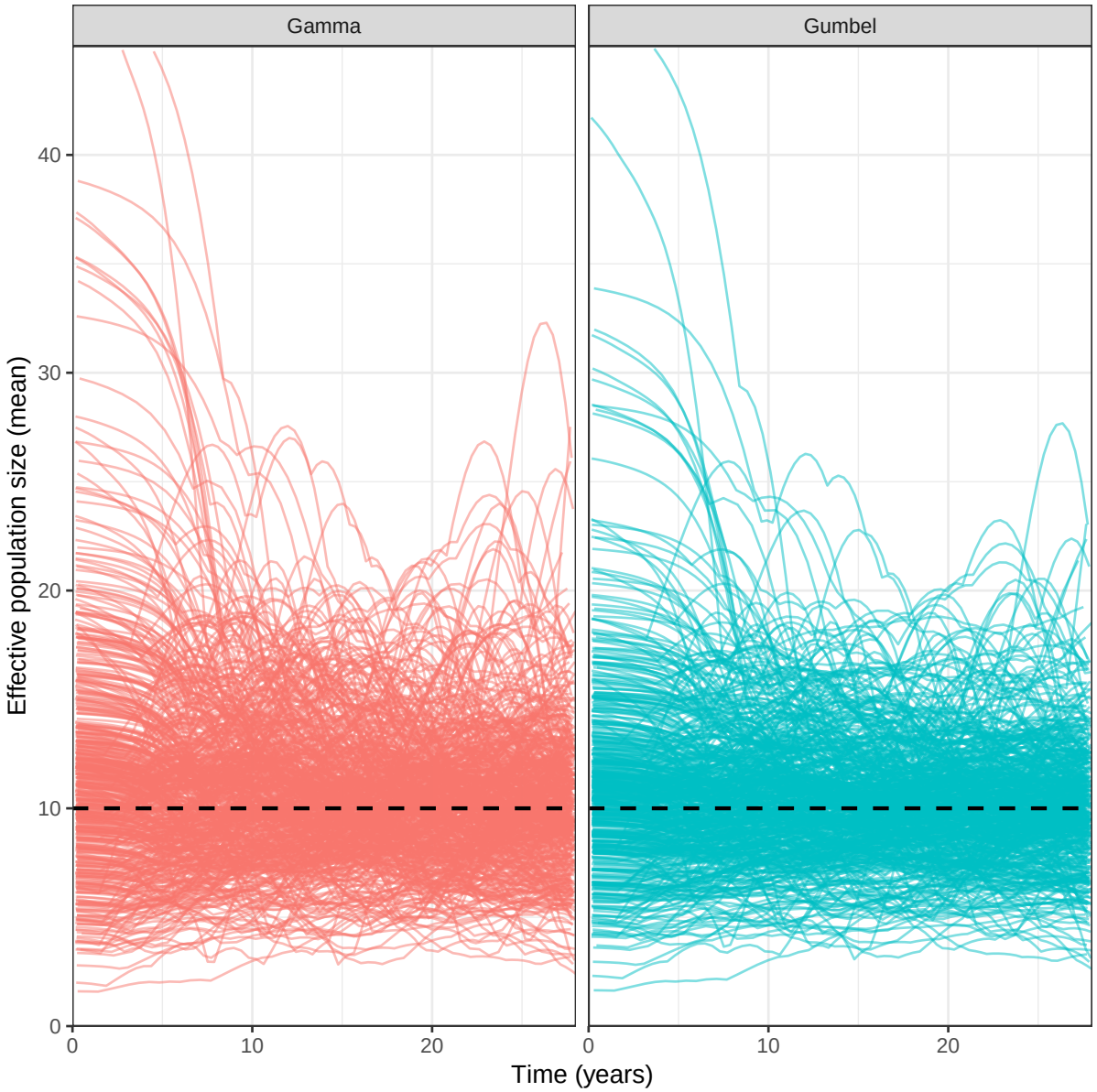

Figure 5: **Posterior mean of effective population size trajectories for** 500 **replicates of each number of taxa (20, 200)**. Vertical tiles show different numbers of taxa. Dashed line marks the ground truth, $N_e = 10$.

(Table 2). I also did a similar analysis using the Gumbel prior, in which the varying quantity is the "threshold" standard deviation, S. Table 3 holds the results of this experiment. As you can see, when the data contain little information (17 taxa against the 152 taxa from the original PSA) the prior does matter, at least insofar as the estimates of the precision are concerned – population sizes are fine, as you would expect since τ controls the smoothness of the trajectory but not its magnitude. Moreover, the Gumbel prior seems to be a bit more robust to "prior misspecification", owing perhaps to its tail behaviour. There is not much difference, though. To see this, compute $e_{\text{Gamma}} = (232 - 138)/138 = 0.68$ and $e_{\text{Gumbel}} = (1.58 - 0.97)/0.97 = 0.63$.

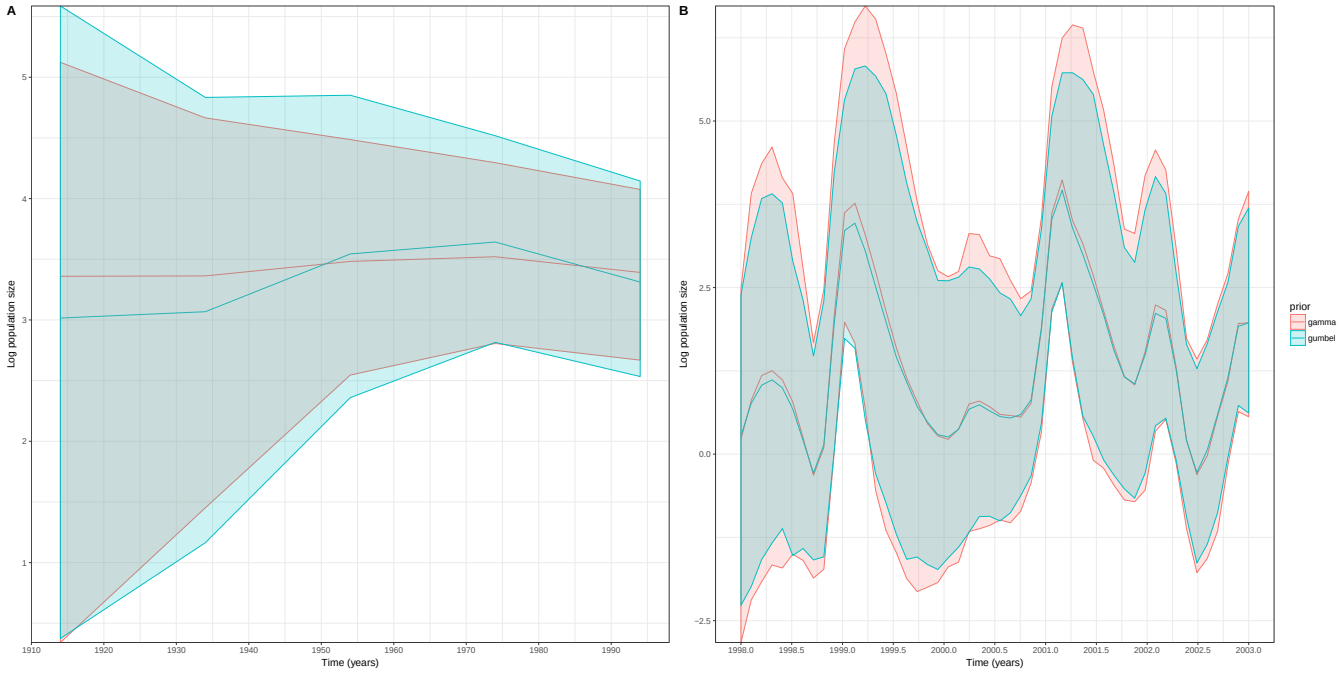

Figure 6: **Skygrid reconstructions for two data sets under the current and proposed prior**. I show the reconstructions for the Dengue serotype 4 *env* (panel A) and Influenza H3N2 HA (panel B) data sets.

Conclusion

While the differences between the two priors compared here are not dramatic, the PC prior does seem to have better empirical properties. It is also justified at a conceptual level. To hell with "non-informative" priors. I have added the new prior to both BEAST[4] and phylodyn, should anyone want to try it out. The implementation in phylodyn is even in terms of the "substantive parameters", S and p.

More results can be found at https://github.com/maxbiostat/CODE/tree/master/skygrid_hyperprior.

Acknowledgements

I thank Mandev Gill (Leuven) for helpful discussions.

[4] https://github.com/maxbiostat/beast-mcmc/tree/gumbel_typeII.

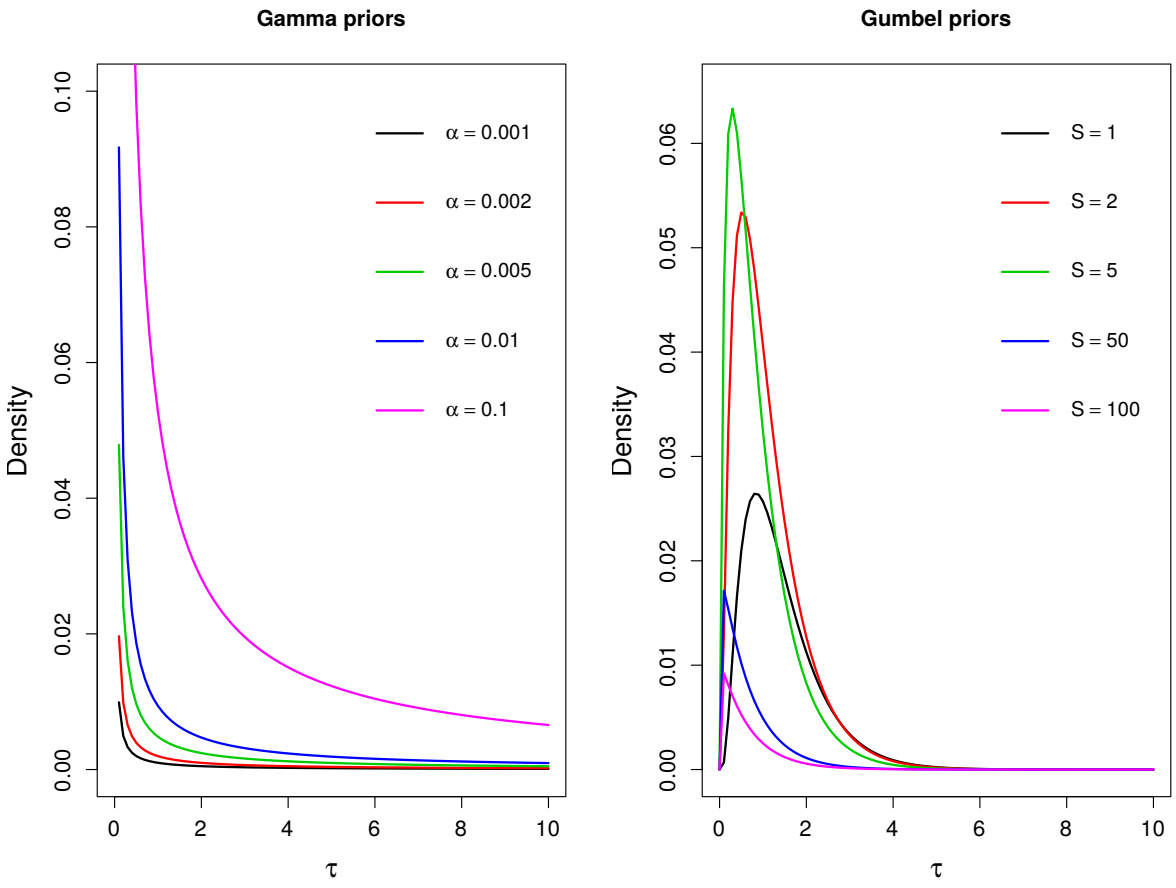

Figure 7: **Densities used in the prior sensitivity analysis for the precision parameter,** τ. I show the same Gamma priors as Gill et al. (2012) ($\alpha = 0.001, 0.002, 0.005, 0.01$ and 0.1 with $\beta = 1000$. For the Gumbel densities I used $a = 1/2$ and $S = 1, 2, 5, 50$ and 100 which lead to $b = 2.30, 1.15, 0.46, 0.046$ and 0.023. Please note that y-axes are different in scale.

References

Gill, M. S., Lemey, P., Faria, N. R., Rambaut, A., Shapiro, B., and Suchard, M. A. (2012). Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Molecular biology and evolution*, 30(3):713–724.

Hall, M. D., Woolhouse, M. E., and Rambaut, A. (2016). The effects of sampling strategy on the quality of reconstruction of viral population dynamics using bayesian skyline family coalescent methods: A simulation study. *Virus evolution*, 2(1).

Lan, S., Palacios, J. A., Karcher, M., Minin, V. N., and Shahbaba, B. (2015). An efficient bayesian inference framework for coalescent-based nonparametric phylodynamics. *Bioinformatics*, 31(20):3282–3289.

Minin, V. N., Bloomquist, E. W., and Suchard, M. A. (2008). Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Molecular biology and evolution*, 25(7):1459–1471.

Pybus, O. G., Rambaut, A., and Harvey, P. H. (2000). An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics*, 155(3):1429–1437.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., Sørbye, S. H., et al. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, 32(1):1–28.

Table 1: **Comparison of current and proposed prior on τ for a range of real world data sets**. All MCSE were below 5% of the estimated posterior mean.

						Posterior mean τ (HPD)		ESS τ		Average ESS γ	
Data set	number of taxa	sites	sampling span	M	K (years)	Gamma	Gumbel	Gamma	Gumbel	Gamma	Gumbel
Dengue 4	17	1485	1956-1994	5	80	156 (0, 791)	1.58 (0.28, 3.39)	834	3748	2284	4737
YFV	71	654	1953-2009	30	3000	130 (0, 697)	1.57 (0.25, 3.36)	1358	2865	284	256
Influenza H3N2	165	1698	2000-2003	50	5	0.65 (0.13, 1.36)	0.83 (0.27, 1.55)	1504	2548	2048	2589
FMDV A	184	649	1955-2013	50	100	1.45 (0.18, 3.54)	1.31 (0.36, 2.49)	865	1430	1348	1577
FMDV O	225	649	1958-2011	50	100	2.11 (0.23, 4.83)	1.57 (0.39, 2.92)	1280	2077	1517	1930

Table 2: **Prior sensitivity analysis for the Dengue4 data set, Gamma priors**. I used the same priors as in Table 4 of Gill et al. (2012).

α	posterior mean τ	posterior median τ	posterior hpd	mcse	ESS τ
0.001	138	20	0, 716	11	844
0.002	157	22	0, 806	13	782
0.005	154	24	0, 792	10	935
0.01	173	29	0, 902	13	757
0.1	232	53	0, 1088	15	781

Table 3: **Prior sensitivity analysis for the Dengue4 data set, Gumbel type II priors**. I devised priors with increasing values of S, see text.

S (b)	posterior mean τ	posterior median τ	posterior hpd	mcse	ESS τ
1 (2.302)	1.58	1.4	0.24, 3.34	0.01	3525
2 (1.151)	1.3	1.12	0.15, 2.92	0.01	3255
5 (0.460)	1.11	0.92	0.04, 2.64	0.01	2677
50 (0.046)	1.01	0.81	0.01, 2.58	0.01	2617
100 (0.023)	0.97	0.78	0.01, 2.44	0.02	2123