Received: 28 October 2019 | Accepted: 6 January 2020

DOI: 10.1111/biom.13275

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

DISCUSSION

Discussion on "Horseshoe-based Bayesian nonparametric estimation of effective population size trajectories" by James R. Faulkner, Andrew F. Magee, Beth Shapiro, and Vladimir N. Minin

Lorenzo Cappello | **Swarnadip Ghosh** | **Julia A. Palacios**

Department of Statistics, Stanford University, Stanford, California

Correspondence
Julia A. Palacios, Department of Statistics, Stanford University, Stanford, CA 94305.
Email: juliapr@stanford.edu

Funding information
Alfred P. Sloan Foundation; NIH, Grant/Award Number: R01 GM131404

1 | INTRODUCTION

The authors present an attractive solution to a long-standing problem of local adaptivity of Gaussian process priors for phylodynamic inference. While Gaussian process–based phylodynamics have been used for over 10 years (Minin *et al.*, 2008; Gill *et al.*, 2013; Palacios and Minin, 2013), these methods a priori assume a single precision parameter that controls smoothness over the whole population size history, limiting precision of posterior estimates in cases of variable smoothness over time or abrupt changes. The authors propose a horseshoe Markov random field (HSMRF) prior of order p on the log-effective population size trajectory at a regular fixed grid of $H + 1$ time points. The HSMRF is flexible to local adaptivity modeling each pth-order forward difference of the log-trajectory with a prior that spikes at 0 with Cauchy-like heavy tails. The HSMRF favors small variance of small population size jumps and large variance of large population size jumps. We discuss two aspects of the proposed method: (a) posterior checks and model selection that can accompany HSMRF modeling tools, and (b) the ability of the HSMRF model to differentiate between alternative population size trajectories that can be translated into meaningful scientific discoveries. In this discussion, we assume the inference setting in which a genealogy is observed.

2 | MODEL CHECKS AND MODEL SELECTION

As clearly stated by the authors in their introduction, the toolkit available to infer population sizes from molecular DNA data has dramatically expanded. Different models may be more suitable than others in extracting evolutionary patterns from observed data. Here, we argue that it is essential to develop tailored techniques for model checking to pair with model selection. The issue is self-evident by looking at the different outputs of the four model specifications studied in the paper: Gaussian Markov random field (GMRF) and HSMRF, both of orders 1 and 2. GMRF priors enforce constant precision across the trajectory, whereas HSMRF priors allow for varying levels of precision. Similarly, first-order models target piecewise constant trajectories, whereas second-order models target piecewise linear functions. For example, in the study of Egyptian Hepatitis C Virus, HSMRF of order 1 showed a very rapid increase in $N_e(t)$ around 1920, that is, a spike that none of the other models recovered.

The authors use the Watanabe—Aikake information criteria and Bayes factors for model comparison. While both statistics have clear statistical interpretation, they lack the ability to detect model fit in terms of biologically meaningful variables. Posterior-predictive checks are widely used to assess model fit (Gelman *et al.*, 2013); it involves sampling from the posterior-predictive distribution

$$P(\mathbf{T}'|\mathbf{t}) = \int P(\mathbf{T}'|\theta)\pi(\theta|\mathbf{t})d\theta,$$

where θ is the random vector of size H that parametrizes the effective population size as a piecewise constant or linear trajectory, $\mathbf{t}$ is the observed vector of coalescent times, $\pi(\theta|\mathbf{t})$ is the posterior distribution, and $\mathbf{T}'$ is a random vector of coalescent times with posterior-predictive distribution $P(\mathbf{T}'|\mathbf{t})$. The posterior-predictive distribution is the distribution of

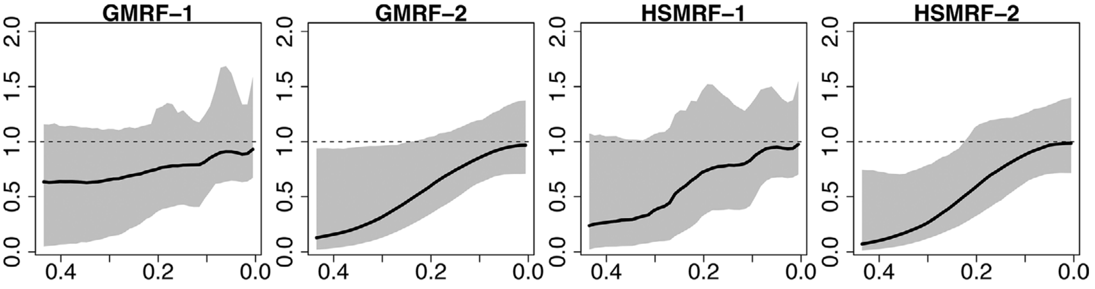

FIGURE 1 Example 1: effective population size estimates from an atypical constant trajectory. Estimated curves with GMRF-1, GMRF-2, HSMRF-1, and HSMRF-2 (left to right). Black lines depict the posterior median, gray-shaded areas correspond to 95% credible regions, and dashed lines depict the true $N_e(t)$.

coalescent times after accounting for effective population size uncertainty. The discrepancy between the model and the data is measured via an appropriately chosen test quantity $T(\mathbf{T}', \theta)$, which is a statistic of parameters and data. Then, one proceeds either through visual comparison of the empirical distribution of $T(\mathbf{T}', \theta)$ and the observed test statistic $T(\mathbf{t}', \theta)$, or by computing posterior-predictive p-values

$$p_B = \int\int I_{T(\mathbf{T}',\theta)\geq T(\mathbf{t},\theta)} P(\mathbf{T}'|\theta)\pi(\theta|\mathbf{t})d\mathbf{T}'d\theta.$$

Although the choice of a test statistic is not a trivial endeavor (Gelman *et al.*, 2013; Diaconis and Wang, 2018), we chose two closely related statistics, *total tree length* and *time to the most common ancestor* (TMRCA), to compare the four models: GMRF and HSMRF, both of orders 1 and 2.

Example 1: Atypical genealogy. We simulated a genealogy with $n = 50$ samples at $t = 0$ with $N_e(t) = 1$ for all $t \geq 0$. We labeled this example "atypical genealogy" because the sampled genealogy is not generated at random from the coalescent distribution: we picked the genealogy in the lowest 0.001 quantile out of 10,000 realizations. Our atypical genealogy's TMRCA is around 0.45 (expected value under this model is 1.96). Estimated population sizes with the four models are shown in Figure 1 using the authors' R code implementation with model parameters as suggested by the authors. All models are biased downward: this was expected as we have chosen an atypical genealogy. Only GMRF-1 and HMRF-1 credible intervals fully cover the true trajectory and the former outperforms the latter. Similarly, GMRF-2 outperforms HSMRF-2, with the latter covering solely 51% of the true trajectory. For our posterior-predictive check, we generated 2000 genealogies from the posterior predictive of each model and computed the corresponding tree length and TMRCA.

Our posterior checks are displayed in Figure 2A as boxplots of the deviations from the true values (sampled tree length − true tree length, sampled TMRCA − true TMRCA) of sampled values obtained from the predictive posterior distribution. The four distributions of deviations from the truth are skewed: having picked an outlier, this is expected. The p-values of tree length confirm this: $p_B^{\text{HSMRF-2}} = 0.911$, $p_B^{\text{GMRF-2}} = 0.915$, $p_B^{\text{HSMRF-1}} = 0.921$, and $p_B^{\text{GMRF-1}} = 0.929$. As expected, we see little discrepancy between the two statistics and across the four models.

This example illustrates a discrepancy that may happen in real applications: GMRF-1 is the model that more closely recovers the true trajectory and it is closest to the true model while HSMRF-2 exhibits the worst performance. However, in terms of predictive posterior checks, HSMRF-2 is the model whose predictive posterior samples are closer to the observed TRMCA and tree length (lowest median and sample variance in Figure 2A).

Example 2: Average genealogy. We replicated the simulation conditions of Example 1, except that this time we picked a genealogy whose TMRCA was in the 47th quantile (tree length in the 62nd quantile). Figure 3 shows the estimated trajectories for the four models and Figure 2B plots the posterior-predictive distribution of deviations from the true values. For a typical genealogy, the models perform exactly how it is expected theoretically, with GMRF-1 outperforming all others models, and HSMRF-2 being the worst of the four. Posterior checks completely agree this view and show how GMRF-1 is now the model that provides the best fit to the data, that is, posterior samples are more concentrated around 0.

Two final remarks. The conclusions obtained from predictive checks can be, to certain extent, anticipated from the observed credible regions, however a posterior-predictive distribution of a meaningful test statistic may offer better interpretability. Also, these considerations can be largely extended to the case of unknown genealogies. Here we assumed an observed timed genealogy and, consequently, we discussed sampling vectors of coalescent times from the posterior predictive to compare with the observed data. In a similar manner, one could sample DNA sequences and proposed new test statistic such as observed number of segregating sites.

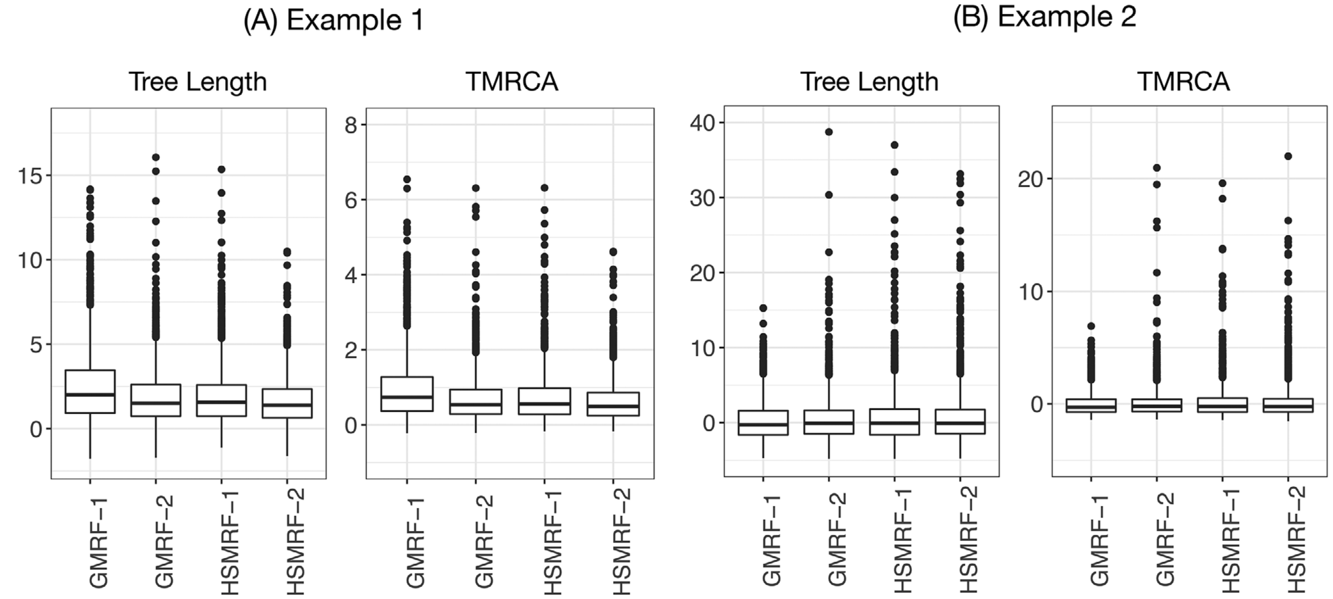

FIGURE 2 Boxplots of the posterior-predictive deviation of the tree length and TMRCA from the true values. We sampled 2000 sets of coalescent times from each posterior-predictive distribution $P_{\text{GMRF-1}}(\mathbf{T}'|\mathbf{t})$, $P_{\text{GMRF-2}}(\mathbf{T}'|\mathbf{t})$, $P_{\text{HSMRF-1}}(\mathbf{T}'|\mathbf{t})$, and $P_{\text{HSMRF-2}}(\mathbf{T}'|\mathbf{t})$. Each plot shows the deviations from the true tree lengths and TMRCAs. Panel A refers to Example 1, Panel B to Example 2.

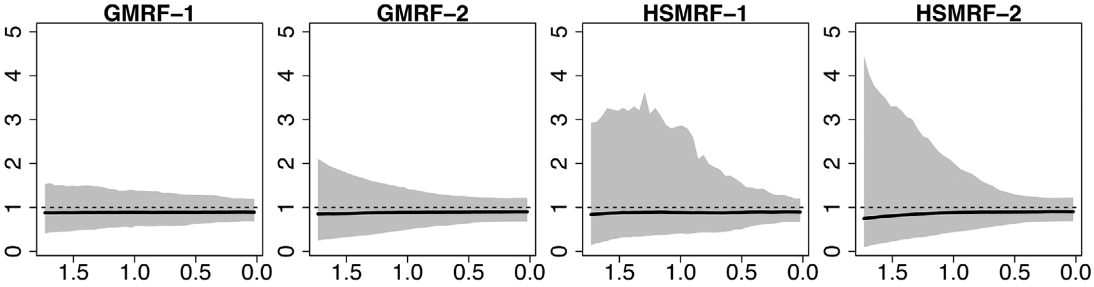

FIGURE 3 Example 2: effective population size estimates from a typical constant trajectory. Estimated curves with GMRF-1, GMRF-2, HSMRF-1, and HSMRF-2 (left to right). Black lines depict the posterior median, gray-shaded areas correspond to 95% credible regions, and dashed lines depict the true $N_e(t)$.

3 | DISTINGUISHING BETWEEN ALTERNATIVE POPULATION SIZE TRAJECTORIES

The authors claim that HSMRF priors are more flexible to respond to rapid changes in the effective population size. To provide an alternative empirical assessment of such claim, we simulated genealogical realizations from a population that experiences a bottleneck. More precisely, we assumed the following population size trajectory:

$$N_0(t) = \begin{cases} 0.1 & 2 \le t \le 4 \\ 1 & \text{elsewhere.} \end{cases} \tag{1}$$

We evaluated the ability of the four methods: HSMRF of orders 1 and 2, and GMRF of orders 1 and 2 to distinguish between the true trajectory and a shifted by s trajectory, that is, we consider alternative population size trajectories:

$$N_s(t) = \begin{cases} 0.1 & 2+s \le t \le 4, \text{ for } s \in \{0.1, 0.2, 0.3, 0.4, 0.5\} \\ 1 & \text{elsewhere.} \end{cases}$$

To evaluate the four methods we first computed the Envelope statistic as defined in Faulkner *et al.* (2019), and the Envelope restricted to the interval (1.5,4.5), that is,

$$\text{Envelope}_{(1.5,4.5)} = \frac{100}{B} \sum_{j=1}^{B} \sum_{x_i \in (1.5,4.5)} \times \frac{I\left(\hat{N}^{(j)}_{.025}(x_i) < N_0(x_i) < \hat{N}^{(j)}_{.975}(x_i)\right)}{\#\{x_i \in (1.5, 4.5)\}},$$

TABLE 1 Envelope and confusion statistics. Bold depicts the method with the best performance

Statistic	GMRF-1	HSMRF-1	GMRF-2	HSMRF-2
Envelope	91.4	**95.8**	80.5	84.9
Envelope (1.5,4.5)	85.8	**93.8**	69.4	77.3
$C_{0,0.1}$	51	**8**	99	71
$C_{0,0.2}$	7	**4**	79	29
$C_{0,0.3}$	3	**2**	37	13
$C_{0,0.4}$	**0**	1	16	10
$C_{0,0.5}$	**0**	1	6	6

Comparison of the ability of the models to distinguish between alternative population bottlenecks. Average across $B = 100$ simulations from a coalescent model of $n = 200$ samples with population size trajectory N_0 (Equation 1).

where $I()$ is the indicator function, $\hat{N}^{(j)}_{.025}(x_i)$ and $\hat{N}^{(j)}_{.0975}(x_i)$ are the 2.5 and 97.5 percentiles of the posterior distribution at time x_i from the jth simulated genealogy according to $N_0(t)$. B is the number of simulated genealogies, and $x_1, \ldots, x_m$ is the corresponding regular grid of m points for each simulation. We evaluate our statistics only for the interval (1.5,4.5) that covers the bottleneck.

We computed the following confusion statistic:

$$C_{0,s} = \frac{100}{B}\sum_{j=1}^{B} I\Big(\text{Envelope}_{(1.5,4.5)}\big(N_0^{(j)}\big) = \text{Envelope}_{(1.5,4.5)}\big(N_s^{(j)}\big)\Big),$$

which is the percentage of simulations in which the two curves N_0 and N_s are equally covered by the 95% credible intervals for N_0.

For our results in Table 1 we simulated $B = 100$ genealogies of 200 leaves with 50 samples at time 0 and 150 uniformly sampled in the interval (0,2). The first row in Table 1 shows a mean envelope that mirrors the same patterns of Figure 2 in (Faulkner *et al.*, 2020): the best envelope ≥95% is achieved with HSMRF-1. The envelope restricted to the interval (1.5,4.5) achieved with HSMRF-1 is reduced to 93.8%. The 95% credible regions in the interval (1.5,4.5) estimated with HSMRF-1 have the same envelope for both the truth N_0 and $N_{0.1}$, in only 8% of the simulations. This is a striking advantage over its counterpart GMRF-1 (51%) and all other models (>70%). The gains are reduced as s increases. For $s = 0.4$ and $s = 0.5$, HSMRF-1, and GMRF-1 are very similar, with GMRF-1 being the model with a slightly better performance. Both HSMRF-1 and GMRF-1 outperform the order 2 models, with GMRF-2 being consistently the worst of the four models.

Our empirical results confirm the authors' claim that HSMRF respond to rapid changes through the use of the horseshoe prior. This is shown by the fact that in the presence of a sharp change in $N_e(t)$ HSMRF-1 and HSMRF-2 outperform the GMRF of the corresponding order. However, the choice of the order appears to be essential given that GMRF-1 largely outperforms HSMRF-2.

ORCID

Lorenzo Cappello https://orcid.org/0000-0001-6682-908X
Julia A. Palacios https://orcid.org/0000-0003-4501-7378

REFERENCES

Diaconis, P. and Wang, G. (2018) Bayesian goodness of fit tests: a conversation for David Mumford. *Annals of Mathematical Sciences and Applications*, 3, 287–301.

Faulkner, J.R., Magee, A.F., Shapiro, B. and Minin, V.N. (2020) Horseshoe-based Bayesian nonparametric estimation of effective population size trajectories. *Biometrics*. [In press]. Available at: arXiv:1808.04401.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. and Rubin, D.B. (2013) *Bayesian Data Analysis*. Boca Raton, FL: Chapman and Hall/CRC.

Gill, M., Lemey, P., Faria, N., Rambaut, A., Shapiro, B. and Suchard, M. (2013) Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Molecular Biology and Evolution*, 30, 713–724.

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, 1459–1471.

Palacios, J.A. and Minin, V.N. (2013) Gaussian process-based Bayesian nonparametric inference of population trajectories from gene genealogies. *Biometrics*, 63, 8–18.

How to cite this article: Cappello L, Ghosh S, Palacios JA. Discussion on "Horseshoe-based Bayesian nonparametric estimation of effective population size trajectories" by James R. Faulkner, Andrew F. Magee, Beth Shapiro, and Vladimir N. Minin. *Biometrics*. 2020;1–4. https://doi.org/10.1111/biom.13275