Are skyline plot-based demographic estimates overly dependent on smoothing prior assumptions?

Kris V Parag[1, 2, *], Oliver G Pybus[2], and Chieh-Hsi Wu[3]

[1]MRC Centre for Global Infectious Disease Analysis, Imperial College London, London, W2 1PG, UK
[2]Department of Zoology, University of Oxford, Oxford, OX1 3SY, UK
[3]Mathematical Sciences, University of Southampton, Highfield, Southampton SO17 1BJ, UK

[*]Email: k.parag@imperial.ac.uk

***Abstract*—In Bayesian phylogenetics, the coalescent process provides an informative framework for inferring dynamical changes in the effective size of a population from a sampled phylogeny (or tree) of its sequences. Popular coalescent inference methods such as the *Bayesian Skyline Plot*, *Skyride* and *Skygrid* all model this population size with a discontinuous, piecewise-constant likelihood but apply a smoothing prior to ensure that posterior population size estimates transition gradually with time. These prior distributions implicitly encode extra population size information that is not available from the observed coalescent tree (data). Here we present a novel statistic, Ω, to quantify and disaggregate the relative contributions of the coalescent data and prior assumptions to the resulting posterior estimate precision. Our statistic also measures the additional mutual information introduced by such priors. Using Ω we show that, because it is surprisingly easy to over-parametrise piecewise-constant population models, common smoothing priors can lead to overconfident and potentially misleading conclusions, even under robust experimental designs. We propose Ω as a useful tool for detecting when posterior estimate precision is overly reliant on prior choices.**

Key words: coalescent processes, skyline plots, prior assumptions, effective population size, phylodynamics, information theory.

I. Introduction

The coalescent process models how changes in the effective size of a target population influence the phylogenetic patterns of sequences sampled from that population. First derived by Kingman [1] under the assumption of a constant sized population, the coalescent process has since been extended to account for temporal variation in the population size [2], structured demographics [3] and multi-loci sampling [4]. Inference under these models aims to statistically recover the unknown effective population size (or demographic) history from the reconstructed phylogeny (or tree) and has provided insights in macro- and molecular biology [5], [6], [7]. Here we focus on coalescent processes that describe the genealogies of serially sampled individuals from deterministically varying populations. These are widely applied to study the phylodynamics of infections diseases [2] [8].

Early approaches to inferring effective population size from coalescent phylogenies used pre-defined parametric models (e.g. exponential or logistic growth functions) to represent temporal demographic changes [9], [7]. While these formulations required only a few variables and led to interpretable estimates, justifying their restrictive parametric assumptions was often difficult or computationally prohibitive [10]. This motivated the introduction of the *classic skyline plot* [11], which, by proposing an independent, piecewise-constant demographic change at every coalescent event (i.e at branching times in the phylogeny), maximised flexibility and removed all a-priori restrictions. However, this flexibility came at the cost of increased estimate noise and the risk of population size over-fitting [12].

Efforts to redress these issues, within a piecewise-constant framework, subsequently spawned a family of skyline plot-based methods [12]. Among these, the most popular and commonly-used are the *Bayesian Skyline Plot* (BSP) [13], the *Skyride* [10] and the *Skygrid* [14] (we denote the latter two S/S). All three attempted to regulate the sharp fluctuations of the inferred piecewise-constant demographic function by enforcing *a priori* assumptions about the smoothness (i.e. the level of autocorrelation among piecewise-constant segments) of real population dynamics. This was seen as a biologically sensible compromise between noise regulation and model flexibility.

The BSP limited overfitting by predefining fewer piecewise demographic changes than coalescent events and smoothed noise by asserting *a priori* that the population size after a change-point was exponentially distributed around the population size before it. This method was questioned by [10] for making strong smoothing and change-point assumptions and motivated development of the Skyride, which embeds the flexible classic skyline plot within a tunable Gaussian smoothing field. The Skygrid, which extends the Skyride to multiple loci and allows arbitrary change-points (the BSP and Skyride change-times coincide with coalescent events), also uses this prior. The S/S methods aimed to better tradeoff prior influence with noise reduction, and while somewhat effective, are still imperfect [15].

As a result, studies continue to address the non-trivial problem of optimising this tradeoff, either by searching for less-restrictive and more adaptive priors or by deriving new data-driven skyline change-point grouping strategies [15], [16]. The evolution of coalescent model inference thus reflects a desire to understand and fine-tune how prior assumptions and observed phylogenetic data interact to yield posterior population size estimates. Surprisingly, and in contrast to this desire, no study has yet tried to directly and rigorously measure the relative influence of the priors and data on these estimates.

Here we present a novel coalescent information theoretic ratio, Ω, to formally quantify and disaggregate the contributions of both priors and data on the uncertainty around the posterior estimates of skyline-based methods. Using Ω we illustrate how widely-used smoothing priors can lead to overconfident population size inferences and provide guidelines against such pitfalls. Our statistic can help detect when prior assumptions are inadvertently and overly influencing demographic estimates and will hopefully serve as a diagnostic tool that future methods can employ to optimise and validate their prior-data tradeoffs.

II. Preliminaries

A. Coalescent Inference

We provide an overview of the coalescent process and statistical inference under skyline plot-based demographic models. The coalescent is a stochastic process that describes the ancestral genealogy of sampled individuals or lineages from a target population [1]. Under the coalescent, a tree or phylogeny of relationships among these individuals is reconstructed backwards in time with coalescent events defined as the points where pairs of lineages merge (i.e. coalesce) into their ancestral lineage. This tree, $\mathcal{T}$, is rooted at time T into the past, which is the time to the most recent common ancestor (TMRCA) of the sample. The tips of $\mathcal{T}$ correspond to sampled individuals.

The rate at which coalescent events occur (i.e. the rate of branching in $\mathcal{T}$) is determined by and hence informative about the effective size of the target population. We assume that a total of $n \geq 2$ samples are taken from the target population at $n_s \geq 1$ distinct sampling times, which are independent of population size changes [13]. We do not specify the sample generating process as it does not affect our analysis by this independence assumption [17]. We let c_i be the time of the i^{th} coalescent event in $\mathcal{T}$ with $1 \leq i \leq n-1$ and $c_{n-1} = T$ (n samples can coalesce $n-1$ times before reaching the TMRCA).

We use l_t to count the number of lineages in $\mathcal{T}$ at time $t \geq 0$ into the past; l_t then decrements by 1 at every c_i and increases at sampling times. Here $t = 0$ is the present. The effective population size or demographic function at t is $N(t)$ so that the coalescent rate underlying $\mathcal{T}$ is $\binom{l_t}{2}N(t)^{-1}$ [1]. While $N(t)$ can be described using appropriate parametric formulations [18], it is more common to represent $N(t)$ by some tractable p-dimensional piecewise-constant approximation [12]. Thus, we can write $N(t) := \sum_{j=1}^{p} N_j 1(\epsilon_{j-1} \leq t < \epsilon_j)$, with $p \geq 1$ as the number of piecewise-constant segments. Here N_j is the constant population size of the j^{th} segment which is delimited by times $[\epsilon_{j-1},\, \epsilon_j)$, with $\epsilon_0 = 0$ and $\epsilon_p \geq T$ and $1(x)$ is an indicator function. The rate of producing coalescent events is then $\sum_{j=1}^{p} N_j^{-1}\binom{l_t}{2} 1(\epsilon_{j-1} \leq t < \epsilon_j)$. Note that Kingman's coalescent model is obtained by setting $p = 1$ (constant population of N_1).

When reconstructing the phylodynamic history of infectious diseases, it is often of interest to infer $N(t)$ from $\mathcal{T}$ [12], which forms our data generating process. If $\boldsymbol{N} = [N_1, ..., N_p]$ denotes the vector of demographic parameters to be estimated then the log-likelihood $\ell(\boldsymbol{N}) := \log \mathrm{P}(\mathcal{T} \mid \boldsymbol{N})$ can be obtained from [18] [19] as

$$\ell(\boldsymbol{N}) = \sum_{j=1}^{p} m_j \log N_j^{-1} - N_j^{-1}\omega_j + \log C_j, \tag{1}$$

with C_j as a constant that depends on the times and lineage counts of the m_j coalescent events that fall within the j^{th} segment duration $[\epsilon_{j-1},\, \epsilon_j)$, and $\sum_{j=1}^{p} m_j = n - 1$. Eq. (1) is equivalent to the standard temporally sampled skyline log-likelihood from [13], except that we do not restrict $N(t)$ to change only at coalescent times.

In Bayesian phylogenetic inference, skyline-based methods such as the BSP and S/S combine this likelihood with a prior distribution $\mathrm{P}(\boldsymbol{N})$, which encodes *a priori* beliefs about the demographic function. This yields a population size posterior by Bayes law as:

$$\mathrm{P}(\boldsymbol{N} \mid \mathcal{T}) \propto \mathrm{P}(\mathcal{T} \mid \boldsymbol{N})\mathrm{P}(\boldsymbol{N}). \tag{2}$$

Here we assume that the phylogeny, $\mathcal{T}$, is known without error. In some instances, only sampled sequence data, $\boldsymbol{D}$, is available and a distribution over $\mathcal{T}$ must be reconstructed from $\boldsymbol{D}$ under a model of molecular evolution with parameters $\boldsymbol{\theta}$. Eq. (2) is then embedded in the more complex Bayesian expression $\mathrm{P}(\mathcal{T}, \boldsymbol{\theta}, \boldsymbol{N} \mid \boldsymbol{D}) \propto \mathrm{P}(\boldsymbol{D} \mid \mathcal{T}, \boldsymbol{\theta}, \boldsymbol{N})\mathrm{P}(\boldsymbol{N} \mid \mathcal{T})\mathrm{P}(\boldsymbol{\theta})$, which involves inferring both the tree and population size. While we do not consider this extension here we note that results from this work are still applicable.

B. *Information and Estimation Theory*

We review and extend some concepts from information and estimation theory as applied to skyline-based coalescent inference. We consider a general parametrisation of the effective population size $\boldsymbol{\psi} = [\psi_1,\, \dots,\, \psi_p]$, where $\psi_i = \phi(N_i)$ for all $i \in \{1, ..., p\}$ and $\phi\,(\cdot)$ is a differentiable function. Popular skyline-based methods usually choose the identity function (e.g. BSP) or the natural logarithm (e.g. S/S) for ϕ. Eq. (1) and Eq. (2) are then reformulated with $\ell(\boldsymbol{\psi}) = \log \mathrm{P}(\mathcal{T} \mid \boldsymbol{\psi})$ as the log-likelihood and $\mathrm{P}(\boldsymbol{\psi})$ as the demographic prior. The Bayesian posterior, $\mathrm{P}(\boldsymbol{\psi} \mid \mathcal{T})$ combines this likelihood and prior, and hence is influenced by both the coalescent data and prior beliefs. We can formalise these influences using information theory.

The expected Fisher information, $\boldsymbol{\mathcal{I}}(\boldsymbol{\psi})$, is a $p \times p$ matrix with $(i,\, j)^{\text{th}}$ element $\boldsymbol{\mathcal{I}}(\boldsymbol{\psi})_{ij} := -\mathbb{E}_{\mathcal{T}}\left[\nabla_{ij}\ell(\boldsymbol{\psi})\right]$ [20]. The expectation is taken over the coalescent tree branches and $\nabla_{ij} := {\partial^2}/{\partial\psi_i\partial\psi_j}$. As observed in [17], $\boldsymbol{\mathcal{I}}(\boldsymbol{\psi})$ quantifies how precisely we can estimate the demographic parameters, $\boldsymbol{\psi}$, from the coalescent data, $\mathcal{T}$. Precision is defined as the inverse of variance [20]. The BSP and S/S parametrisations yield $\boldsymbol{\mathcal{I}}(\boldsymbol{N}) = [m_1 N_1^{-2},\, \dots,\, m_p N_p^{-2}]\,\mathrm{I}_p$ and $\boldsymbol{\mathcal{I}}(\log \boldsymbol{N}) = [m_1,\, \dots,\, m_p]\,\mathrm{I}_p$, with I_p as a $p \times p$ identity matrix [17]. These matrices provide several useful insights that we will exploit in later sections [17]. First, $\boldsymbol{\mathcal{I}}(\boldsymbol{\psi})$ is orthogonal (diagonal), meaning that the coalescent process over the j^{th} segment $[\epsilon_{j-1},\, \epsilon_j)$ can be treated as deriving from an independent Kingman coalescent with constant population size N_j [18]. Second, the number of coalescent events in that segment, m_j, controls the Fisher information available about N_j. Last, working under $\log N_j$ removes any dependence of this Fisher information component on the unknown parameter N_j.

The prior distribution, $\mathrm{P}\,(\boldsymbol{\psi})$, that is placed on the demographic parameters can alter and impact both estimate bias and precision. We can gauge prior-induced bias by comparing the maximum likelihood estimate (MLE), $\hat{\boldsymbol{\psi}} = \arg\max_{\boldsymbol{\psi}}\{\log \mathrm{P}(\mathcal{T} \mid \boldsymbol{\psi})\}$ with the maximum a posteriori estimate (MAP), $\tilde{\boldsymbol{\psi}} = \arg\max_{\boldsymbol{\psi}}\{\log \mathrm{P}(\mathcal{T} \mid \boldsymbol{\psi}) + \log \mathrm{P}(\boldsymbol{\psi})\}$ [21]. The difference $\tilde{\boldsymbol{\psi}} - \hat{\boldsymbol{\psi}}$ measures this bias. We can account for prior-induced precision by computing Fisher-type matrices for the prior and posterior as $\boldsymbol{\mathcal{P}}(\boldsymbol{\psi})_{ij} = -\nabla_{ij}\log \mathrm{P}(\boldsymbol{\psi})$ and $\boldsymbol{\mathcal{J}}(\boldsymbol{\psi})_{ij} = -\mathbb{E}_{\mathcal{T}}\left[\nabla_{ij}\log \mathrm{P}(\boldsymbol{\psi} \mid \mathcal{T})\right]$ [22] [23]. Combining these gives

$$\boldsymbol{\mathcal{J}}(\boldsymbol{\psi}) = \boldsymbol{\mathcal{I}}(\boldsymbol{\psi}) + \boldsymbol{\mathcal{P}}(\boldsymbol{\psi}). \tag{3}$$

Eq. (3) shows how the posterior Fisher information matrix, $\boldsymbol{\mathcal{J}}(\boldsymbol{\psi})$, relates to the standard Fisher information $\boldsymbol{\mathcal{I}}(\boldsymbol{\psi})$ and the prior second derivative $\boldsymbol{\mathcal{P}}(\boldsymbol{\psi})$. We make the common regularity assumptions (see [23] for details) that ensure $\boldsymbol{\mathcal{J}}(\boldsymbol{\psi})$ is positive definite and that all Fisher matrices exist. These assumptions are valid for exponential families such as the piecewise-constant coalescent [20][17]. Eq. (3) will prove fundamental to resolving the relative impact of the prior and data on the best precision achievable using $\mathrm{P}(\boldsymbol{N} \mid \mathcal{T})$. We also define expectations on these matrices with respect to the prior as $\boldsymbol{\mathcal{J}}_{\mathbf{0}}$, $\boldsymbol{\mathcal{I}}_{\mathbf{0}}$ and $\boldsymbol{\mathcal{P}}_{\mathbf{0}}$, with $\boldsymbol{\mathcal{J}}_{\mathbf{0}} = \mathbb{E}_0\left[\boldsymbol{\mathcal{J}}(\boldsymbol{\psi})\right] = \int \boldsymbol{\mathcal{J}}(\boldsymbol{\psi})\mathrm{P}(\boldsymbol{\psi})\,\mathrm{d}\boldsymbol{\psi}$, for example. These matrices are now constants instead of functions of $\boldsymbol{\psi}$. Eq. (3) also holds for these matrices [22].

These Fisher information matrices set theoretical upper bounds on the precision attainable by all possible statistical inference methods. For any unbiased estimate of $\boldsymbol{\psi}$, $\bar{\boldsymbol{\psi}}$, the Cramer-Rao bound (CRB) states that $\mathbb{E}_{\mathcal{T}}\left[(\bar{\boldsymbol{\psi}} - \boldsymbol{\psi})(\bar{\boldsymbol{\psi}} - \boldsymbol{\psi})^{\intercal} \mid \boldsymbol{\psi}\right] = \mathrm{var}(\bar{\boldsymbol{\psi}} \mid \boldsymbol{\psi}) \geq \boldsymbol{\mathcal{I}}(\boldsymbol{\psi})^{-1}$ with $\intercal$ indicating transpose. If we relax the unbiased requirement and include prior (distribution) information then the Bayesian or posterior Cramer-Rao lower bound (BCRB) controls the best estimate precision [21]. If $\bar{\boldsymbol{\psi}}$ is any estimator of $\boldsymbol{\psi}$ then the BCRB states that $\mathbb{E}_0\left[\mathbb{E}_{\mathcal{T}}\left[(\bar{\boldsymbol{\psi}} - \boldsymbol{\psi})(\bar{\boldsymbol{\psi}} - \boldsymbol{\psi})^{\intercal} \mid \boldsymbol{\psi}\right]\right] \geq \boldsymbol{\mathcal{J}}_{\mathbf{0}}^{\mathbf{-1}}$. This bound is not dependent on $\boldsymbol{\psi}$ due to the extra expectation over the prior [22].

The CRB describes how precisely we can estimate demographic parameters using just the coalescent data and is achieved (asymptotically) with equality for skyline (piecewise-constant) coalescent models [17]. The BCRB, instead, defines the precision limit for the combined contributions of the data and the prior. The CRB is a frequentist bound that assumes a true fixed $\boldsymbol{\psi}$, while the BCRB is a Bayesian bound treats $\boldsymbol{\psi}$ as a random parameter. The expectation over the prior connects the two formalisms [24]. Given their importance in delimiting precision, the $\boldsymbol{\mathcal{J}}(\boldsymbol{\psi})$ and $\boldsymbol{\mathcal{I}}(\boldsymbol{\psi})$ Fisher matrices will be central to our analysis, which focuses on resolving the individual contributions of the data versus prior assumptions.

III. Results

A. *The Coalescent Information Ratio, Ω*

We propose and derive the coalescent information ratio, Ω, as a statistic for evaluating the relative contributions of the prior and data to the posterior estimates obtained as solutions to Bayesian skyline inference problems (see Preliminaries). Consider such a problem in which the n-tip phylogeny $\mathcal{T}$ is used to estimate the p-element demographic parameter vector $\boldsymbol{\psi}$. Let $\hat{\boldsymbol{\psi}}$ be the MLE of $\boldsymbol{\psi}$ given the data $\mathcal{T}$. Asymptotically, the uncertainty around this MLE can be described with a multivariate Gaussian distribution with covariance matrix $\boldsymbol{\mathcal{I}}(\boldsymbol{\psi})^{-1}$. The Fisher information, $\boldsymbol{\mathcal{I}}(\boldsymbol{\psi})$ then defines a confidence ellipsoid that circumscribes the total uncertainty from this distribution. In [17] this ellipsoid was found central to understanding the statistical properties of skyline-based estimates.

The volume of this ellipsoid is $V_1 = C \det [\boldsymbol{\mathcal{I}}(\boldsymbol{\psi})]^{-\frac{1}{2}}$, with C as a p-dependent constant. Decreasing V_1 increases the best estimate precision attainable from the data $\mathcal{T}$ [17]. In a Bayesian framework, the asymptotic posterior distribution of $\boldsymbol{\psi}$ also follows a multivariate Gaussian distribution with covariance matrix of $\boldsymbol{\mathcal{J}}(\boldsymbol{\psi})^{-1}$. We can therefore construct an analogous ellipsoid from $\boldsymbol{\mathcal{J}}(\boldsymbol{\psi})$ with volume $V_2 = C \det [\boldsymbol{\mathcal{J}}(\boldsymbol{\psi})]^{-\frac{1}{2}}$ that measures the uncertainty around the MAP estimate $\tilde{\boldsymbol{\psi}}$. This volume includes the effect of both prior and data on estimate precision. Accordingly, we propose the ratio

$$\Omega := \frac{V_2}{V_1} = \sqrt{\frac{\det [\boldsymbol{\mathcal{I}}(\boldsymbol{\psi})]}{\det [\boldsymbol{\mathcal{I}}(\boldsymbol{\psi}) + \boldsymbol{\mathcal{P}}(\boldsymbol{\psi})]}}, \tag{4}$$

as a novel statistic for dissecting the relative impact of data and prior on posterior estimate precision. A visual example of Ω, as applied to a later smoothing-prior problem, is provided in Fig. 2.

Further, observe $0 \leq \Omega \leq 1$ and importantly that

$$\Omega^2 \leq \frac{1}{2} \iff \det [\boldsymbol{\mathcal{I}}(\boldsymbol{\psi})] \leq \frac{1}{2} \det [\boldsymbol{\mathcal{P}}(\boldsymbol{\psi}) + \boldsymbol{\mathcal{I}}(\boldsymbol{\psi})] . \tag{5}$$

At this threshold value $\boldsymbol{\mathcal{P}}(\boldsymbol{\psi})$ contributes at least as much information as the data. Moreover, $\lim_{n \to \infty} \Omega = 1$ since the prior contribution becomes negligible with increasing data and Ω is undefined when $\boldsymbol{\psi}$ is unidentifiable from $\mathcal{T}$ (i.e. when $\boldsymbol{\mathcal{I}}(\boldsymbol{\psi})$ is singular [25]). Consequently, we posit that a smaller Ω implies the prior provides a greater contribution to estimate precision.

We define Ω as an information ratio due to its close connection to both the Fisher and mutual information. The mutual information between $\boldsymbol{\psi}$ and $\mathcal{T}$, $\mathbb{I}(\boldsymbol{\psi};\, \mathcal{T})$, measures how much information (in bits for example) $\mathcal{T}$ contains about $\boldsymbol{\psi}$ [26]. This is distinct but related to $\boldsymbol{\mathcal{I}}(\boldsymbol{\psi})$, which quantifies the precision of estimating $\boldsymbol{\psi}$ from $\mathcal{T}$ [27]. Recent work from [23] into the connection between the Fisher and mutual information has yielded two key approximations to $\mathbb{I}(\boldsymbol{\psi};\, \mathcal{T})$. These can be obtained by substituting either $\boldsymbol{\mathcal{I}}$ or $\boldsymbol{\mathcal{J}}$ for $\boldsymbol{\mathcal{X}}$ in

$$\mathbb{I}(\boldsymbol{\mathcal{X}}) = \mathcal{H}(\boldsymbol{\psi}) + \mathbb{E}_0 \left[\log \sqrt{\det [\boldsymbol{\mathcal{X}}(\boldsymbol{\psi})]} - p \log \sqrt{2\pi e} \right] . \tag{6}$$

Here $\mathcal{H}(\boldsymbol{\psi}) := \mathbb{E}_0 [- \log \mathrm{P}(\boldsymbol{\psi})]$ is the differential entropy of $\boldsymbol{\psi}$ [26].

These approximations were used in [23] to characterise how much information, about a stimulus $\boldsymbol{\psi}$, is encoded by a large population of neurons. The outputs of those neurons would be equivalent to $\mathcal{T}$ in our notation. For a flat prior or many observations, $\mathbb{I}(\boldsymbol{\psi};\, \mathcal{T}) \approx \mathbb{I}(\boldsymbol{\mathcal{I}}) \approx \mathbb{I}(\boldsymbol{\mathcal{J}})$, as the prior is contributing little or no information [27]. For sharper priors, $\mathbb{I}(\boldsymbol{\psi};\, \mathcal{T}) \approx \mathbb{I}(\boldsymbol{\mathcal{J}})$ as the prior contribution is not negligible. In that case, using $\mathbb{I}(\boldsymbol{\mathcal{I}})$ can lead to large errors [23].

Eq. (6) is predicated on (i) regularity assumptions for the distributions used (i.e. that the second derivatives exist), (ii) conditional dependence of the observed data given $\boldsymbol{\psi}$ and (iii) that the likelihood is peaked around its most probable value [20], [27], [23]. The skyline-based inference problems that we consider here automatically satisfy (i) and (ii) as these models belong to an exponential family. Stipulation (iii) is satisfied for moderate to large trees (and hence asymptotically) [20], [17]. Using the above approximations, we find

$$\Delta \mathbb{I} = \mathbb{I}(\boldsymbol{\mathcal{I}} + \boldsymbol{\mathcal{P}}) - \mathbb{I}(\boldsymbol{\mathcal{I}}) = \mathbb{E}_0 [- \log \Omega] . \tag{7}$$

Eq. (7) suggests that our ratio directly measures the excess mutual information introduced by the prior, providing a substantive link between how sharper estimate precision is attained with extra mutual information. Observe that both sides of Eq. (7) diminish when $\mathcal{P}(\boldsymbol{\psi}) \ll \mathcal{I}(\boldsymbol{\psi})$. Because the mutual information and its approximations (see Eq. (6)) are invariant to invertible parameter transformations [23], our coalescent information ratio does not depend on whether we infer $\boldsymbol{N}$, its inverse, or its logarithm.

Moreover, we can use normalising transformations to make Ω valid at even small tree sizes. In [28] several such transformations for exponentially distributed models like the coalescent are derived. Among them, the log transform can achieve approximately normal log-likelihoods for about 7 observations and above ($n \geq 8$). Thus, $\log \boldsymbol{N}$, which is also optimal for experimental design [17], ensures the validity of Ω on small trees. This is the parametrisation adopted by the S/S methods [10]. Other (cubic-root) parametrisations under which Ω would be valid at even smaller n also exist [28].

Eq. (4)–Eq. (7) are not restricted to coalescent inference problems and are generally applicable to statistical models that involve exponential families [20]. We now specify Ω for skyline-based models, which all possess piecewise-constant population sizes and orthogonal $\boldsymbol{\mathcal{I}}(\boldsymbol{\psi})$ matrices [17]. These properties permit the expansion [29]:

$$\det [\boldsymbol{\mathcal{I}}(\boldsymbol{\psi}) + \boldsymbol{\mathcal{P}}(\boldsymbol{\psi})] = \det [\boldsymbol{\mathcal{I}}(\boldsymbol{\psi})] + \det [\boldsymbol{\mathcal{P}}(\boldsymbol{\psi})] + \sum_{j=1}^{p-1} \gamma_j ,$$

$$\text{with } \gamma_j = \sum d_{i_1} \dots d_{i_j} \det \left[\boldsymbol{\mathcal{P}}(\boldsymbol{\psi})_{\bar{\boldsymbol{i}}_1 \dots \bar{\boldsymbol{i}}_j} \right] ,$$

where d_k are the diagonal elements of $\boldsymbol{\mathcal{I}}(\boldsymbol{\psi})$ with $1 \leq i_1 < \dots < i_j \leq p$, and $\boldsymbol{\mathcal{P}}(\boldsymbol{\psi})_{\bar{\boldsymbol{i}}_1 \dots \bar{\boldsymbol{i}}_j}$ is the sub-matrix formed by deleting the $(i_1, \, \dots, \, i_j)^{\text{th}}$ rows and columns of $\boldsymbol{\mathcal{P}}(\boldsymbol{\psi})$ [29].

This allows us to formulate a prior signal-to-noise ratio

$$r = \prod_{j=1}^{p} d_j^{-1} \left(\det [\mathcal{P}(\boldsymbol{\psi})] + \sum_{k=1}^{p-1} \gamma_k \right) \implies \Omega = \sqrt{\frac{1}{1+r}}, \tag{8}$$

which quantifies the relative excess Fisher information (the ‘signal’) that is introduced by the prior. This ratio signifies when the prior contribution overwhelms that of the data i.e. $r > 1 \iff \Omega^2 < \frac{1}{2}$. Having derived theoretically meaningful metrics for resolving prior-data precision contributions, we next investigate their ramifications.

B. The Kingman Conjugate Prior

Kingman’s coalescent process [1], which describes the phylogeny of a constant sized population N_1, is the foundation of all skyline model formulations. Specifically, a p-dimensional skyline model is analogous to having p Kingman coalescent models, the j^{th} of which is valid over $[\epsilon_{j-1}, \epsilon_j)$ and describes the genealogy under population size N_j. Here we use Kingman’s coalescent to validate and clarify the utility of Ω as a measure of relative data-prior precision contributions.

We assume an n-tip Kingman coalescent tree, $\mathcal{T}$ and initially work with the inverse parametrisation, N_1^{-1}. We scale $\mathcal{T}$ at t by $\binom{l_t}{2}$ as in [18] so that $\binom{l_{c_{i-1}}}{2}(c_i - c_{i-1}) \sim \exp(N_1^{-1})$ for $1 \leq i \leq n-1$ with $c_0 = 0$. If y defines the space of N_1^{-1} values, and has prior distribution $\mathrm{P}(y)$, then, by [19], [18], its posterior is

$$\mathrm{P}(y \,|\, \mathcal{T}) = \frac{A y^{n-1} e^{-y\bar{T}} \mathrm{P}(y)}{\int_0^\infty A y^{n-1} e^{-y\bar{T}} \mathrm{P}(y) \, \mathrm{d}y} \quad \text{with} \quad A = \prod_{i=2}^{n} \binom{i}{2},$$

where A is a constant and $\bar{T}$ is the scaled TMRCA of $\mathcal{T}$.

The likelihood function embedded within $\mathrm{P}(y \,|\, \mathcal{T})$ is proportional to a shape-rate parametrised gamma distribution, with known shape n. The conjugate prior for N_1^{-1} is also gamma [30] i.e. $N_1^{-1} \sim \mathrm{Gam}\left(m_0, \, \bar{T}_0\right)$ with $m = n - 1$ counting the coalescent events in $\mathcal{T}$. The posterior distribution is then $N_1^{-1} \,|\, \mathcal{T} \sim \mathrm{Gam}\left(m + m_0, \, \bar{T} + \bar{T}_0\right)$ [31]. Transforming back to N_1, this implies that $N_1 \,|\, \mathcal{T} \sim \mathrm{Gam}^{-1}\left(m + m_0, \, \bar{T} + \bar{T}_0\right)$. This is an inverse gamma distribution with mean $\frac{\bar{T} + \bar{T}_0}{m + m_0 - 1}$. If x describes the space of possible N_1 values and $\Gamma(s) := \int_0^\infty z^{s-1} e^{-z} \, \mathrm{d}z$ then

$$\mathrm{P}(x \,|\, \mathcal{T}) = \frac{(\bar{T} + \bar{T}_0)^{(m+m_0)}}{\Gamma(m + m_0)} x^{-(m+m_0+1)} e^{-\frac{\bar{T} + \bar{T}_0}{x}} .$$

We can interpret the parameters of the gamma posterior distribution as involving a prior contribution of $m_0 - 1$ coalescent events from a virtual tree, $\mathcal{T}_0$, with scaled TMRCA $\bar{T}_0$. This is then combined with the actual data, which contributes m coalescent events from $\mathcal{T}$, with scaled TMRCA of $\bar{T}$ [31]. This offers a very clear breakdown of how our posterior estimate precision is derived from prior and likelihood contributions, and suggests that if $\mathcal{T}_0$ has more tips than $\mathcal{T}$ then we are depending more on the prior than the data. We now calculate Ω to determine if we can formalise this intuition.

The Fisher information values of N_1^{-1} are $\mathcal{I}(N_1^{-1}) = mN_1^2$ and $\mathcal{J}(N_1^{-1}) = (m + m_0 - 1)N_1^2$. The information ratio and mutual information difference, $\Delta\mathbb{I}$, which hold for all parametrisations, then follow from Eq. (4), Eq. (7) and Eq. (8) as

$$\Omega^2 = \frac{1}{1+r} \approx 1 - r, \qquad \Delta\mathbb{I} = \frac{1}{2}\log(1+r) \approx \frac{1}{2}r, \tag{9}$$

with $r = \frac{m_0-1}{m}$, as the signal-to-noise ratio. The approximations shown are valid when $r \ll 1$. Interestingly, when $m_0 - 1 = m$ so that $r = 1$, we get $\Omega^2 = {}^1\!/_2$ (see Eq. (5)). This exactly quantifies the relative impact of real and virtual observations described previously. At this point we are being equally informed by both the conjugate prior and the likelihood. Prior over-reliance can be defined by the threshold condition of $r > 1 \implies \Omega^2 < {}^1\!/_2$.

The expression of $\Delta\mathbb{I}$ confirms our interpretation of r as an effective signal-to-noise ratio controlling the extra mutual information introduced by the conjugate prior. This can be seen by comparison with the standard Shannon mutual information expressions from information theory [26]. At small r, where the data dominates, we find that the prior linearly detracts from Ω^2 and linearly increases $\Delta\mathbb{I}$. We also observe that $\bar{T}_0$, the gamma prior shape parameter, has no effect on estimate precision or mutual information.

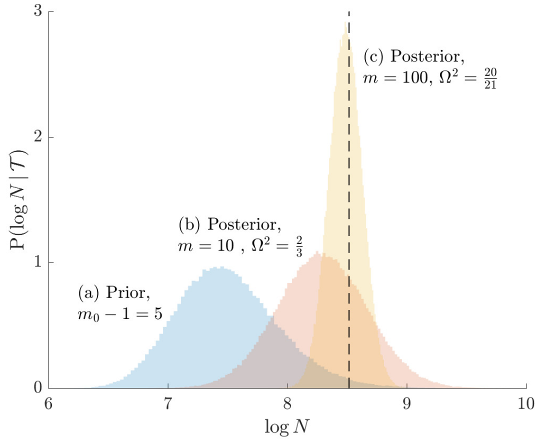

Fig. 1: **Effect of conjugate prior on Kingman coalescent estimation.** We examine the relative impact on estimate precision of a conjugate Kingman prior that contributes $m_0 - 1 = 5$ virtual observations. We work in $\log N_1$ for convenience. We compare this prior to posteriors, which are obtained under observed trees with $m = 10$ (red) and $m = 100$ (yellow) coalescent events. The true value is in black. The prior contribution decays as Ω^2 increases.

Our ratio Ω therefore provides a systematic decomposition of the estimate precision and generalises the virtual observation idea to any prior distribution. In essence, the prior is contributing an effective sample size, which for the conjugate Kingman prior is $m_0 - 1$. We summarise these points in Fig. 1, which shows the conjugate prior and two posteriors together with their corresponding Ω^2 values.

C. Skyline Smoothing Priors

We specialise Ω for the BSP [13] and S/S coalescent inference methods [10], [14]. These popular skyline-based approaches couple a piecewise-constant demographic likelihood with a smoothing prior to produce population size estimates that change more continuously with time. The smoothing prior achieves this by assuming informative relationships between N_j and its neighbouring parameters (N_{j-1}, N_{j+1}). Such *a priori* correlation implicitly introduces additional demographic information that is not available from the data $\mathcal{T}$. While these priors can embody sensible biological assumptions, we show that they can also engender overconfident statements or obscure parameter non-identifiability. We propose Ω as a simple but meaningful analytic for diagnosing these problems.

We first define a uniquely objective (i.e. uninformative) reference skyline priors, $\mathrm{P}^*(\boldsymbol{\psi})$. Finding objective priors for multivariate statistical models is generally non-trivial, but [32] states that if $\boldsymbol{\mathcal{I}}(\boldsymbol{\psi})$ has form $[f_1(\psi_1)g_1(\boldsymbol{\psi_{-1}}), \ldots, f_p(\psi_1)g_p(\boldsymbol{\psi_{-p}})]\ \mathrm{I}_p$ then $\mathrm{P}^*(\boldsymbol{\psi}) \propto \prod_{j=1}^{p}\sqrt{f_j(\psi_j)}$. Here f_j and g_j represent some functions and $\boldsymbol{\psi_{-j}}$ symbolises the vector $\boldsymbol{\psi}$ excluding ψ_j. Following this, we obtain

$$\mathrm{P}^*(\boldsymbol{\psi} = \boldsymbol{N}) = Z_1^{-1}\prod_{j=1}^{p} N_j^{-1} \text{ and } \mathrm{P}^*(\boldsymbol{\psi} = \log \boldsymbol{N}) = Z_2^{-1},$$

with Z_1, Z_2 as normalisation constants. Given its optimal properties [17], we only consider $\boldsymbol{\psi} = \log \boldsymbol{N}$, and drop explicit notational references to it. Under this parametrisation, $\boldsymbol{\mathcal{I}}$ and its expectation with respect to the prior are equal, i.e. $\mathbb{E}_0[\boldsymbol{\mathcal{I}}] = \boldsymbol{\mathcal{I}_0}$. In addition, the reference prior in this case is $\boldsymbol{\mathcal{P}^*} = \boldsymbol{0_p}$, with $\boldsymbol{0_p}$ as a matrix of zeros. This yields $\Omega = 1$ by Eq. (4). A uniform prior over log-population space is hence uniquely objective for skyline inference.

Other prior distributions, which are termed subjective, necessarily introduce extra information and contribute to posterior estimate precision. This contribution will be reflected by an $\Omega < 1$. The two most widely-used, subjective, skyline plot smoothing priors are:

(i) the *Sequential Markov Prior* (SMP) used in the BSP [13], and
(ii) the *Gaussian Markov Random Field* (GMRF) prior employed in the S/S methods [10] [14].

As the SMP and GMRF both propose nearest neighbour autocorrelations among elements of $\boldsymbol{\psi}$, tridiagonal posterior Fisher information matrices result. We denote these $\boldsymbol{\mathcal{J}}_{\text{SMP}}$ and $\boldsymbol{\mathcal{J}}_{\text{GMRF}}$, respectively.

The SMP is defined as: $\mathrm{P}(\boldsymbol{N}) = {}^1\!/_{N_1}\prod_{j=2}^{m}{}^1\!/_{N_{j-1}}\, e^{N_j/N_{j-1}}$ [13]. It assumes that $N_j \sim \exp(N_{j-1}^{-1})$ with a prior mean of N_{j-1}. An objective prior is used for N_1. To adapt this for $\log \boldsymbol{N}$, we define $l_j = e^{\log N_{j+1} - \log N_j} = {}^{N_{j+1}}\!/_{N_j}$ for $j \in \{1, \ldots, p-1\}$. In Appendix A we show how this expression yields Eq. (A.1) and hence the transformed prior $\mathrm{P}(\log \boldsymbol{N}) = \prod_{j=1}^{p-1} l_j e^{-l_j}$. We then take relevant derivatives to obtain $\boldsymbol{\mathcal{J}}_{\text{SMP}}$, which for the minimally representative $p = 3$ case is written as:

$$\boldsymbol{\mathcal{J}}_{\text{SMP}} = \begin{bmatrix} m_1 + \frac{N_2}{N_1} & -\frac{N_2}{N_1} & 0 \\ -\frac{N_2}{N_1} & m_2 + \frac{N_2}{N_1} + \frac{N_3}{N_2} & -\frac{N_3}{N_2} \\ 0 & -\frac{N_3}{N_2} & m_3 + \frac{N_3}{N_2} \end{bmatrix}. \tag{10}$$

The $p > 3$ matrices simply extend the pattern in Eq. (10).

An issue with the SMP is its dependence on the unknown 'true' demographic parameter values. Consequently, we cannot evaluate (or control) *a priori* how much information is contributed by this smoothing prior. Exponentially growing populations could feature ${}^{N_{j+1}}\!/_{N_j} > m_j$, for example, which would result in prior over-reliance. Conversely, rapidly declining populations would be more data-dependent. This likely reflects the asymmetry in using sequential exponential distributions. The only control we have on smoothing implicitly emerges from choosing the number of segments, p.

The possibility of strong or inflexible prior assumptions under the BSP motivated the development of the GMRF for the S/S methods [10]. The GMRF works directly with $\log \boldsymbol{N}$ and models the autocorrelation between neighbouring segments with multivariate Gaussian distributions. The GMRF prior is defined as $\mathrm{P}(\log \boldsymbol{N}) = Z^{-1}\tau^{\frac{p-2}{2}}e^{-\frac{\tau}{2}\sum_{j=1}^{p-1}\delta_j^{-1}(\log N_{j+1} - \log N_j)^2}$ [10]. In this model, Z is a normalisation constant, τ a smoothing parameter, to which a gamma prior is often applied, and the δ_j values adjust for the duration of the piecewise-constant skyline segments. Usually either (i) δ_j is chosen based on the inter-coalescent midpoints in $\mathcal{T}$ or (ii) a uniform GMRF is assumed with $\delta_j = 1$ for every $j \in \{1, \ldots, m-1\}$.

Similarly, we calculate $\boldsymbol{\mathcal{J}}_{\text{GMRF}}$ for the $p = 3$ case, which is:

$$\boldsymbol{\mathcal{J}}_{\text{GMRF}} = \begin{bmatrix} m_1 + \frac{\tau}{\delta_1} & -\frac{\tau}{\delta_1} & 0 \\ -\frac{\tau}{\delta_1} & m_2 + \frac{\tau}{\delta_1} + \frac{\tau}{\delta_2} & -\frac{\tau}{\delta_2} \\ 0 & -\frac{\tau}{\delta_2} & m_3 + \frac{\tau}{\delta_2} \end{bmatrix}. \tag{11}$$

Appendix A provides the general derivation for any $p \geq 3$. As τ is arbitrary and the δ_j depend only on $\mathcal{T}$, the GMRF is insensitive to the unknown parameter values. This property makes it more desirable than the SMP and gives us some control (via τ) of the level of smoothing introduced. Nevertheless, the next section demonstrates that this model still tends to over-smooth demographic estimates.

We diagonalise $\boldsymbol{\mathcal{J}}_{\text{GMRF}}$ and $\boldsymbol{\mathcal{J}}_{\text{SMP}}$ to obtain matrices of form $\boldsymbol{\mathcal{J}} = \boldsymbol{S}\boldsymbol{\mathcal{Q}}\boldsymbol{S}^{\intercal}$. Here $\boldsymbol{S}$ is an orthogonal transformation matrix (i.e. $|\det[\boldsymbol{S}]| = 1$) and $\boldsymbol{\mathcal{Q}} = [\lambda_1, \ldots, \lambda_p]\,\mathrm{I}_p$ with λ_j as the j^{th} eigenvalue of $\boldsymbol{\mathcal{J}}$. Since $\det[\boldsymbol{J}] = \det[\boldsymbol{\mathcal{Q}}]$, we can use Eq. (4) to find that $\Omega = \prod_{j=1}^{p}\sqrt{m_j/\lambda_j}$. This equality reveals that λ_j acts as a prior perturbed version of m_j. When objective reference priors are used we recover $m_j = \lambda_j$ and $\Omega = 1$. We can use the $\boldsymbol{S}$ matrix to gain insight into how the GMRF and SMP encode population size correlations. The principal components of our posterior demographic estimates (which are obtained from $\mathrm{P}(\log \boldsymbol{N} \,|\, \mathcal{T})$) are the vectors forming the axes of the uncertainty ellipsoid described by $\boldsymbol{\mathcal{J}}$.

These principal component vectors take the form $\{e_1, \ldots, e_p\} = \{(\log N_1, 0, \ldots, 0)^{\intercal}, \ldots (0, 0, \ldots, \log N_p)^{\intercal}\}$ when we apply the reference prior $\mathrm{P}^*(\log \boldsymbol{N})$. Thus, as we would expect, our uncertainty ellipses are centred on the parameters we wish to infer. However, if we use the GMRF prior these axes are instead transformed to $\{\boldsymbol{S}e_1, \ldots, \boldsymbol{S}e_p\}$. These new axes are linear combinations of $\log \boldsymbol{N}$ and elucidate how smoothing priors share information (i.e. introduce autocorrelations) about $\log \boldsymbol{N}$ across its elements. These geometrical changes hint at how smoothing priors influence the statistical properties of our coalescent inference problem.

Lastly, we provide a visualisation of Ω and an example of $\boldsymbol{S}$. We consider the $p = 2$ case, where the posterior Fisher information and Ω for both the GMRF and SMP take the form:

$$\boldsymbol{\mathcal{J}} = \begin{bmatrix} m_1 + a & -a \\ -a & m_2 + a \end{bmatrix} \implies \Omega^2 = \frac{1}{1 + a\frac{m_1 m_2}{m_1 + m_2}}, \tag{12}$$

with $a = \tau/\delta_1$ for the GMRF and $a = N_2/N_1$ for the SMP. The signal-to-noise ratio is $r = a\frac{m_1 m_2}{m_1 + m_2}$ (see Eq. (9)) and performance clearly depends on how the m coalescent events in $\mathcal{T}$ are apportioned between the two population size segments.

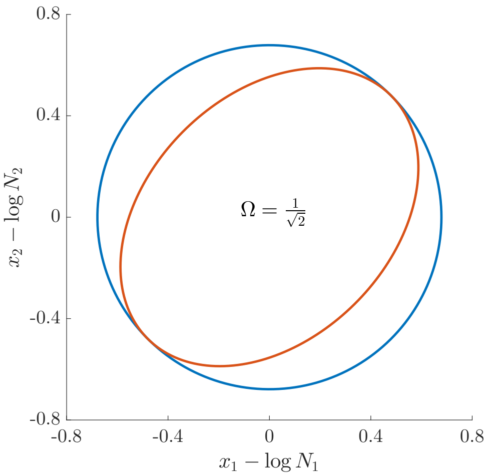

Fig. 2: **Uncertainty ellipses for SMP and GMRF.** We show the improvement in asymptotic precision rendered by use of a smoothing prior for a $p = 2$ segment skyline inference problem. The prior informed ellipse (red) is smaller in volume and has skewed principal axes relative to the purely data informed one (blue). All ellipses represent 99% confidence. The covariance that smoothing introduces controls this skew. Here $\Omega^2 = 1/2$, $m = 40$ and $a = 10$. We posit that larger a values lead to over-reliance on the smoothing prior.

We can lower bound the contribution of these priors to Ω under any (m_1, m_2) settings by using the robust coalescent design from [17]. This design stipulates that we define our skyline segments such that $m_1 = m_2 = m/2$ in order to optimise estimate precision under $\mathcal{T}$. At this robust point we also find that $\max_{\{m_j\}}\Omega^2$ (or $\min_{\{m_j\}} r$) is attained. Fig. 2 gives the uncertainty ellipses for this robust $p = 2$ model at $a = m/4$. Moreover, at $m_1 = m_2 = m/2$, for any r, we can calculate the diagonalisable transformation for $\boldsymbol{\mathcal{J}}$ to get:

$$\boldsymbol{\mathcal{Q}} = \begin{bmatrix} \frac{m}{2} & 0 \\ 0 & \frac{m}{2} + 2a \end{bmatrix}, \qquad \boldsymbol{S} = \begin{bmatrix} \cos(\frac{\pi}{4}) & -\sin(\frac{\pi}{4}) \\ \sin(\frac{\pi}{4}) & \cos(\frac{\pi}{4}) \end{bmatrix}. \tag{13}$$

Applying $\boldsymbol{S}$, we find that the axes of our uncertainty ellipse change from $\{\binom{\log N_1}{0}, \binom{0}{\log N_2}\}$ to $\{\binom{\log N_1 - \log N_2}{0}, \binom{0}{\log N_1 + \log N_2}\}$. Sums and differences of log-populations are now the parameters that can be most naturally estimated under the SMP and GMRF.

D. The Dangers of Smoothing

Having defined ratios for measuring the contribution of smoothing priors to estimate precision, we now use them to explore and expose the conditions under which prior over-reliance is likely to occur in practice. We assume that skyline segments are chosen to satisfy the robust design $m_j = m/p$ for $1 \leq j \leq p$ [17]. We previously proved that robust designs, at $p = 2$, minimise dependence on the prior (maximise Ω). While this is not the case for $p > 2$, in Fig. B.1 of the Appendix we illustrate that the maximal Ω point is generally well approximated by this robust setting. The Ω values computed here are therefore conservative for most $\{m_j\}$ settings. Other experimental designs rely more on the prior.

As in Eq. (5), we use the $\Omega^2 = 1/2$ threshold to diagnose when the data $\mathcal{T}$ (likelihood) and prior are equally influencing demographic posterior estimate precision. At $\Omega^2 = 1/2$ the total Fisher information doubles since $\det[\boldsymbol{\mathcal{J}}] = 2\det[\boldsymbol{\mathcal{I}}]$. We previously uncovered the importance of this threshold in the Kingman conjugate prior problem, where it signified an equality between the number of pseudo and real samples contributed by the prior and data, respectively. As $\Omega^2 = \frac{1}{1+r}$ (see Eq. (8)), this setting os also meaningful because it achieves a unit signal-to-noise ratio for any skyline-based model.

We first reconsider the $p = 2$ case of Eq. (12). Here $\Omega^2 = 1/2$ suggests $a = m/4$, which implies that we are overly-reliant on smoothing when a is larger than $1/4$ of the total observed coalescent events. This occurs when $N_2 \geq m/4\,N_1$ or $\tau \geq m/4\,\delta_1$, for the SMP and GMRF respectively. The improved precision due to the prior at this $^m/_4$ threshold is shown in Fig. 2. The relative ellipse area (and hence Ω) will shrink further as we deviate from robust designs.

As p increases, smoothing becomes more influential and can promote misleading conclusions. For the $p > 2$ cases, we will only examine the GMRF, since the SMP has the undesirable property of dependence on the unknown N_j values. To better expose the impact of the smoothing parameter τ, we will assume a uniform GMRF ($\{\delta_j\} = 1$) so that $\mathcal{J}_{\text{GMRF}}$ then only depends on $\{m_j\}$ and τ. We compute r and hence Ω, at various p. For example we find that

$$r \mid_{p=3} = \left(^{27}/_{m^2}\right) \tau^2 + \left(^{12}/_m\right) \tau;$$
$$r \mid_{p=4} = \left(^{256}/_{m^3}\right) \tau^3 + \left(^{160}/_{m^2}\right) \tau^2 + \left(^{24}/_m\right) \tau,$$

under the robust design. Interestingly, the order of the polynomial dependence of r (and hence Ω) on τ increases with p. We find that this trend holds for any $\{m_j\}$ design. We will use the term robust Ω for when Ω is calculated under a robust design.

Fig. 3 plots the robust Ω against τ and p for the uniform GMRF. A key feature of Fig. 3 is the steep p-dependent decay of Ω relative to the $\Omega^2 = {}^1/_2$ threshold, which exposes how easily we can be unduly reliant on the prior, as p increases. Given a phylogeny $\mathcal{T}$, increasing the complexity of a skyline-based model enhances the dependence of our posterior estimate precision on the smoothing prior. This pattern is intuitive as fewer coalescent events now inform each demographic parameter [17]. However, Ω decays with surprising speed. For example, at $p = 20$ (the lowest curve in Fig. 3) we get $\Omega < 0.1$ for $\tau = 1$ and $m = 100$. Usually, τ has a gamma-prior with mean of 1 [10]. We show the corresponding mutual information increases due to these GMRF priors in Fig. B.2 of the Appendix.

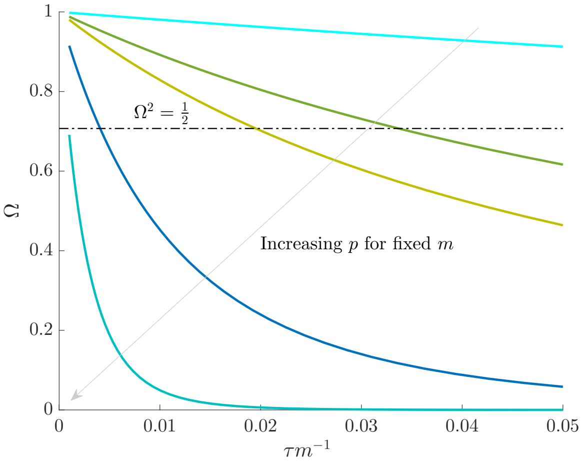

Fig. 3: **The impact of smoothing priors increases with skyline complexity.** For the GMRF, we find that for a fixed $^\tau/_m$, Ω significantly depends on the complexity, p, of our skyline. The coloured Ω curves are (along the arrow) for $p = [2, 4, 5, 10, 20]$ at $m = 100$ with $m_j = {}^m/_p$ (robust design point [17]). The dashed $\Omega^2 = {}^1/_2$ line depicts the threshold below which the prior contributes more than the data to posterior precision (asymptotically). For a given tree and τ, the larger the number of demographic parameters we choose to estimate, the stronger the influence of the prior on those estimates.

While Fig. 3 might seem specific to the uniform GMRF, it is broadly applicable to the BSP and S/S methods. We now outline the implications of Fig. 3 for each of these skyline-based approaches.

(1) Bayesian Skyline Plot: This method uses the SMP, which depends on the unknown N_j values. However, the results of Fig. 3 are valid if we set τ to $\min_{\{1 \leq j \leq p-1\}} {}^{N_{j+1}}/_{N_j}$, which results in the smallest non-data contribution to Eq. (10). This follows as $\mathcal{J}_{\text{GMRF}}$ and $\mathcal{J}_{\text{SMP}}$ have similar forms. While this choice underestimates the impact of the SMP, it still cautions against high-p skylines and confirms known BSP issues related to poor estimation precision when skylines are too complex, or the data are not sufficiently informative [12]. However, strong use of BSP grouping parameter [13], which sets $p < m$, could alleviate some of these problems.

(2) Skyride: When this method uses the uniform GMRF, all results are exactly applicable. In its full implementation, the Skyride employs a time-aware GMRF that sets δ_j based on $\mathcal{T}$ and estimates τ from the data [10]. However, even with these adjustments, the GMRF can over-smooth, and fail to recover population size changes [12], [15]. Our results provide a theoretical grounding for this observation. The Skyride constrains $p = m$ and then smooths this noisy piecewise model. Consequently, it constructs a skyline which is too complex by our measures (the lowest curve in Fig. 3 is at $p = {}^m/_5$). By rescaling the smoothing parameter to $\min_{\{1 \leq j \leq p-1\}} {}^\tau/_{\delta_j}$, the Ω curves in Fig. 3 upper bound the true Ω values of the time-aware GMRF.

(3) Skygrid: This method uses a scaled GMRF. For a tree with TMRCA T, the Skygrid assumes new population size segments every $^T/_p$ time units [14]. As a result, every $\delta_j = {}^T/_p$ and the time-aware GMRF becomes uniform with rescaled smoothing parameter $^\tau/_p$. Therefore, the conclusions of Fig. 3 hold exactly for the Skygrid, provided the horizontal axis is scaled by p. This setup reduces the rate of decay but the Ω curves still caution strongly against using skylines with $p \approx m$. Unfortunately, as its default formulation sets p to 1 less than the number of sampled taxa (or lineages) [14], the Skygrid is also be vulnerable to prior over-reliance.

The popular skyline-based coalescent inference methods therefore all tend to over-smooth, resulting in population size estimates that can be overconfident or misleading. This issue can be even more severe than Fig. 3 suggests since in current practice p is often close to m and non-robust designs are generally employed. Further, skylines are only statistically identifiable if every segment has at least 1 coalescent event [17], [33]. Consequently, if $p > m$ is set, smoothing priors can even mask identifiability problems. We recommend that $\frac{m}{p} \geq \kappa > 1$ must be guaranteed and in the next section derive a model rejection guideline for finding κ and diagnosing prior over-reliance.

E. Prior Informed Model Rejection

We previously demonstrated how commonly-used smoothing priors can dominate the posterior estimate precision when coalescent inference involves complex, highly parametrised (large-p) skyline models. Since data is more influential than the prior when $\Omega^2 > {}^1/_2$, we can use this threshold to define a simple p-rejection policy to guard against prior over-reliance. Assume that the $\mathcal{J}$ matrix resulting from our prior of interest is symmetric and positive definite. This holds for the GMRF and SMP. The standard arithmetic-geometric mean inequality, $\det[\mathcal{J}] \leq ({}^1/_p \, \text{tr}\, [\mathcal{J}])^p$, then applies with tr denoting the matrix trace. Since $\text{tr}\,[\mathcal{J}] = m + \text{tr}\,[\mathcal{P}]$ we can expand this inequality and substitute in Eq. (4) to get $\Omega^2 \geq ({}^1/_p \, (m + \text{tr}\,[\mathcal{P}]))^{-p} \prod_{j=1}^p m_j$.

Since this inequality applies to all $\{m_j\}$, we can maximise its right hand side to get a tighter lower bound on Ω^2. This bound, termed ω^2, is achieved at the robust design $m_j = {}^m/_p$ and is given by

$$\omega^2 = \left(\frac{m}{m + \text{tr}\,[\mathcal{P}]}\right)^p \implies p^* = \arg\max_{p \geq 1} \omega^2 \geq b. \tag{14}$$

We define $b \geq {}^1/_2$ as a conservative model rejection criteria with $\omega^2 \geq b$ implying that $\Omega^2 \geq b$. If p^* is the largest p satisfying these inequalities (see Eq. (14), arg indicates argument), then any skyline with more than p^* segments is likely to be overly-dependent on the prior and should be rejected under the current tree data.

Alternatively, we recommend that skylines using a smoothing prior (with matrix $\mathcal{P}$) should have at least $\kappa = {}^m/_{p^*}$ events per segment to avoid prior reliance. The $p \geq 1$ condition in Eq. (14) ensures skyline identifiability [17] and generally $p^* > 2$ (i.e. $\kappa > 1$). The dependence of ω^2 on $\mathrm{tr}[\boldsymbol{\mathcal{P}}]$ means that additions to the diagonals of $\boldsymbol{\mathcal{P}}$ necessarily increase the precision contribution from the prior. This insight supports our previous analysis, which used τ from the uniform GMRF to bound the performance of the SMP and time-aware GMRF. In the Appendix (see Eq. (7)) we derive analogous rejection bounds based on the excess mutual information, $\Delta\mathbb{I}$ (see Eq. (7)). There we find that p acts like an information-theoretic bandwidth, controlling the prior-contributed mutual information.

Eq. (14) can be computed and is valid for any smoothing prior of interest. In the case of the uniform GMRF where $\mathrm{tr}\,[\boldsymbol{\mathcal{P}}] = 2\tau(p-1)$, we get $\omega^2 = \left(\frac{m}{m+2\tau(p-1)}\right)^p$. Note that $\omega^2 = 1$ here whenever $p = 1$ or $\tau = 0$, as expected (i.e there is no smoothing at these values). In Fig. C.1 of the Appendix, we confirm that ω^2 is a good lower bound of Ω^2. We enumerate ω^2 across τ and p, for an observed tree with $m = 100$, to get Fig. 4, which recommends using no more than $p^* = 19$ segments ($\kappa \approx 5.3$). In Fig. C.2 we plot p^* curves for various m and τ, defining boundaries beyond which skyline estimates will be overly-dependent on the GMRF.

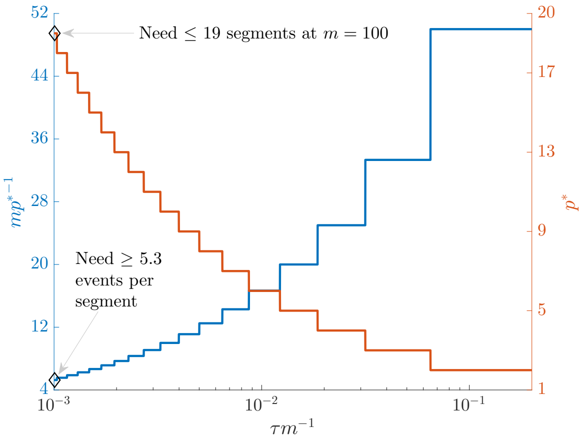

Fig. 4: **Bounding skyline complexity using the prior-data tradeoff.** For the GMRF with uniform smoothing, we show how the maximum number of recommended skyline segments, p^* (red), decreases with prior contribution (level of smoothing i.e. increasing τ/m). Hence the minimum recommended number of coalescent events per segment, m/p^* (blue), rises. Here we use the $\omega^2 \geq b = 1/2$ boundary. At larger b the p^* at a given τ/m decreases. The p^* measure provides a model rejection tool, suggesting that models with $p > p^*$ should not be used, as they would risk being overly informed by the prior.

In the Appendix we further analyse Eq. (14) for the uniform GMRF to discover that Ω^2 is bounded by curves with exponents linear in τ and quadratic in p (see Eq. (C.2)). This explains how the influence of smoothing increases with skyline complexity and yields a simple transformation $\tau \rightarrow \tau/2p(p-1)$, which can negate prior over-reliance. For comparison, the *Skyride* implements $\tau \rightarrow \tau/p$. The marked improvement, relative to Fig. 3, is striking in Fig. B.3. Other revealing prior-specific insights can be obtained from Eq. (14), reaffirming its importance as a model rejection statistic.

Our model rejection tool of Eq. (14) can serve as a useful diagnostic for skyline over-parametrisation, and as a precaution against prior over-reliance. However, we do not propose p^* as the sole measure of optimal skyline complexity. Our reason is that it does not guarantee any absolute estimate precision. Furthermore, it only focuses on the information from the prior relative to the data, and when the prior contribution is negligible, p^* can be unbounded. Choosing an optimal p is an open problem that is still under active study [16].

F. Case Study: Egyptian HCV

We validate the practical utility of Ω, as a diagnostic of prior over-dependence, on the well-studied Egyptian HCV-4 dataset, which consists of 63 sampled sequences [7]. Previous analyses reconstructed a demographic trend involving periods of constant population size separated by a phase of exponential growth from this dataset. We used the software MASTER [34] to simulate 100 coalescent trees, each with $m + 1 = n = 63$ tips, according to this demographic function [7]. We then inferred $\log \boldsymbol{N}$ from every tree using skyline models with time-aware GMRF smoothing priors, as in [10], [14].

We varied the relative contributions of the data and GMRF to the posterior log-population estimates by changing either the skyline dimension, p, or the smoothing parameter τ. Since m was fixed and robust designs applied, varying m_j effectively changes p. We analysed every tree over all combinations of $m_j \in \{1, 4, 8\}$ and a range of τ between 10^{-5} and 1. For comparison, we also generated purely data-informed estimates of $\log \boldsymbol{N}$, across the same range of m_j group sizes, by replacing the subjective GMRF with a uniform, objective prior. Each tree was additionally analysed with a time-aware *Skyride*, which means that $m_j = 1$ and τ is estimated from the data.

For every simulated tree we computed Ω (via Eq. (4)) and two ancillary statistics based on the 95% highest posterior density (HPD) intervals of the $\log \boldsymbol{N}$ estimates. These are the median HPD ratio $\mathrm{q}_{0.5}$ and the relative HPD product $\mathbb{H}_{\tau,m}$, which are formulated as:

$$\mathrm{q}_{0.5} = \mathrm{med}_j \left\{ \mathbb{H}^j_{\tau,m} := \frac{H^j_{\tau,m}}{H^j_m} \right\} \text{ and } \mathbb{H}_{\tau,m} = \textstyle\prod_{j=1}^m \mathbb{H}^j_{\tau,m},$$

with med as the median value of a set. Here $H^j_{\tau,m}$ is the 95% HPD interval of $\log N_j$ under a GMRF with smoothing parameter τ and H^j_m is the equivalent HPD with the uniform prior.

The 95% HPD interval is closely connected to the inverse of the Fisher information matrices that define Ω and, further, describes the most visually conspicuous representation of the uncertainty present in skyline-plot estimates. Comparing Ω to these ancillary statistics, which evaluate the median and total 95% uncertainty of a skyline plot, allows us to contextualise Ω against more relatable (though different) and obvious visualisations of posterior performance. We present these comparisons in Fig. D.1 of the Appendix.

There we find that all statistics monotonically decay with τ i.e. as the time-aware GMRF becomes more informative. The sharpness of this decay is highly sensitive to m_j. Larger m_j means that the more coalescent data is informing each estimated parameter and implies smaller p. The reduced decay with m_j supports our assertion that p acts as an exponent controlling prior over-reliance (see Fig. 3). The gentler decay of $\mathrm{q}_{0.5}$ (relative to Ω and $\mathbb{H}_{\tau,m}$), which largely does not account for p, confirms that we could be misled in our understanding of the impact of smoothing if we neglected skyline dimension.

In contrast Ω and $\mathbb{H}_{\tau,m}$, which both measure, in some sense, the relative volumes of uncertainty across the entire skyline-plot due to the data alone and the data and prior, fall more significantly and consistently. At $m_j = 1$ ($p = m$), which is the most common setting in the S/S methods, both statistics are markedly below $\frac{1}{2}$ and posterior estimates are expected to almost always be too dependent on the prior. This high-p behaviour to also indicative of model over-fitting. Our metric Ω therefore relates sensibly to visible proxies of uncertainty.

Having empirically confirmed Ω as a credible measure of relative uncertainty, we explore the behaviour of our proposed model rejection approximation, ω, which can be used to practically guard against prior over-reliance by defining a maximum viable p^*. We compute ω^2 for the HCV trees in Fig. 5 and observe that, as above, it decreases with both p and τ. Each dark red line in Fig. 5 depicts the posterior median τ estimated from each HCV tree. Practical analyses of this dataset with the S/S methods would use these τ values with $p = m$.

However, Fig. 5 shows that $p^* < m$ (i.e. $m_j > 1$) is necessary to achieve $\omega^2 \geq {}^1\!/_2$. This raises questions about the validity of using the S/S methods with default settings.

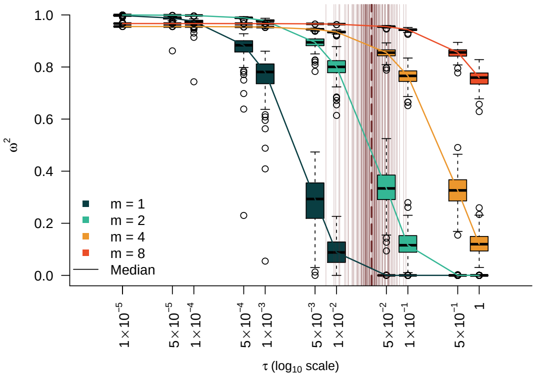

Fig. 5: **Model rejection statistics for the HCV dataset** The metric ω^2 is calculated for each tree under a time-aware GMRF at various (m_j, τ) settings. The box-plots summarise the resulting ω^2 over 100 HCV trees. The solid lines link the median values across boxes for a given m_j and hence p value ($m_j = {}^m\!/_p$). Vertical dark red lines highlight the τ that would be estimated by S/S approaches from these trees. We reject skyline settings achieving $\omega^2 < {}^1\!/_2$.

Fig. 5 confirms that the allowable p^* and hence minimum m_j inflates with τ. We demonstrate the qualitative difference between skyline-based estimates either side of the p^* criteria for a single HCV tree in Fig. 6. In panel A we present the *Skyride* estimate which uses $m_j = 1$ and implements a $p > p^*$, at one of its estimated τ values. We compare this to an equivalent skyline at $m_j = 4$, which achieves $p < p^*$ at this same τ (see Fig. 5) in panel B. In each panel, we overlay the corresponding skyline obtained with an objective uniform prior, to visualise the uncertainty available from the data alone.

At $m_j = 1$, the uniform prior produces a skyline that infers much more rapid demographic fluctuations through time than that estimated from applying the GMRF prior. Further, the 95% HPD intervals from the uniform prior (red) are much wider than those from the GMRF prior (blue) in the recent period ($t < 100$ years), highlighting the marked contribution of the GMRF prior to posterior estimate precision. While this smoother trajectory might look more reliable we argue that it is not justified by the data and that the skyline is over-fitting ($\kappa = {}^m\!/_p = 1$ data-points inform each parameter).

In contrast, at $m_j = 4$, both prior distributions yield similar skylines, implying that GMRF smoothing has not substantially inflated posterior estimate precision. At this setting we have less demographic fluctuations than $m_j = 1$ because four times more data is used for each parameter. For the period $t > 100$ years the 95% HPD interval of the GMRF estimate is much wider for $m_j = 1$ than $m_j = 4$ and some may argue that the latter case underestimates uncertainty. However, in this period there are few coalescent events. Maintaining $\kappa = 1$ when there is very little neighbouring information to regulate the log $\boldsymbol{N}$ estimates strongly suggests that the inflated uncertainty in panel A is instead symptomatic of over-parametrisation.

It contextualising these results it is important to note that skyline-plots provide harmonic mean and not point estimates of population size [11]. Consequently, we are really inferring a sequence of means from our coalescent data. Fig. 6 shows that over $t > 100$ years there are so few events that it is more sensible to estimate a single mean (panel B), which we are confident in across this period as opposed to several less certain means (panel A). In general, deciding on how to balance uncertainty with model complexity is non-trivial and, as shown in this example, caution is needed to avoid overconfident or misleading conclusions. We posit that Ω (and ω) can help clarify and formalise this decision-making.

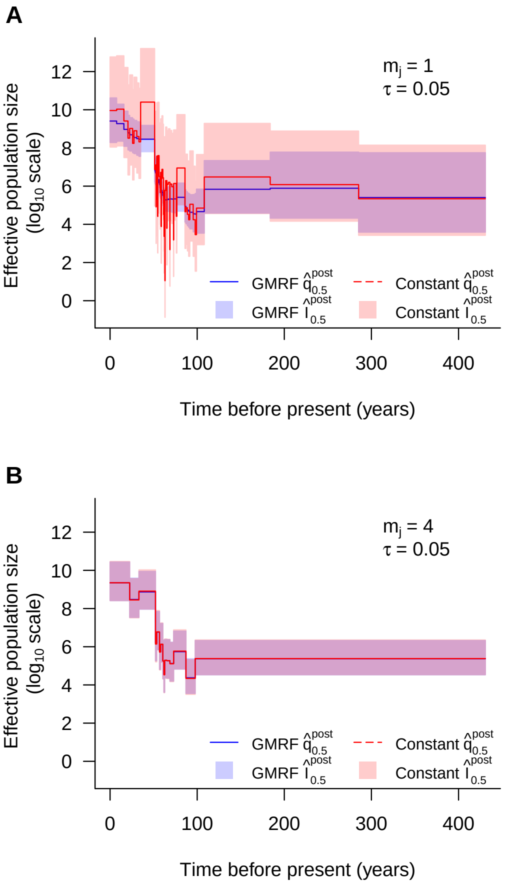

Fig. 6: **HCV demographic estimates under GMRF and uniform priors.** We analyse demographic estimates under time-aware GMRF priors (red) and objective uniform priors (blue) for a single tree generated from the Egyptian HCV dataset. In panel A we present *Skyride* estimates, which use $m_j = 1$ and $\tau = 0.05$ and implement a $p > p^*$ as computed from Fig. 5. In panel B we re-estimate population size at $m_j = 4$ which achieves $p < p^*$ as justified by our model rejection metric (see Eq. (14)). Solid lines are posterior medians while semi-transparent blocks are the 95% HPD intervals.

IV. Discussion

Popular approaches to coalescent inference, such as the BSP and S/S methods, all rely on combining a piecewise-constant population size likelihood function with prior assumptions that enforce continuity. This combination, which is meant to maximise descriptive flexibility without sacrificing the supposed smoothness exhibited by real population size curves over time, has led to many insights in phylodynamics [12]. However, it has also spawned several issues related to over-smoothing and lack of methodological transparency [10] [15]. In this work we attempted to address these issues by deriving metrics for diagnosing and clarifying the existing assumptions and obscurities present in current best practice.

By capitalising on (mutual) information theory and (Fisher) information geometry we formulated the novel coalescent information ratio, Ω, which is our main contribution. This ratio describes both the proportion of the asymptotic uncertainty around our posterior estimates that is due solely to the data and the additional mutual information that the prior introduces. It is simple to compute for piecewise-constant likelihoods and standard smoothing priors and has an exact interpretation as the ratio of real coalescent events to the sum of real and virtual (prior-contributed) ones in a Kingman coalescent model. As observed in our empirical Egyptian HCV analysis, Ω also correlates well with standard and visible indicators of estimate uncertainty such as relative HPDs.

Our ratio is therefore theoretically justified and practically useful. Using $\Omega^2 = 1/2$ as a threshold delimiting when the prior contributes as much information as the data, we examined widely-used SMP and GMRF smoothing priors and found that it is deceptively easy to become overly dependent on prior assumptions as skyline dimension, p, increases. This central result emerges from the drastic reduction in the number of coalescent events informing on any population size parameter as p rises. Per parameter, the BSP and *Skyride* use only a few or one event respectively [10], [13], while the *Skygrid* may have no events informing on some parameters [14].

These issues can be obscured by current Bayesian implementations, which would nonetheless produce seemingly reasonable population size estimates, at least visually, as illustrated in our Egyptian HCV case study. However, failing to diagnose these issues can be problematic, not just for prior over-dependence. Low coalescent event counts, for example, can lead to poor statistical identifiability [25] which might manifest in spurious MCMC mixing. Consequently, we proposed a practical p^* rejection criteria for ensuring that the data is the main source of inferential information.

This criteria bounds the maximum allowable skyline dimension for a given dataset (tree) size, only depends on computing the sum of the diagonals of the prior Fisher matrix and provides insight into how we can counter the dramatic impact of skyline complexity on prior over-reliance. When specialised to the GMRF, for example, it revealed that we could completely and surprisingly negate over-smoothing by scaling the precision parameter τ with a quadratic of p.

Moreover, this criteria shows that only by increasing the information available from the sampled phylogeny (i.e. the data) can we truly and sensibly allow for more complex piecewise-constant functions under a given prior. Recent methods, such as the *epoch sampling skyline plot* [33], which doubles the Fisher information extracted from a given phylogeny by accounting for the informativeness of sampling times, should therefore be able to support higher dimensional skylines than more standard approaches.

Thus, we have devised and validated a rigorous means of better understanding, diagnosing and preventing prior over-dependence. We hope that our statistic, which clarifies and quantifies the often inscrutable impact of the prior and data, will help researchers make more active and considered design decisions when adapting popular skyline-based techniques. While we recommend that data-driven conclusions are generally the most justifiable we note that in the context of skyline plots, this is open to interpretation.

Acknowledgments

We thank Louis du Plessis for his useful comments and insights on this project. This study was funded by the UK Medical Research Council (MRC) and the UK Department for International Development (DFID) under the MRC/DFID Concordat agreement and is also part of the EDCTP2 programme supported by the European Union (grant reference MR/R015600/1). The European Research Council and the European Commission Seventh Framework Programme (FP7/2007-2013)/European Research Council grant agreement 614725-PATHPHYLODYN is also acknowledged.

References

[1] J. Kingman, On the Genealogy of Large Populations, Journal of Applied Probability 19 (1982) 27–43.

[2] R. Griffiths, S. Tavare, Sampling Theory for Neutral Alleles in a Varying Environment, Philosophical Transactions Royal Society B 344 (1994) 403–10.

[3] P. Beerli, J. Felsenstein, Maximum Likelihood Estimation of Migration Rates and Effective Population Numbers in Two Populations using a Coalescent Approach, Genetics 152 (1999) 763–73.

[4] H. Li, R. Durbin, Inference of Human Population History from Individual Whole-genome Sequences, Nature 475 (7357) (2011) 493–6.

[5] B. Shapiro, A. Drummond, A. Rambaut, et al., Rise and Fall of the Beringian Steppe Bison, Science 306 (5701) (2004) 1561–1565.

[6] J. Wakeley, Coalescent Theory: An Introduction, Roberts and Company Publishers, 2008.

[7] O. Pybus, A. Drummond, T. Nakano, et al., The Epidemiology and Iatrogenic Transmission of Hepatitis C Virus in Egypt: A Bayesian Coalescent Approach, Molecular Biology and Evolution 20 (3) (2003) 381–7.

[8] A. Rodrigo, J. Felsenstein, Coalescent Approaches to HIV-1 Population, The Evolution of HIV, Johns Hopkins University Press, 1999.

[9] M. Kuhner, J. Yamato, J. Felsenstein, Maximum Likelihood Estimation of Population Growth Rates based on the Coalescent, Genetics 149 (1998) 429–34.

[10] V. Minin, E. Bloomquist, M. Suchard, Smooth Skyride through a Rough Skyline: Bayesian Coalescent-Based Inference of Population Dynamics, Molecular Biology and Evolution 25 (7) (2008) 1459–71.

[11] O. Pybus, A. Rambaut, P. Harvey, An Integrated Framework for the Inference of Viral Population History from Reconstructed Genealogies, Genetics 155 (2000) 1429–37.

[12] S. Ho, B. Shapiro, Skyline-plot Methods for Estimating Demographic History from Nucleotide Sequences, Molecular Ecology Resources 11 (2011) 423–34.

[13] A. Drummond, A. Rambaut, B. Shapiro, O. Pybus, Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences, Molecular Biology and Evolution 22 (5) (2005) 1185–92.

[14] M. Gill, P. Lemey, N. Faria, et al., Improving Bayesian Population Dynamics Inference: A Coalescent-Based Model for Multiple Loci, Molecular Biology and Evolution 30 (3) (2012) 713–24.

[15] J. Faulkner, A. Magee, B. Shapiro, et al., Horseshoe-based Bayesian Nonparametric Estimation of Effective Population Size Trajectories., Biometrics (2019) In Press.

[16] K. Parag, C. Donnelly, Adaptive Estimation for Epidemic Renewal and Phylogenetic Skyline Models, BioRxiv (2019) [Preprint].

[17] K. Parag, O. Pybus, Robust Design for Coalescent Model Inference, Systematic Biology 68 (5) (2019) 730–43.

[18] K. Parag, O. Pybus, Optimal Point Process Filtering and Estimation of the Coalescent Process, Journal of Theoretical Biology 421 (2017) 153–67.

[19] D. Snyder, M. Miller, Random Point Processes in Time and Space, 2nd Edition, Springer-Verlag, 1991.

[20] E. Lehmann, G. Casella, Theory of Point Estimation, 2nd Edition, Springer-Verlag, 1998.

[21] H. van Trees, Detection, Estimation, and Modulation Theory, Part I, John Wiley and Sons Inc, 1968.

[22] P. Tichavsky, C. Muravchik, A. Nehorai, Posterior Cramer-Rao Bounds for Discrete-Time Nonlinear Filtering, IEEE Transactions on Signal Processing 46 (5) (1998) 1386–95.

[23] W. Huang, K. Zhang, Information-Theoretic Bounds and Approximations in Neural Population Coding, Neural Computation 30 (4) (2018) 885–944.

[24] Z. Ben-Haim, Y. Eldar, A Lower Bound on the Bayesian MSE Based on the Optimal Bias Function, IEEE Transactions on Information Theory 55 (11) (2009) 5179–96.

[25] T. Rothenburg, Identification in Parametric Models, Econometrica 39 (3).

[26] T. Cover, J. Thomas, Elements of Information Theory Second Edition, John Wiley and Sons, 2006.

[27] N. Brunel, J. Nadal, Mutual Information, Fisher Information, and Population Coding, Neural Computation 10 (1998) 1731–57.

[28] E. Slate, Parameterizations for Natural Exponential Families with Quadratic Variance Functions, Journal of the American Statistical Association 89 (428) (1994) 1471–81.

[29] I. Ipsen, R. Rehman, Perturbation Bounds for Determinants and Characteristic Polynomials, SIAM J. Matrix Anal. Appl 30 (2) (2008) 762–76.

[30] D. Fink, A Compendium of Conjugate Priors, Tech. rep., Montana State University (1997).

[31] C. Robert, The Bayesian Choice, Springer Texts in Statistics, Springer Science + Business Media, 2007.

[32] J. Berger, J. Bernardo, D. Sun, Overall Objective Priors, Bayesian Analysis 10 (1) (2015) 189–221.

[33] K. Parag, L. du Plessis, O. Pybus, Jointly inferring the dynamics of population size and sampling intensity from molecular sequences, Molecular Biology and Evolution (In Press) (2020).

[34] T. Vaughan, A. Drummond, A Stochastic Simulator of Birth–Death Master Equations with Application to Phylodynamics, Molecular Biology and Evolution 30 (6) (2013) 1480–93.

APPENDICES

A. *Smoothing Prior Fisher Information Matrices*

Here we derive the prior-informed Fisher information matrices for the SMP and GMRF smoothing priors. We start by finding the log-population size transformed version of the SMP smoothing prior. We then calculate its Hessian to get $\boldsymbol{\mathcal{P}}$, and so obtain the general form of Eq. (10). The SMP is given in [13] as $f(\boldsymbol{N}) = {}^{1}/_{N_1} \prod_{j=2}^{m} {}^{1}/_{N_{j-1}}\, e^{N_j/N_{j-1}}$. We define $\boldsymbol{\eta} = \rho(\boldsymbol{N}) := \log \boldsymbol{N}$ so that its inverse $\rho^{-1}(\boldsymbol{\eta}) = e^{\boldsymbol{\eta}}$. These expressions are in vector form so $\boldsymbol{\eta} = [\eta_1, \, \dots, \, \eta_p] = [\log N_1, \, \dots, \, \log N_p]$. We want the transformed prior $g(\boldsymbol{\eta})$. Applying the multivariate change of variables formula gives $g(\boldsymbol{\eta}) = f(e^{\boldsymbol{\eta}})|\det\left[\Delta\rho^{-1}\right]|$, with $\Delta\rho^{-1} = [e^{\eta_1}, \, \dots, \, e^{\eta_p}]\, \mathrm{I}_p$ as the Jacobian of ρ^{-1}. This implies that $|\det\left[\Delta\rho^{-1}\right]| = e^{\sum_{j=1}^{p} \eta_j}$. Substituting and expanding gives the SMP log-prior:

$$\log g(\boldsymbol{\eta}) = \eta_p - \eta_1 + \sum_{j=2}^{p} -e^{\eta_j - \eta_{j-1}}. \tag{A.1}$$

We can then obtain $\boldsymbol{\mathcal{P}} = -\nabla \boldsymbol{G}$, with $\boldsymbol{G} = \log g(\boldsymbol{\eta})$. The diagonals of $\boldsymbol{\mathcal{P}}$ are: ${}^{\partial^2 \boldsymbol{G}}/_{\partial \eta_j^2} = -e^{\eta_j - \eta_{j-1}} - e^{\eta_{j+1} - \eta_j}$ for $2 \leq j \leq p-1$, ${}^{\partial^2 \boldsymbol{G}}/_{\partial \eta_1^2} = -e^{\eta_2 - \eta_1}$ and ${}^{\partial^2 \boldsymbol{G}}/_{\partial \eta_p^2} = -e^{\eta_p - \eta_{p-1}}$. The non-zero off-diagonal terms are: ${}^{\partial^2 \boldsymbol{G}}/_{\partial \eta_j \eta_{j+1}} = e^{\eta_{j+1} - \eta_j}$ and ${}^{\partial^2 \boldsymbol{G}}/_{\partial \eta_j \eta_{j-1}} = e^{\eta_j - \eta_{j-1}}$. The result is a symmetric tridiagonal matrix that has zero row and column sums. The $\boldsymbol{\mathcal{P}}$ matrix is then added to the Fisher information matrix $\boldsymbol{\mathcal{I}} = [m_1, \, \dots, \, m_p]\, \mathrm{I}_p$ (with m_j as the number of coalescent events informing on the j^{th} parameter), to get $\boldsymbol{\mathcal{J}}_{\text{SMP}}$.

We now compute $\boldsymbol{\mathcal{J}}_{\text{GMRF}}$, which is given in the main text as Eq. (11). For the GMRF $g(\boldsymbol{\eta}) = Z^{-1} \tau^{\frac{p-2}{2}} e^{-\frac{\tau}{2} \sum_{j=1}^{p-1} \delta_j^{-1} (\eta_{j+1} - \eta_j)^2}$ [10] and so $\boldsymbol{G} = -\log Z + \frac{m-2}{2} \log \tau - \frac{\tau}{2} \sum_{j=1}^{p-1} \frac{(\eta_{j+1} - \eta_j)^2}{\delta_j}$. Taking second derivatives we get diagonal terms of the Hessian, $\nabla \boldsymbol{G}$, as: ${}^{\partial^2 \boldsymbol{G}}/_{\partial \eta_j^2} = -\tau \left({}^{1}/_{\delta_j} + {}^{1}/_{\delta_{j-1}}\right)$ for $2 \leq j \leq p-1$, ${}^{\partial^2 \boldsymbol{G}}/_{\partial \eta_1^2} = -{}^{\tau}/_{\delta_1}$ and ${}^{\partial^2 \boldsymbol{G}}/_{\partial \eta_p^2} = -{}^{\tau}/_{\delta_{p-1}}$. The non-zero off diagonal terms are: ${}^{\partial^2 \boldsymbol{G}}/_{\partial \eta_j \eta_{j+1}} = {}^{\tau}/_{\delta_j}$ and ${}^{\partial^2 \boldsymbol{G}}/_{\partial \eta_j \eta_{j-1}} = {}^{\tau}/_{\delta_{j-1}}$. The GMRF also gives a symmetric tridiagonal $\boldsymbol{\mathcal{P}}$ with row and column sums of zero. Adding $-\nabla \boldsymbol{G}$ to the diagonal $\boldsymbol{\mathcal{I}}$ matrix yields $\boldsymbol{\mathcal{J}}_{\text{GMRF}}$.

B. *Further Smoothing Results*

We previously asserted that the Ω computed at the robust point of $m_j = {}^{m}/_{p}$ [17] generally upper bounds the achievable Ω values at other m_j settings. Here we provide evidence for this assertion. While strictly $\arg\max_{\{m_j\}} \Omega \neq {}^{m}/_{p}$ (except for $p = 2$), we numerically find that $\max_{\{m_j\}} \Omega \approx \Omega|_{\{m_j = \frac{m}{p}\}}$. We show this for the GMRF under uniform smoothing in Fig. B.1. This makes sense as while (for fixed smoothing parameters) $\arg\max_{\{m_j\}} \det[\boldsymbol{\mathcal{I}}] = {}^{m}/_{p}$ and $\arg\max_{\{m_j\}} \det[\boldsymbol{\mathcal{J}}] = {}^{m}/_{p}$, there is no reason to believe that this also maximises their ratio. The sawtooth Ω curves in Fig. B.1 reflect changes in the other $\{m_j\}$ values, given a fixed m_1.

Hence we used the robust design point in our calculation of the Ω^2 curves for the GMRF in Fig. 3. The corresponding additional mutual information ($\Delta\mathbb{I}$) curves for this case are provided in Fig. B.2. These show how larger values of the smoothness parameter, τ, directly lead to increases in the relative mutual information contribution from the prior. Observe that $\Delta\mathbb{I}$ is highly sensitive to the skyline complexity, p, thus clarifying how estimates from over-parametrised skyline plots can be dominated by prior information.

Interestingly, we can largely negate the impact of skyline complexity by making τ a function of p. In Section III-D we explained how the Skyride implicitly implements the scaling $\tau \rightarrow {}^{\tau}/_{p}$. While this reduces some of the effect of p, it still leads to decaying curves that can, for a given τ, be deceptively dependent on smoothing. Here we propose the key transformation $\tau \rightarrow {}^{\tau}/_{2p(p-1)}$, as a means of reducing our smoothing in line with our skyline complexity. This transformation was inspired by the dependence of a lower bound on Ω^2, which we derived in Eq. (C.2) of Section C. Its impact on the spread of curves from Fig. 3 is given in Fig. B.3.

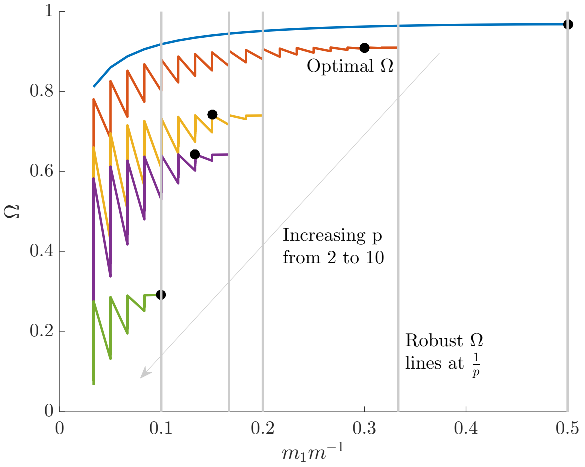

Fig. B.1: **Robust and Ω optimal designs.** For the GMRF smoothing prior with $\delta_i = 1$ for all i and $\tau = 1$, we show that the optimal Ω design point is not always the same as the robust design point, at which $\frac{m_1}{m} = \frac{1}{p}$. The coloured Ω curves are (along the dashed arrow) for $p = [2, \, 3, \, 5, \, 6, \, 10]$ at $m = 60$, and computed across all partitions for any given m_1 (hence the zig-zagged form). The grey vertical lines mark the robust point for each Ω curve, and the black circles give the optimal Ω points. While these lines and circles do not always match, both generally feature approximately the same Ω values. We found this to be the case across several m and τ values.

C. *Further Model Selection Bounds*

In Section III-E we derived lower bounds on Ω^2, which led to the model rejection parameter, p^*. Here we extend and support those results. In Fig. C.1 we first show that the bound of Eq. (14) is a good measure of the true Ω^2 value, for a skyline with uniform GMRF smoothing. We used this bound to define a maximum p, p^*, above which the skyline would be over-parametrised and susceptible to prior induced overconfidence. We explore p^* over τ and m for this GMRF in Fig. C.2 and observe that p^* becomes more restrictive with fewer observed data (coalescent events) or increased smoothing. This supports Ω as a useful measure of prior-data contribution.

Lower bounds on Ω^2 imply upper bounds on the excess mutual information, $\Delta\mathbb{I}$ (see Eq. (7)). We manipulate Eq. (14) (under a robust design) to obtain the first inequality in Eq. (C.1), with $q = {}^{\mathrm{tr}[\boldsymbol{\mathcal{P}}]}/_{m}$.

$$\Delta\mathbb{I} \leq \frac{1}{2} p \log(1 + q) \leq \frac{1}{2} pq \tag{C.1}$$

This expression reveals that p is akin to a signal bandwidth (by comparison with standard Shannon-Hartley theory [26]) and is therefore a key controlling factor in defining how much additional information the prior will introduce. This supports our proposed p^* criteria.

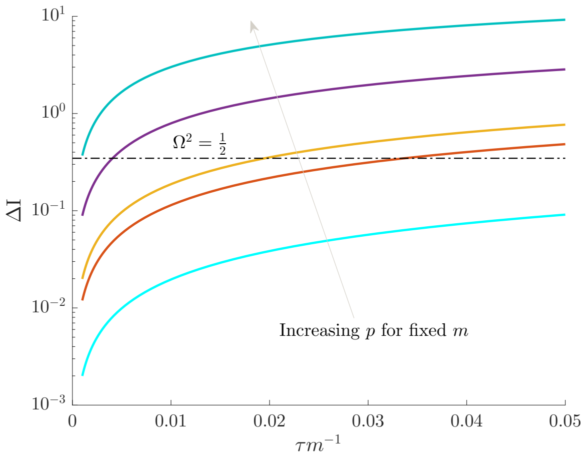

Fig. B.2: **Prior mutual information increases with skyline complexity.** For the uniform GMRF, we show that under fixed smoothing (and hence $^{\tau}/_{m}$), the additional mutual information introduced by the prior, $\Delta\mathbb{I} = \mathbb{E}_0[-\log\Omega]$, significantly increases with the complexity, p, of our skyline. The coloured Ω curves are (along the grey arrow) for $p = [2, 4, 5, 10, 20]$ at $m = 100$ with $m_j = {}^{m}/_{p}$ (robust design point [17]). The dashed $\Omega^2 = {}^1/_2$ threshold is also given for comparison. Clearly, the more skyline segments we have for a given tree, the more likely we are being overly informed by our prior.

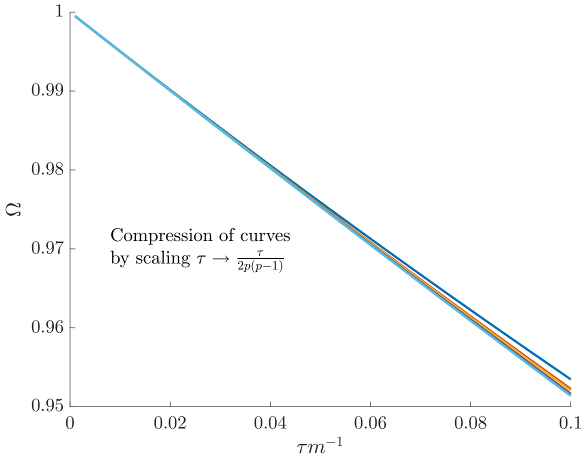

Fig. B.3: **Negating the impact of skyline complexity.** We show how an appropriate quadratic scaling of the GMRF precision parameter, τ, can remove the complexity (p) induced smoothing contribution portrayed in Fig. 3 of the main text. This scaling significantly compresses the coloured Ω curves shown, which are for $p = [2, 4, 5, 10, 20]$ at $m = 100$ with $m_j = {}^{m}/_{p}$ (robust design point [17]). The resulting Ω^2 values are now all comfortably above the ${}^1/_2$ threshold.

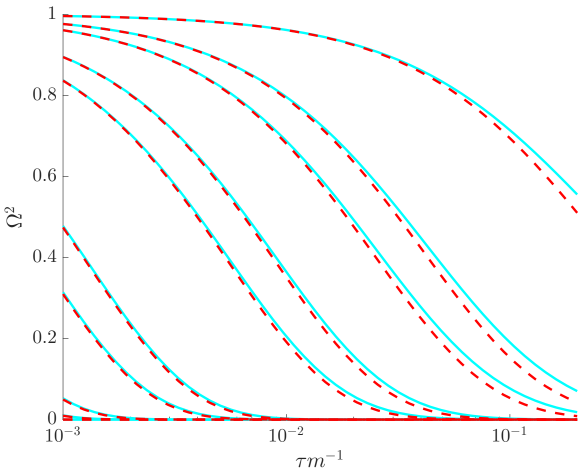

Fig. C.1: **Lower bounds on Ω^2.** For the GMRF smoothing prior with $\delta_j = 1$ for all j and $m = 200$, we compare the lower bound on Ω^2 (red, dashed, see Eq. (14)) with the actual value of Ω^2 (cyan) at the robust design point of $m_j = {}^{m}/_{p}$. We examine all integer p values that are factors of m, and find that qualitatively similar comparisons hold for different τ and m settings. In general the lower bound (denoted ω^2 in the main text) is a good approximation to Ω^2.

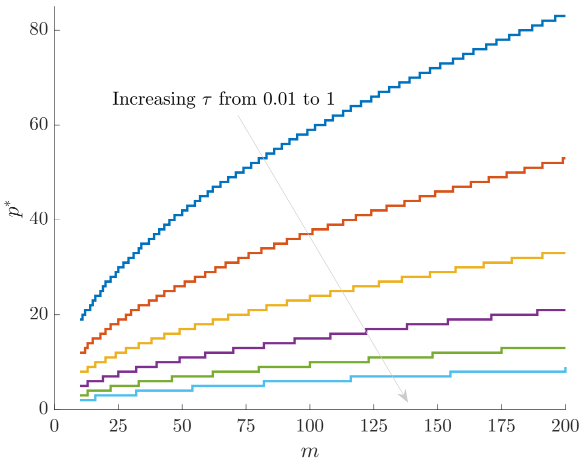

Fig. C.2: **Maximum p model selection boundary.** For the GMRF smoothing prior with $\delta_j = 1$ for all j and at the robust point $m_j = {}^{m}/_{p}$, we compute the maximum allowed number of skyline segments, p^*, such that $\Omega^2 \geq {}^1/_2$. These curves increase with m and decrease with τ, indicating how the prior-data contribution can be used to define model rejection regions. Skylines with $p > p^*$ would be overly informed by the prior and hence should not be used.

Under the log $\boldsymbol{N}$ parametrisation, $\boldsymbol{\mathcal{I}}$ and $\boldsymbol{\mathcal{J}}$ are symmetric, positive definite matrices. For such matrices we can apply a theorem from [23], which states that $\Delta\mathbb{I} \leq {}^{\zeta}/_{2}$, with $\zeta = \text{tr}[\boldsymbol{\mathcal{I}}^{-\frac{1}{2}}\boldsymbol{\mathcal{P}}\boldsymbol{\mathcal{I}}^{-\frac{1}{2}}]$. At the robust point, we get $\zeta = \text{tr}[\boldsymbol{\mathcal{I}}^{-1}\boldsymbol{\mathcal{P}}]$, which leads to the second inequality in Eq. (C.1). Thus, our bound is tighter than that in [23], and hence useful for broader, future mathematical analyses of $\Delta\mathbb{I}$.

We can also use the bound of [23] to derive alternate (but slacker) lower bounds on Ω^2. This gives the first inequality in Eq. (C.2). Applying this to the uniform GMRF gives the second inequality.

$$\Omega^2 \geq e^{-pq} \implies \Omega^2 \geq e^{-\frac{2}{m}p(p-1)\tau} \tag{C.2}$$

Interestingly, Eq. (C.2) shows that the dependence of Ω^2 on the precision parameter τ is at most only linear, while the dependence on complexity p can be quadratic. This provides further theoretical backing for the use of p^* to reject models and emphasises how smoothing can play a deceptively prominent role in the resulting estimate precision produced under complex skyline plots.

D. Ancillary Uncertainty Statistics

In Section III-F we defined two 95% HPD based ancillary statistics for characterising the visual uncertainty present in a skyline-plot demographic estimate. In Fig. D.1 we plot these statistics and Ω^2 for various τ and m_j values under a time-aware GMRF. We discuss the implications of Fig. D.1 in the main text.

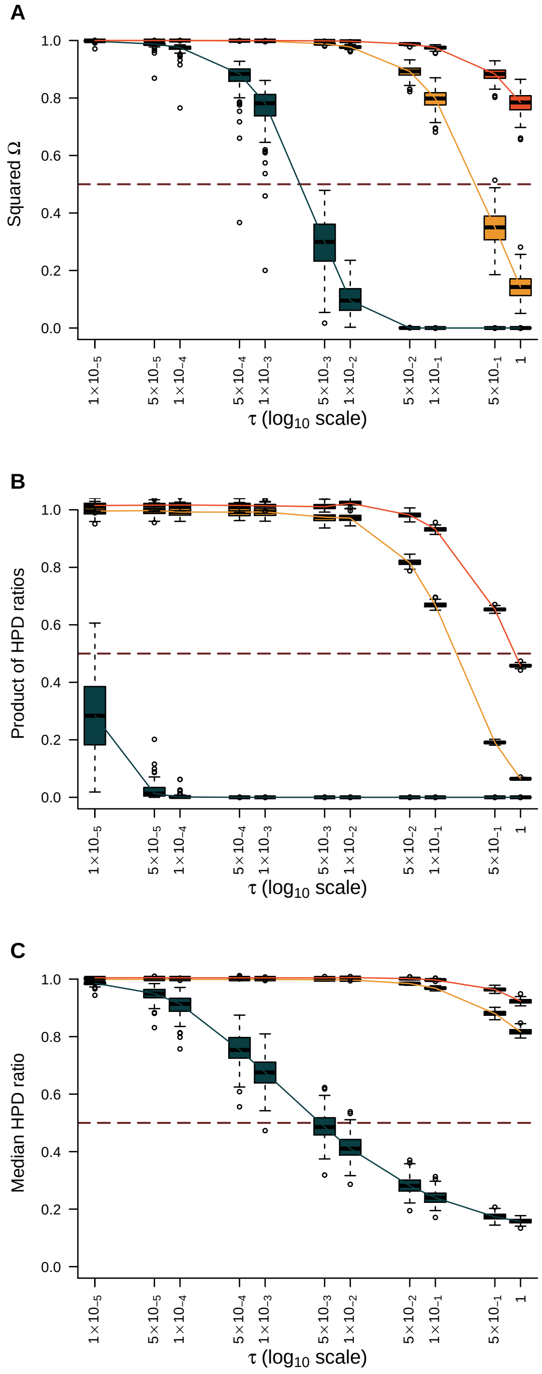

Fig. D.1: **Trends in HPD-based statistics and Ω^2 under various time-aware GMRF settings.** The Ω^2 (panel A), median HPD ratio of $\log N_j$ (panel B) and HPD product (panel C) statistics are computed across $\log N_j$ over various combinations of m_j and τ. Box-plots summarise our results over 100 observed coalescent trees consistent with the Egyptian HCV dataset. Analyses with $m_j = 1$ are in dark green, $m_j = 4$ yellow and $m_j = 8$ orange. The solid lines link the median values across boxes for a given m_j value. The dashed line is positioned at the threshold $\Omega^2 = {}^1/_2$.