Systematic Biology (2021), **0**, 0, *pp.* 1–44
doi:10.1093/sysbio/main˙paper

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*

**Correspondence to be sent to: k.parag@imperial.ac.uk*

Abstract

In Bayesian phylogenetics, the coalescent process provides an informative framework for inferring changes in the effective size of a population from a phylogeny (or tree) of sequences sampled from that population. Popular coalescent inference approaches such as the *Bayesian Skyline Plot*, *Skyride* and *Skygrid* all model these population size changes with a discontinuous, piecewise-constant function but then apply a smoothing prior to ensure that their 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 data i.e., the tree. 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 inference, even under robust experimental designs. We propose Ω as a useful tool for detecting when effective population size estimates are overly reliant on prior assumptions and for improving quantification of the uncertainty in those estimates.

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

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 in (Kingman, 1982) under the assumption of a constant sized population, the coalescent process has since been extended to account for temporal variation in the population size (Griffiths and Tavare, 1994), structured demographics (Beerli and Felsenstein, 1999) and multi-locus sampling (Li and Durbin, 2011). 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 into infectious disease epidemiology, population genetics and molecular ecology (Shapiro *et al.*, 2004; Wakeley, 2008; Pybus *et al.*, 2003). Here we focus on coalescent processes that describe the genealogies of serially-sampled individuals from populations with deterministically varying size. These are widely applied to study the phylodynamics of infectious diseases (Griffiths and Tavare, 1994; Rodrigo and Felsenstein, 1999).

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 (Kuhner *et al.*, 1998; Pybus *et al.*, 2003). While these formulations required only a few variables and provided interpretable estimates, selecting the most appropriate parametric description could be challenging and risk underfitting complex trends (Minin *et al.*, 2008). This motivated the introduction of the *classic skyline plot* (Pybus *et al.*, 2000), 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 parametric restrictions. However, this flexibility came at the cost of increased estimation noise and potential overfitting of changes in effective population size (Ho and Shapiro, 2011).

Efforts to redress these issues within a piecewise-constant framework subsequently spawned a family of skyline plot-based methods (Ho and Shapiro, 2011). Among these, the most popular and commonly-used are the *Bayesian Skyline Plot* (BSP) (Drummond *et al.*, 2005), the *Skyride* (Minin *et al.*, 2008) and the *Skygrid* (Gill *et al.*, 2013) approaches. 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 (Parag and Donnelly, 2020; Strimmer and Pybus, 2001).

The BSP limited overfitting by (i) predefining fewer piecewise demographic changes than coalescent events and (ii) smoothing 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 (Minin *et al.*, 2008) for making strong smoothing and change-point assumptions and stimulated the 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 Skyride and Skygrid methods aimed to better trade off prior influence with noise reduction, and while somewhat effective, are still imperfect because they can fail to recover genuinely abrupt demographic changes such as bottlenecks (Faulkner *et al.*, 2019).

As a result, studies continue to explore and address the non-trivial problem of optimising this tradeoff, either by searching for less-restrictive and more adaptive priors (Faulkner *et al.*, 2019) or by deriving new data-driven skyline change-point grouping strategies (Parag and Donnelly, 2020). 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 reliable 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 develop and present a novel information theoretic statistic, Ω, to formally quantify and disaggregate the contributions of both priors and data on the uncertainty around the posterior demographic estimates of popular skyline-based coalescent methods. Using Ω we show how widely-used smoothing priors can result in overconfident population size inferences (i.e. estimates with unjustifiably small credible intervals) and provide practical guidelines against such circumstances. We illustrate the utility of this approach on well-characterised datasets describing the population size of HCV in Egypt (Pybus *et al.*, 2003) and ancient Beringian steppe Bison (Shapiro *et al.*, 2004).

To our knowledge, Ω, which in theory can be adapted to any prior-data comparison problem, is new not only to the field of phylogenetics but also across statistics and data science. While inference that is strongly driven by prior assumptions can be beneficial, for example when a prior encodes expert knowledge or salient dynamics, having a measure of the relative information introduced by data and prior distributions can improve the reproducibility and interpretability of analyses. Our statistic will help to 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.

Materials and Methods

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 (Kingman, 1982). 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 $n \geqslant 2$ samples are taken from the target population at $n_s \geqslant 1$ distinct sampling times, which are independent of and uninformative about population size changes (Drummond *et al.*, 2005). We do not specify the sample generating process as it does not affect our analysis by this independence assumption (Parag and Pybus, 2019). We let c_i be the time of the i^{th} coalescent event in $\mathcal{T}$ with $1 \leqslant i \leqslant n-1$ 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 $t \geqslant 0$ 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 $\binom{l_t}{2}N(t)^{-1}$ (Kingman, 1982). While $N(t)$ can be described using appropriate parametric formulations (Parag and Pybus, 2017), it is more common to represent $N(t)$ by some tractable p-dimensional piecewise-constant approximation (Ho and Shapiro, 2011). Thus, we can write $N(t) := \sum_{j=1}^{p} N_j 1(\epsilon_{j-1} \leqslant t < \epsilon_j)$, with $p \geqslant 1$ 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 $\epsilon_p \geqslant T$ and $1(x)$ is an indicator function. The rate of producing new coalescent events is then $\sum_{j=1}^{p} N_j^{-1}\binom{l_t}{2}1(\epsilon_{j-1} \leqslant t < \epsilon_j)$. Kingman's coalescent model is obtained by setting $p = 1$ (constant population of N_1).

When reconstructing the population size history of infectious diseases, it is often of interest to infer $N(t)$ from $\mathcal{T}$ (Ho and Shapiro, 2011), which forms our coalescent data generating process. If $\boldsymbol{N} = [N_1, ..., N_p]$ denotes the vector of demographic parameters to be estimated then the coalescent data log-likelihood $\ell(\boldsymbol{N}) := \log \mathrm{P}(\mathcal{T} \mid \boldsymbol{N})$ can be obtained from (Parag and Pybus, 2019) (Snyder and Miller, 1991) as

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

with A_j and B_j as constants that depend 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 serially-sampled skyline log-likelihood in (Drummond *et al.*, 2005), except that we do not restrict $N(t)$ to change only at coalescent event times.

In Bayesian phylogenetic inference, skyline-based methods such as the BSP, Skyride and Skygrid 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, from Bayes law, which depends on both the prior and coalescent data-likelihood 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, $\boldsymbol{D}$, are 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 expression $\mathrm{P}(\mathcal{T}, \boldsymbol{\theta}, \boldsymbol{N} \mid \boldsymbol{D}) \propto \mathrm{P}(\boldsymbol{D} \mid \mathcal{T}, \boldsymbol{\theta})\mathrm{P}(\mathcal{T} \mid \boldsymbol{N})\mathrm{P}(\boldsymbol{N})\mathrm{P}(\boldsymbol{\theta})$, which involves inferring both the tree and population size (Drummond *et al.*, 2002).

While we do not consider this extension here we note that results presented here are still applicable and relevant. This follows because the output of the more complex Bayesian analysis above (i.e. when sequence data $\boldsymbol{D}$ are used directly) is a posterior distribution over tree space. We can sample from this posterior and treat each sampled tree effectively as a fixed tree. Consequently, we expect any summary statistic that we derive here, under the assumption of a fixed-tree will be usable in studies that incorporate genealogical uncertainty by computing the distribution of that statistic over this covering set of sampled posterior trees.

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 $\boldsymbol{\psi} = [\psi_1, \ldots, \psi_p]$, 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. the Skyride and Skygrid) for ϕ. Eq. (1) and Eq. (2) are then reformulated with $\ell(\boldsymbol{\psi}) = \log \mathrm{P}(\mathcal{T} \mid \boldsymbol{\psi})$ as the coalescent data 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, $\boldsymbol{\mathcal{I}(\psi)}$, is a $p \times p$ matrix with $(i, j)^{\text{th}}$ element $\boldsymbol{\mathcal{I}(\psi)}_{ij} := -\mathbb{E}_{\mathcal{T}}\left[\nabla_{ij}\ell(\boldsymbol{\psi})\right]$ (Lehmann and Casella, 1998). The expectation is taken over the coalescent tree branches and $\nabla_{ij} := {}^{\partial^2}\!/_{\partial\psi_i\partial\psi_j}$. As observed in (Parag and Pybus, 2019), $\boldsymbol{\mathcal{I}(\psi)}$ 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 (Lehmann and Casella, 1998). The BSP, Skyride and Skygrid parametrisations all yield $\boldsymbol{\mathcal{I}(N)} = [m_1 N_1^{-2}, \ldots, m_p N_p^{-2}]\, \mathrm{I}_p$ and $\boldsymbol{\mathcal{I}}(\log \boldsymbol{N}) = [m_1, \ldots, m_p]\, \mathrm{I}_p$, with I_p as a $p \times p$ identity matrix (Parag and Pybus, 2019). These matrices provide several useful insights that we will exploit in later sections. First, $\boldsymbol{\mathcal{I}(\psi)}$ 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 (Parag and Pybus, 2017). 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 (Parag and Pybus, 2019).

The prior distribution, $\mathrm{P}\left(\boldsymbol{\psi}\right)$, 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), $\tilde{\boldsymbol{\psi}} = \arg\max_{\boldsymbol{\psi}}\{\log \mathrm{P}(\mathcal{T} \mid \boldsymbol{\psi}) + \log \mathrm{P}(\boldsymbol{\psi})\}$ (van Trees, 1968). 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 $\boldsymbol{\mathcal{J}}(\boldsymbol{\psi})_{ij} = -\mathbb{E}_{\mathcal{T}}\left[\nabla_{ij} \log \mathrm{P}(\boldsymbol{\psi} \mid \mathcal{T})\right]$ (Tichavsky *et al.*, 1998; Huang and Zhang, 2018). 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 (Huang and Zhang, 2018) 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 (Lehmann and Casella, 1998; Parag and Pybus, 2019). 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 (Tichavsky *et al.*, 1998).

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 $\mathbb{E}_{\mathcal{T}}\left[(\bar{\boldsymbol{\psi}} - \boldsymbol{\psi})(\bar{\boldsymbol{\psi}} - \boldsymbol{\psi})^{\intercal} \mid \boldsymbol{\psi}\right] = \mathrm{var}(\bar{\boldsymbol{\psi}} \mid \boldsymbol{\psi}) \geqslant \boldsymbol{\mathcal{I}}(\boldsymbol{\psi})^{-1}$ 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 (van Trees, 1968). If $\bar{\boldsymbol{\psi}}$ is any estimator of $\boldsymbol{\psi}$ then the BCRB states that $\mathbb{E}_0\left[\mathbb{E}_{\mathcal{T}}\left[(\bar{\boldsymbol{\psi}} - \boldsymbol{\psi})(\bar{\boldsymbol{\psi}} - \boldsymbol{\psi})^{\intercal} \mid \boldsymbol{\psi}\right]\right] \geqslant \boldsymbol{\mathcal{J}}_{\mathbf{0}}^{\mathbf{-1}}$. This bound is not dependent on $\boldsymbol{\psi}$ due to the extra expectation over the prior (Tichavsky *et al.*, 1998).

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 (Parag and Pybus, 2019). 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 that treats $\boldsymbol{\psi}$ as a random parameter. The expectation over the prior connects the two formalisms (Ben-Haim and Eldar, 2009). 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.

Results

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 coalescent data to the posterior estimates obtained as solutions to Bayesian skyline inference problems (see Materials and Methods). 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 coalescent 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 (Parag and Pybus, 2019) this ellipsoid was found central to understanding the statistical properties of skyline-based estimates.

The volume of this ellipsoid is $V_1 = C \det\left[\boldsymbol{\mathcal{I}}(\boldsymbol{\psi})\right]^{-\frac{1}{2}}$, with C as a p-dependent constant. Decreasing V_1 increases the best estimate precision attainable from the data $\mathcal{T}$ (Lehmann and Casella, 1998). 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 $V_2 = C \det\left[\boldsymbol{\mathcal{J}}(\boldsymbol{\psi})\right]^{-\frac{1}{2}}$ that measures the uncertainty around the MAP estimate $\tilde{\boldsymbol{\psi}}$ (Tichavsky *et al.*, 1998). This volume includes the effect of both prior and data on estimate precision. Accordingly, we propose the ratio

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

as a novel and natural statistic for dissecting the relative impact of the data and prior on posterior estimate precision.

From Eq. (4), we observe that $0 \leqslant \Omega \leqslant 1$ with $\Omega = 1$ signifying that the information from our prior distribution is negligible in comparison to that from the data and $\Omega = 0$ indicating the converse. Importantly, we find

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

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 (Rothenburg, 1971)). 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 $\boldsymbol{\psi}$ (Cover and Thomas, 2006). This is distinct but related to $\boldsymbol{\mathcal{I}}(\boldsymbol{\psi})$, which quantifies the precision of estimating $\boldsymbol{\psi}$ from $\mathcal{T}$ (Brunel and Nadal, 1998). Recent work from (Huang and Zhang, 2018) 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

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

with $\mathcal{H}(\boldsymbol{\psi}) := \mathbb{E}_0\left[-\log \mathrm{P}(\boldsymbol{\psi})\right]$ as the differential entropy of $\boldsymbol{\psi}$ (Cover and Thomas, 2006).

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 contributes little or no information (Brunel and Nadal, 1998). For sharper priors, $\mathbb{I}(\boldsymbol{\psi};\, \mathcal{T}) \approx \mathbb{I}(\boldsymbol{\mathcal{J}})$ as the prior contribution is significant – using $\mathbb{I}(\boldsymbol{\mathcal{I}})$ would lead to large errors (Huang and Zhang, 2018). 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 (Lehmann and Casella, 1998; Brunel and Nadal, 1998; Huang and Zhang, 2018). The skyline-based inference problems that we consider here automatically satisfy (i) and (ii) as these models belong to an exponential family. Condition (iii) is satisfied for moderate to large trees (and asymptotically) (Lehmann and Casella, 1998; Parag and Pybus, 2019).

Using the above approximations, we derive the interesting expression

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

which 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 (Huang and Zhang, 2018), 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 (Slate, 1994) 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 ($n \geqslant 8$). Thus, $\log \boldsymbol{N}$, which is also optimal for experimental design (Parag and Pybus, 2019), ensures the validity of Ω on small trees. This is the parametrisation adopted by the Skyride and Skygrid methods (Minin *et al.*, 2008). Other (cubic-root) parametrisations under which Ω would be valid at even smaller n also exist (Slate, 1994).

Eq. (4)–Eq. (7) are not restricted to coalescent inference problems and are generally applicable to statistical models that involve exponential families (Lehmann and Casella, 1998). We now specify Ω for skyline-based models, which all possess piecewise-constant population sizes and orthogonal $\boldsymbol{\mathcal{I}}(\boldsymbol{\psi})$ matrices (Parag and Pybus, 2019). These properties permit the expansion (Ipsen and Rehman, 2008):

$$\det\left[\boldsymbol{\mathcal{I}}(\boldsymbol{\psi}) + \boldsymbol{\mathcal{P}}(\boldsymbol{\psi})\right] = \det\left[\boldsymbol{\mathcal{I}}(\boldsymbol{\psi})\right] + \det\left[\boldsymbol{\mathcal{P}}(\boldsymbol{\psi})\right] + \sum_{j=1}^{p-1} \gamma_j,$$
$$\text{with } \gamma_j = \sum d_{i_1} \ldots d_{i_j} \det\left[\boldsymbol{\mathcal{P}}(\boldsymbol{\psi})_{\bar{\boldsymbol{i}}_1 \ldots \bar{\boldsymbol{i}}_j}\right],$$

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

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

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

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.

The Kingman Conjugate Prior

Kingman's coalescent process (Kingman, 1982), 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 (Parag and Pybus, 2017) so that $\binom{l_{c_{i-1}}}{2}(c_i - c_{i-1}) \sim \exp(N_1^{-1})$ for $1 \leqslant i \leqslant n-1$ with $c_0 = 0$. If y defines the space of N_1^{-1} values, and has prior distribution $\mathrm{P}(y)$, then, by (Snyder and Miller, 1991), 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 (Fink, 1997) i.e. $N_1^{-1} \sim \mathrm{Gam}\left(m_0, \, \bar{T}_0\right)$ with shape m_0 and rate $\bar{T}_0$. The posterior distribution is then $N_1^{-1} \,|\, \mathcal{T} \sim \mathrm{Gam}\left(m + m_0, \, \bar{T} + \bar{T}_0\right)$ with $m = n - 1$ counting coalescent events in $\mathcal{T}$ (Robert, 2007). Transforming to N_1 implies $N_1 \,|\, \mathcal{T} \sim \mathrm{Gam}^{-1}\left(m + m_0, \, \bar{T} + \bar{T}_0\right)$. This is an inverse gamma distribution with mean $\frac{\bar{T}+\bar{T}_0}{m+m_0-1}$, shape $m + m_0$ and inverse rate $\bar{T} + \bar{T}_0$. 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 coalescent data, which contributes m coalescent events from $\mathcal{T}$, with scaled TMRCA of $\bar{T}$ (Robert, 2007). 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 $\mathcal{I}(N_1^{-1}) = m N_1^2$ 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

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

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 (Cover and Thomas, 2006). 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 rate parameter, has no effect on estimate precision or mutual information.

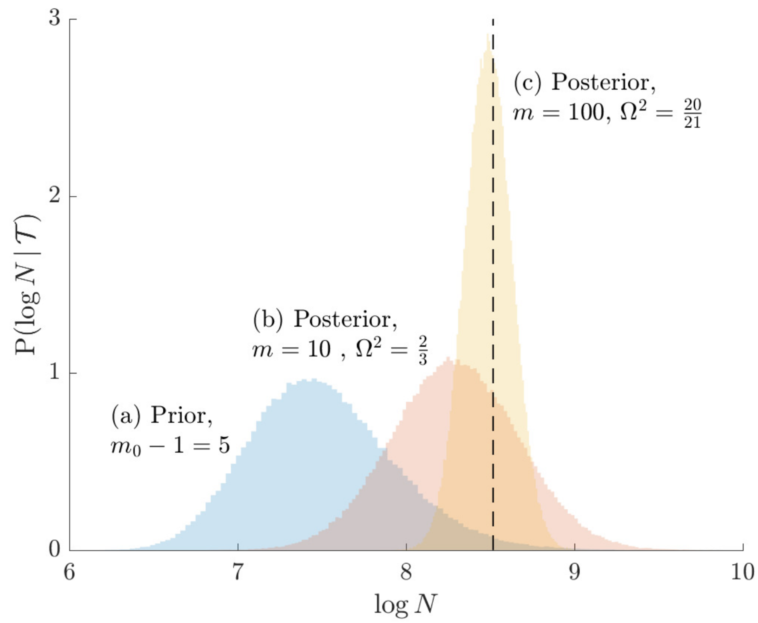

Fig. 1: **Effect of conjugate prior on Kingman coalescent estimation.** We examine the relative impact on estimate precision of a conjugate Kingman prior that contributes $m_0 - 1 = 5$ virtual observations. We work in $\log N_1$ for convenience. We compare this prior to posteriors, which are obtained 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 towards 1.

Our ratio Ω therefore provides a systematic decomposition of the posterior population size 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.

Skyline Smoothing Priors

In this section, we tailor Ω for the BSP, Skyride and Skygrid coalescent inference methods. These popular skyline-based approaches couple a piecewise-constant demographic coalescent data 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 coalescent data $\mathcal{T}$. While these priors can embody sensible biological assumptions, we show that they may 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 uniquely objective (i.e. uninformative) reference skyline priors, which we denote $\mathrm{P}^*(\boldsymbol{\psi})$. Finding objective priors for multivariate statistical models is generally non-trivial, but (Berger *et al.*, 2015) state that if $\boldsymbol{\mathcal{I}(\psi)}$ has form $[f_1(\psi_1)g_1(\boldsymbol{\psi_{-1}}),\, \dots,\, f_p(\psi_1)g_p(\boldsymbol{\psi_{-p}})]\ \mathrm{I}_p$ then $\mathrm{P}^*(\boldsymbol{\psi}) \propto \prod_{j=1}^{p} \sqrt{f_j(\psi_j)}$. Here f_j and g_j are some functions and $\boldsymbol{\psi_{-j}}$ symbolises the vector $\boldsymbol{\psi}$ excluding ψ_j. Following this, we get

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

with Z_1, Z_2 as normalisation constants. Given its optimal properties (Parag and Pybus, 2019), 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. $\mathbb{E}_0\left[\boldsymbol{\mathcal{I}}\right] = \boldsymbol{\mathcal{I}_0}$. In addition, the reference prior in this case is $\boldsymbol{\mathcal{P}^*} = \mathbf{0}_{\boldsymbol{p}}$, with $\mathbf{0}_{\boldsymbol{p}}$ 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 subjective by this definition, 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 (Drummond *et al.*, 2005), and

(ii) the *Gaussian Markov Random Field* (GMRF) prior employed in both the Skyride and Skygrid methods (Minin *et al.*, 2008) (Gill *et al.*, 2013).

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

The SMP is defined as: $\text{P}(\boldsymbol{N}) = {}^{1}\!/_{N_1} \prod_{j=2}^{m} {}^{1}\!/_{N_{j-1}}\, e^{N_j/N_{j-1}}$ (Drummond *et al.*, 2005). 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 $u_j = e^{\log N_{j+1} - \log N_j} = {}^{N_{j+1}}\!/_{N_j}$ for $j \in \{1, \ldots, p-1\}$. In the Appendix we show how this expression yields Eq. (A1) and hence the transformed prior $\text{P}(\log \boldsymbol{N}) = \prod_{j=1}^{p-1} u_j e^{-u_j}$. 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 tridiagonal pattern of Eq. (10).

An issue with the SMP is its dependence on the unknown 'true' demographic parameter values. We cannot evaluate (or control) *a priori* how much information is contributed by this smoothing prior. Rapidly declining populations could feature ${}^{N_{j+1}}\!/_{N_j} > m_j$, for example, which would result in prior over-reliance. Conversely, exponentially growing 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. Some recent implementations of the BSP include an alternative log-normal prior that links N_j with N_{j-1} (Bouckaert *et al.*, 2019), which is conceptually similar to the GMRF below.

The possibility of strong or inflexible prior assumptions under the BSP motivated the development of the GMRF for the Skyride and Skygrid methods (Minin *et al.*, 2008). 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 $\mathrm{P}(\log \boldsymbol{N}) = Z^{-1}\tau^{\frac{p-2}{2}}e^{-\frac{\tau}{2}\sum_{j=1}^{p-1}\delta_j^{-1}(\log N_{j+1}-\log N_j)^2}$ (Minin *et al.*, 2008). 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:

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

The Appendix provides the general derivation for any $p \geqslant 3$. 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 $\mathrm{P}(\log \boldsymbol{N} \mid \mathcal{T})$) 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 also hint at how smoothing priors influence the statistical properties of our coalescent inference problem.

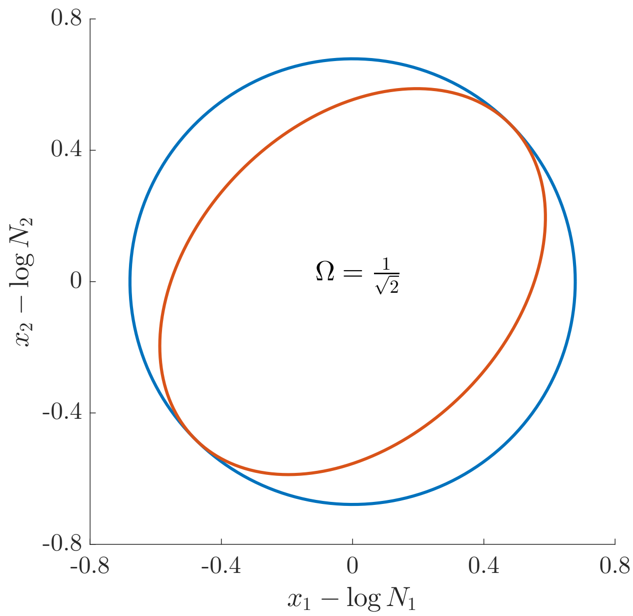

Fig. 2: **Uncertainty ellipses for SMP and GMRF.** We show the improvement in asymptotic precision rendered by use of a smoothing prior for a $p = 2$ segment skyline inference problem. The prior informed ellipse (red) is smaller in volume and has skewed principal axes relative to the purely data informed one (blue). All ellipses represent 99% confidence with the x_j indicating coordinate directions about their means, which are the log population sizes, $\log N_j$. The covariance that smoothing introduces controls the skew of these ellipses. Here $\Omega^2 = {}^1\!/_2$, $m = 40$ (total coalescent event count) and $a = 10$ (this controls the prior influence see Eq. (12)). Larger a values lead to over-reliance on the smoothing prior.

To solidify these ideas, we provide a visualisation of Ω and an example of $\boldsymbol{S}$. We consider the simple $p = 2$ case, where the posterior Fisher information and Ω for the GMRF and SMP both 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.

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 (Parag and Pybus, 2019). This 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 $a = {}^{m}\!/_{4}$. These are constructed in coordinates $\boldsymbol{x} = [x_1,\, \ldots,\, x_p]$ centred about population size means $\log \boldsymbol{N}$ as $(\boldsymbol{x} - \log \boldsymbol{N})^{\intercal} \boldsymbol{\mathcal{X}} (\boldsymbol{x} - \log \boldsymbol{N}) = c$ with c controlling the confidence level.

Here $\boldsymbol{\mathcal{X}}$ is either $\boldsymbol{\mathcal{I}}$ or $\boldsymbol{\mathcal{J}}$. Because $\boldsymbol{\mathcal{I}}$ is diagonal the data-informed confidence ellipse has principal axes aligned with $\log \boldsymbol{N}$. The covariance among population size segments in $\boldsymbol{\mathcal{J}}$, which is induced by the smoothing prior, skews these principal axes. We can see this by diagonalising $\boldsymbol{\mathcal{J}}$ at $m_1 = m_2 = {}^{m}\!/_{2}$ and for every r to obtain:

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

Applying $\boldsymbol{S}$, we find that the axes of our uncertainty ellipse (as visible in Fig. 2) have changed from $\left\{ \left(\begin{smallmatrix} \log N_1 \\ 0 \end{smallmatrix} \right),\, \left(\begin{smallmatrix} 0 \\ \log N_2 \end{smallmatrix} \right) \right\}$ to $\left\{ \left(\begin{smallmatrix} \log N_1 - \log N_2 \\ 0 \end{smallmatrix} \right),\, \left(\begin{smallmatrix} 0 \\ \log N_1 + \log N_2 \end{smallmatrix} \right) \right\}$. Sums and differences of log-populations are now the parameters that can be most naturally estimated under the SMP and GMRF. The reduction in the area of the ellipses of Fig. 2 is a proxy for Ω.

The Dangers of Smoothing

Having defined ratios for measuring the contribution of smoothing priors to the precision of estimates, 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 $1 \leqslant j \leqslant p$ (Parag and Pybus, 2019), with p as the total number of skyline segments. 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. A1 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 coalescent 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 is 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), where a controls the prior contribution to $\boldsymbol{\mathcal{J}}$. 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 $N_2 \geqslant {}^{m}\!/_{4}\, N_1$ or $\tau \geqslant {}^{m}\!/_{4}\, \delta_1$, 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 the number of skyline segments, p, increase, 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 $\boldsymbol{\mathcal{J}}_{\text{GMRF}}$ then only depends on $\{m_j\}$ and τ. We compute r and hence Ω, at various p. For example we find that

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

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 (Parag and Pybus, 2019). 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 (Minin *et al.*, 2008). We show the corresponding mutual information increases due to these GMRF priors in Fig. A2 of the Appendix.

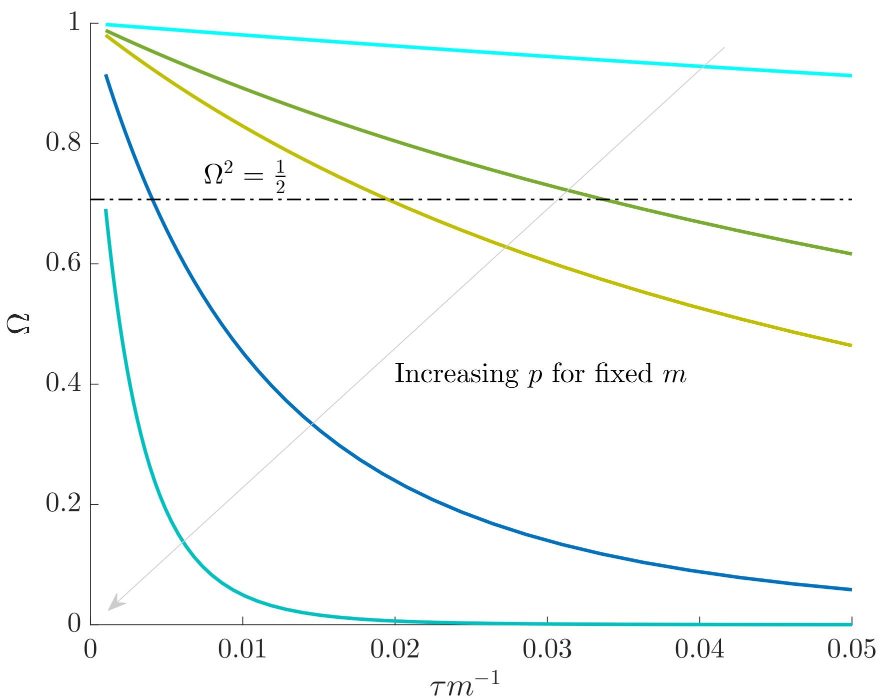

Fig. 3: **The impact of smoothing priors increases with skyline complexity.** For the GMRF, we find that for a fixed τ/m (ratio of smoothing parameter to total coalescent event count), Ω significantly depends on the complexity, p, of our skyline. The coloured Ω curves are (along the arrow) for $p = [2, 4, 5, 10, 20]$ at $m = 100$ with $m_j = m/p$ as the number of coalescent events per skyline segment. The dashed $\Omega^2 = 1/2$ line depicts the threshold below which the prior contributes more than the coalescent data to posterior estimate 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, Skyride and Skygrid 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 $\min_{\{1 \leqslant j \leqslant p-1\}} {}^{N_{j+1}}/_{N_j}$, 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 suspected BSP issues related to poor estimation precision when skylines are too complex, or the coalescent data are not sufficiently informative (Ho and Shapiro, 2011). However, good use of the BSP grouping parameter (Drummond *et al.*, 2005), which sets $p < m$, could alleviate these problems.

(2) Skyride. When this method uses the uniform GMRF, all results apply exactly. 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 (Minin *et al.*, 2008). However, even with these adjustments, the GMRF can over-smooth, and fail to recover population size changes (Ho and Shapiro, 2011; Faulkner *et al.*, 2019). 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 $\min_{\{1 \leqslant j \leqslant p-1\}} {}^{\tau}/_{\delta_j}$, 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 (Gill *et al.*, 2013). 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) (Gill *et al.*,

2013), 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 (Parag and Pybus, 2019; Parag *et al.*, 2020). Consequently, if $p > m$ is set, smoothing priors can even mask identifiability problems. We recommend that $\frac{m}{p} \geqslant \kappa > 1$ must be guaranteed and in the next section derive a model rejection guideline for finding κ, the suggested minimum number of coalescent events per skyline segment, and diagnosing prior over-reliance.

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 are 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 $\boldsymbol{\mathcal{J}}$ 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, $\det[\boldsymbol{\mathcal{J}}] \leqslant (1/p \, \mathrm{tr}[\boldsymbol{\mathcal{J}}])^p$, then applies with tr denoting the matrix trace. Since $\mathrm{tr}[\boldsymbol{\mathcal{J}}] = m + \mathrm{tr}[\boldsymbol{\mathcal{P}}]$ we can expand this inequality and substitute in Eq. (4) to get $\Omega^2 \geqslant (1/p\,(m + \mathrm{tr}[\boldsymbol{\mathcal{P}}]))^{-p} \prod_{j=1}^{p} m_j$.

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

$$\omega^2 = \left(\frac{m}{m + \mathrm{tr}[\boldsymbol{\mathcal{P}}]}\right)^p \implies p^* = \arg\max_{p \leqslant m} \omega^2 \geqslant b. \tag{14}$$

We define $b \geqslant 1/2$ as a conservative model rejection criterion with $\omega^2 \geqslant b$ implying that $\Omega^2 \geqslant b$. 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 coalescent tree data.

Alternatively, we recommend that skylines using a smoothing prior (with matrix $\boldsymbol{\mathcal{P}}$) should have at least $\kappa = {}^{m}/_{p^*}$ events per segment to avoid prior reliance. The $p \leqslant m$ condition in Eq. (14) ensures skyline identifiability (Parag and Pybus, 2019) and generally $p^* \leqslant {}^{m}/_{2}$ (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. (A2)) we derive analogous rejection bounds based on the excess mutual information, $\Delta\mathbb{I}$, from Eq. (7). There we find that p acts like an information-theoretic bandwidth, controlling the prior-contributed mutual information.

Eq. (14), which forms a key contribution of this work, can be computed and is valid for any smoothing prior of interest. For the uniform GMRF where $\mathrm{tr}\left[\boldsymbol{\mathcal{P}}\right] = 2\tau(p-1)$, 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. A4 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. A5 we plot p^* curves for various m and τ, defining boundaries beyond which skyline estimates will be overly-dependent on the GMRF.

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. (A3)). 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. A3. Other revealing prior-specific insights can be obtained from Eq. (14), reaffirming its importance as a model rejection statistic.

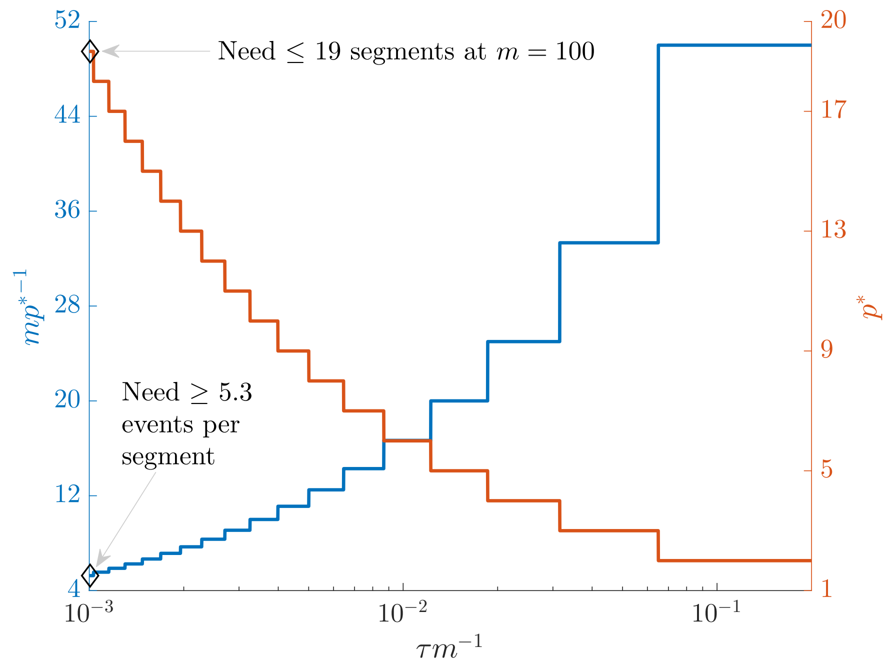

Fig. 4: **Bounding skyline complexity using the prior-data tradeoff.** For the GMRF with uniform smoothing, we show how the maximum number of recommended skyline segments, p^* (red), decreases with prior contribution (level of smoothing i.e. increasing τ/m). Hence the minimum recommended number of coalescent events per segment, $\kappa = m/p^*$ (blue), rises. Here we use the $\omega^2 \geqslant b = 1/2$ boundary (Eq. (14)), which approximates Ω^2 and provides a more easily computed measure of prior-data contributions. 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.

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; because while p^* warns against the prior being too relatively influential, it does not guarantee any absolute estimate precision e.g. a small (m, τ) pair might produce the same p^* as a larger pair. Choosing an optimal p in a data-justified manner is an open problem that is still under active study (Parag and Donnelly, 2020). We next illustrate how Ω^2, via its more easily computed approximation, ω^2, can be practically applied to detect and reject over-smoothed skyline plot models, using datasets that are commonly employed to evaluate the performance of coalescent demographic inference.

Illustrative Examples: Egyptian HCV and Beringian Bison

We validate the practical utility of ω^2 (and hence Ω^2), as a diagnostic of prior over-dependence, by investigating changes in effective population size inferred from the well-studied Egyptian HCV-4 (Pybus *et al.*, 2003) and Beringian steppe bison Shapiro *et al.* (2004) datasets. The first consists of 63 partial sequences of HCV genotype 4 and was previously analysed in (Pybus *et al.*, 2003) using a coalescent model with a parametric demographic function that featured periods of constant population size separated by a phase of exponential growth. The second dataset comprises 152 modern and partial mtDNA and was investigated in Shapiro *et al.* (2004), where skyline plot models confirmed a demographic history of exponential growth then decline (boom-bust) with an additional bottleneck dynamic (Drummond *et al.*, 2005). These two datasets have since been re-examined under various alternate models in (Minin *et al.*, 2008; Gill *et al.*, 2013; Parag *et al.*, 2020) and several other studies.

We simulated 100 trees with $m + 1 = n = 63$ and 152 tips, using the software package MASTER (Vaughan and Drummond, 2013), according to inferred HCV and bison population size trends respectively. The HCV population size trend that we simulated from is provided in (Pybus *et al.*, 2003). We inferred the population size trend of the bison dataset using the BSP (with sequential Markovian prior) in accordance with published analyses (Drummond *et al.*, 2005). We used 20 population groups and the optimal design from (Parag and Pybus, 2019) to ensure that we captured complex bison population dynamics reliably. As our focus is on exploring the behaviour of skylines and ω^2 given a particular underlying population size trend and not the uncertainty associated with that trend, we used the posterior mean (HCV) or median (bison) of these inferred trends for simulating trees and do not consider genealogical uncertainty.

The simulated set of coalescent trees from each dataset provide an approximate measure of the coalescent variance that could arise from the inferred underlying population size trends. We then estimated $\log \boldsymbol{N}$ from every simulated tree using various skyline models with time-aware GMRF smoothing priors, as in (Minin *et al.*, 2008). We varied the relative contributions of the coalescent data and GMRF to our posterior log-population size estimates by changing either the skyline dimension, p, or the GMRF smoothing parameter τ. As m is fixed for a given dataset and robust designs are applied, increasing the number of coalescent events in each segment, m_j, reduces p.

We analysed every tree over all combinations of $m_j \in \{1, 2, 4, 8\}$ across a wide range of τ. For comparison, we also generated purely data-informed estimates of $\log \boldsymbol{N}$, for the same m_j, by replacing the subjective GMRF with a uniform, objective prior. We computed ω^2 from Eq. (14) for these settings in Fig. 5 and observe that, as expected, it decreases with both τ and p (i.e. ω^2 increases with m_j). Practical analyses of these datasets using Skyride or Skygrid approaches, would choose or infer a τ value and set $p \approx m$. However, Fig. 5 shows $\kappa = {}^{m}/_{p^*} > 1$ and hence $m_j > 1$ events per skyline parameter are often necessary to achieve $\omega^2 \geqslant {}^1/_2$. This raises questions about the validity of the common practice of applying these methods using their default settings.

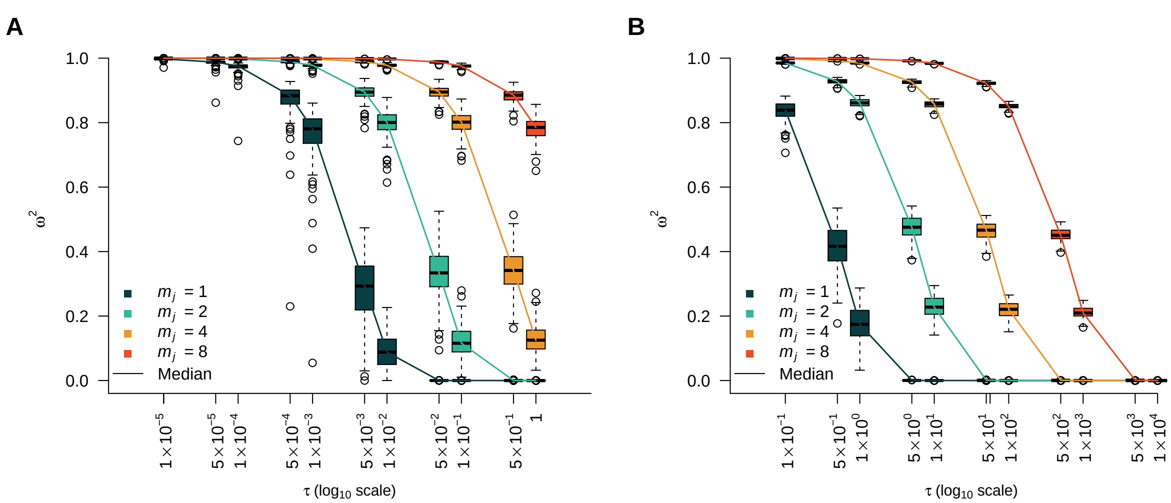

Fig. 5: **Model rejection statistics for the HCV and bison datasets** The metric ω^2 is calculated for each tree (see Eq. (14)) under a time-aware GMRF for various combinations of its smoothing parameter τ and m_j, the number of coalescent events per skyline segment. The box-plots summarise the resulting ω^2 over 100 simulated trees that represent the demographic histories of the (A) Egyptian HCV and (B) Beringian bison datasets. The solid lines link the median values across boxes for a given m_j and hence skyline dimension p ($m_j = {}^{m}/_{p}$). We discourage the use of skyline models with $\omega^2 < {}^1/_2$.

Fig. 5 confirms that the recommended maximum skyline dimension p^* falls and hence the minimum allowable number of coalescent events per segment m_j grows as the smoothing parameter τ increases. We demonstrate the qualitative difference in skyline-based estimates between p values on either side of the p^* criterion for a single simulated HCV and bison tree in Fig. 6. In panels A and C we present the Skyride estimate, which uses $m_j = 1$ and implements $p > p^*$, at the chosen τ values (0.05 and 1). Contrastingly, in B and D, we illustrate an equivalent skyline with a different m_j, which achieves $p < p^*$ at this same τ, according to our ω^2 metric (see the $m_j = 4$ and $m_j = 2$ curves at $\tau = 0.05$ and 1 in panels A and B of Fig. 5) respectively). We overlay the corresponding skyline (with the same m_j) obtained with an objective uniform prior, to visualise the uncertainty engendered from the coalescent data alone.

At $m_j = 1$ (panels A and C of Fig. 6), the uniform prior produces a skyline that infers more rapid demographic fluctuations through time than that estimated with the GMRF prior. Further, the 95% HPD intervals from the uniform prior (red) are substantially wider than those from the GMRF prior (blue) in both examples, highlighting the marked contribution of the time-aware GMRF prior to posterior estimate precision. While this smoothed trajectory looks reliable we argue that, because $p > p^*$ (and hence $\omega^2 < \frac{1}{2}$), it is difficult to justify using the data alone and that the prior is responsible for too much of the estimate precision. In contrast, at $m_j = 4$ and $m_j = 2$ (panels B and D of Fig. 6), which apply $p < p^*$, both prior distributions yield more similar skylines, implying that GMRF smoothing has not substantially inflated posterior estimate precision.

Under these settings we have fewer demographic fluctuations than for $m_j = 1$ because 4 and 2 times more coalescent events are informing each parameter or skyline segment, respectively. We achieve smaller uncertainty than $m_j = 1$ with a uniform prior (which is overfitted) but without excessively relying on the GMRF smoothing, which at $m_j = 1$ is likely underfitting. The ω^2 metric and hence p^* criterion help us better balance data, noise and our prior assumptions. In contextualising these results it is important to

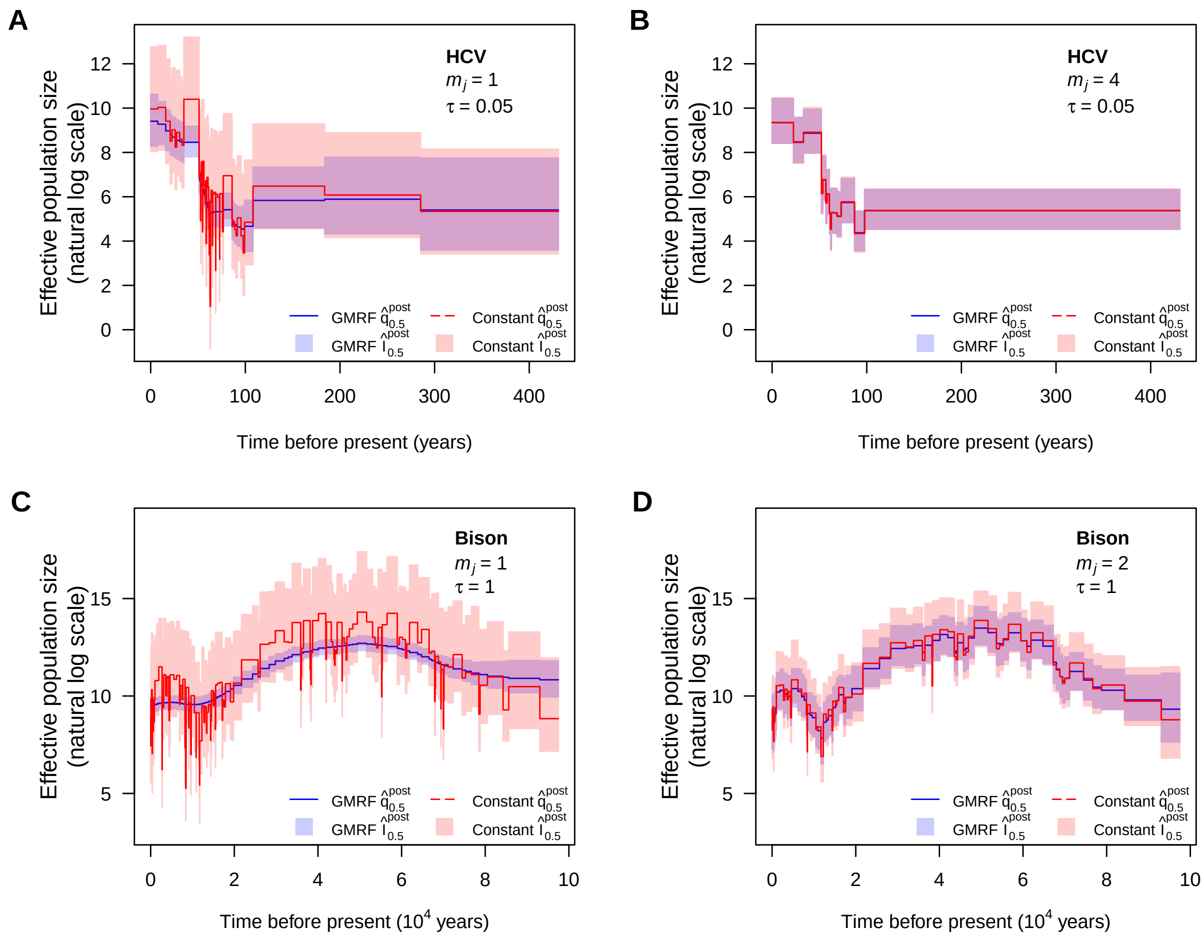

Fig. 6: **HCV and bison demographic estimates under GMRF and uniform priors.** We analyse demographic estimates under time-aware GMRF priors (blue) and objective uniform priors (red) for a single tree simulated under the demographic scenarios inferred from the Egyptian HCV (A and B) and Beringian bison (C and D) datasets. In panels A and C we present Skyride estimates, which use $m_j = 1$ and $\tau = 0.05$ (A) and 1 (C). These skylines have dimension p that is larger than our maximum recommended dimension p^*, which is computed from Fig. 5. In panels B and D we re-estimate population size at $m_j = 4$ (B) and 2 (D). These groupings of coalescent events achieve $p < p^*$ as justified by our ω^2 metric (see Eq. (14)). Solid lines are posterior medians while semi-transparent blocks are the 95% HPD intervals.

note that skyline plots provide harmonic mean and not point estimates of population size (Pybus *et al.*, 2000). Consequently, we are inferring sequences of means from our coalescent data, which *a priori* may not need to conform to a smooth pattern.

The HCV example shows that for times beyond $t > 100$ 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 and overfitted means (panel A). In contrast, for the bison example, the bottleneck over $10^4 < t < 2 \times 10^4$ years is oversmoothed (panel C), despite many coalescent events occurring in that region. The simple correction of extending our harmonic mean over 2 events (panel D) restores the necessary fall in population size. Deciding on how to balance uncertainty with model complexity is non-trivial and, as shown in these examples, caution is needed to avoid misleading conclusions. We posit that ω (and hence Ω) can help formalise this decision-making and improve our quantification of the uncertainty across skyline plots.

Having confirmed Ω as a credible measure of relative uncertainty, we briefly explore how it relates to more easily ascertained measures of uncertainty. For each simulated coalescent tree in the HCV example above 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 (across the skyline segments) $\mathbb{H}_{\tau,m}$, which are formulated as:

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

with med indicating 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 when the objective uniform prior is applied instead.

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. A6 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 more coalescent data are informing each estimated parameter (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 Skyride and Skygrid methods, both statistics are markedly below $\frac{1}{2}$ and posterior estimates will often be too dependent on the prior. This high-p behaviour is also indicative of model over-parametrisation Parag and Donnelly (2020). Our metric Ω therefore relates sensibly to visible and common proxies of uncertainty.

Discussion

Popular approaches to coalescent inference, such as the BSP, Skyride and Skygrid 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 smoothness that is expected to be exhibited by real population size curves over time, has led to many insights in phylodynamics (Ho and Shapiro, 2011). However, it has also spawned concerns related to over-smoothing and lack of methodological transparency (Minin *et al.*, 2008) (Faulkner *et al.*, 2019). In this work we attempted to address these concerns by deriving metrics for diagnosing and clarifying the existing assumptions present in current best practice.

Detecting and correcting for underfitting or over-smoothing is crucial if reliable and meaningful assessments of the effective population size changes of a species or pathogen of interest are to be made from sequence data. Abrupt changes in effective population size are not only biologically plausible but may also signal key events that have shaped the demographic histories of populations (Pyron and Burbink, 2013). In ecology, identifying rapid extinctions and bottlenecks in diversity might signify the impact of environmental change or anthropogenic influences (e.g., hunting or changes in land use) (Stiller *et al.*, 2010; Thomas *et al.*, 2019). Similarly, in epidemiology, sharp fluctuations in the prevalence of an infection might support hypotheses about emergence in novel populations, seasonality, the effect of interventions, vaccines, or drug treatments. Further, rapid exponential growth of any population may, when observed over a longer timescale, appear as a near-stepwise transition in population size.

Underfitting these changes would limit understanding of the dynamics of the study population and could affect conclusions about the potential causative factors that influenced those dynamics. However, recognising when commonly used methods for inferring these demographic trends are over-smoothing is difficult. By capitalising on (mutual) information theory and (Fisher) information geometry we formulated the novel coalescent information ratio, Ω, which provides a rigorous means of solving this over-smoothing problem. 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 assumptions introduce.

We derived analytic expressions for Ω for the BSP, Skyride and Skygrid estimators of effective population size, which combine piecewise skyline likelihoods with either SMP or GMRF smoothing priors. We also showed that Ω has an exact and intuitive interpretation as the ratio of real coalescent events to the sum of real and virtual (prior-contributed) ones in a Kingman coalescent model. Using $\Omega^2 = {}^1/_2$ as a threshold delimiting when the prior contributes as much information as the coalescent data, we found that it is easy to become overly dependent on prior assumptions as the skyline dimension, p, increases (for a fixed tree size). 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 (Minin *et al.*, 2008; Drummond *et al.*, 2005), while the Skygrid may have no events informing some parameters (Gill *et al.*, 2013).

These issues can be obscured by current Bayesian implementations, which can still produce apparently reasonable population size estimates, at least visually, as illustrated in our simulated HCV and bison case studies. Our simulations indicate that analyses that combine maximally parametrised skylines (one event per segment or parameter) with GMRF smoothing can lead to errors in population size inference. For trees simulated according to the HCV demographic scenario, estimates were likely overfitted in the far past, inflating HPDs, but oversmoothed towards the present. The resulting skyline uncertainty contrasted that from the original Pybus *et al.* (2003) and later (Parag and Pybus, 2017) analyses. In the bison example, we found evidence for underfitting. The inferred skyline there emphasised a smoother boom-bust trend with concentrated HPDs. However, this underestimated the depth of a bottleneck during which coalescent events were concentrated.

These mismatches between data and smoothing can be difficult to diagnose and problematic, not just for prior over-dependence. Low coalescent event counts, for example, can lead to poor statistical identifiability (Rothenburg, 1971) which might manifest in spurious MCMC mixing. Consequently, we proposed a practical p^* rejection criterion for ensuring that coalescent data is the main source of inferential information. This criterion, which was based on an approximation to Ω^2, provided a way of regularising skyline complexity. When applied to our examples it recommended a 4-event skyline grouping that resulted in demographic reconstructions that were more consistent with the above mentioned HCV studies. It also suggested a simple 2-event grouping that recovered the bison bottleneck dynamic without generating too much estimate noise.

This p^* criterion bounds the maximum recommended skyline dimension for a given dataset (tree) size and provides a usable means of defining the minimum number of coalescent events, κ, which we should allocate to each skyline segment to guard against too much prior influence. Since κ only requires our computing the sum of the diagonals of the prior Fisher matrix, it can serve as a simple rule-of-thumb for sensibly balancing the prior-data tradeoff in skyline plots (e.g. in the BSP, the grouping parameter might be set to a value above κ to ensure well-regularised estimates). As we found Ω^2 to be lower-bounded by more visible measures of skyline uncertainty, such as the product of relative HPD widths, useful approximations to p^* and κ may also be computed from these measures.

Our Ω metric also provides insight into how we can alleviate the dramatic impact of skyline complexity on prior over-reliance. When specialised to the GMRF, for example, it reveals that we can negate over-smoothing by scaling the smoothing parameter τ with a quadratic of p. Moreover, it shows that only by increasing the information available from the sampled phylogeny can we reasonably allow for more complex piecewise-constant functions under a given prior. Recent methods, such as the *epoch sampling skyline plot* (Parag *et al.*, 2020), which can double the Fisher information extracted from a given phylogeny by exploiting the informativeness of sampling times, would support higher dimensional skylines. Such approaches have the potential to increase the contribution of the data without elevating the influence of the smoothing prior.

While in this paper we have applied Ω to non-parametric, skyline inference problems in population genetics, ecology and epidemiology, its general formulation in Eq. (4) is more widely applicable. It can be also applied to coalescent inference problems where specific parametric models (e.g., exponential/logistic growth) are used, in order to disentangle the contributions of observed data and the prior distributions over these parameters, though numerical solutions will likely be necessary. More generally, our approach is valid for any statistical problem, provided the Hessian matrices necessary for deriving the prior and data Fisher information terms are valid and computable. This is not limited to prior-data tradeoffs. Similar ratio metrics should be derivable by comparing Fisher information terms from different sources (e.g. to test whether one source of data is more informative than another).

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. Our work also aligns with recent studies, which have started to re-examine both model selection and prior definition (Parag and Donnelly, 2020; Faulkner *et al.*, 2019) in an attempt to derive more reliable effective population size estimates from coalescent trees. While we believe that data-driven conclusions are generally the most justifiable we note that, in the context of skyline plots, this can be open to interpretation and the choice of prior is far from trivial.

Funding

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). This work was supported by the Oxford Martin School.

Acknowledgments

We thank Louis du Plessis for his useful comments and insights on this project.

Supplementary Material

Data (and code in Matlab) available from the Dryad Digital Repository:
`https://datadryad.org/stash/dataset/doi:10.5061/dryad.1jwstqjs2`

Literature Cited

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

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

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

Bouckaert, R., Vaughan, T., Barido-Sottani, J., *et al.* (2019). BEAST 2.5: An Advanced Software Platform for Bayesian Evolutionary Analysis. *PLoS Computational Biology*, **15**(4), e1006650.

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

Cover, T. and Thomas, J. (2006). *Elements of Information Theory Second Edition*. John Wiley and Sons.

Drummond, A., Nicholls, G., Rodrigo, A., *et al.* (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, **161**, 1307–20.

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

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

Fink, D. (1997). A Compendium of Conjugate Priors. Technical report, Montana State University.

Gill, M., Lemey, P., Faria, N., *et al.* (2013). Improving Bayesian Population Dynamics Inference: A Coalescent-Based Model for Multiple Loci. *Molecular Biology and Evolution*, **30**(3), 713–24.

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

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

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

Ipsen, I. and Rehman, R. (2008). Perturbation Bounds for Determinants and Characteristic Polynomials. *SIAM Journal on Matrix Analysis and Applications*, **30**(2), 762–76.

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

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

Lehmann, E. and Casella, G. (1998). *Theory of Point Estimation*. Springer-Verlag, second edition.

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

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

Parag, K. and Donnelly, C. (2020). Adaptive Estimation for Epidemic Renewal and Phylogenetic Skyline Models. *Systematic Biology*, **69**(6), 1163–79.

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

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

Parag, K., du Plessis, L., and Pybus, O. (2020). Jointly inferring the dynamics of population size and sampling intensity from molecular sequences. *Molecular Biology and Evolution*, **37**(8), 2414–29.

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

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

Pyron, R. and Burbink, F. (2013). Phylogenetic Estimates of Speciation and Extinction Rates for Testing Ecological and Evolutionary Hypotheses. *Trends in Ecology and Evolution*, **28**(12), 729–36.

Robert, C. (2007). *The Bayesian Choice*. Springer Texts in Statistics. Springer Science + Business Media.

Rodrigo, A. and Felsenstein, J. (1999). *Coalescent Approaches to HIV-1 Population*. The Evolution of HIV. Johns Hopkins University Press.

Rothenburg, T. (1971). Identification in Parametric Models. *Econometrica*, **39**(3).

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

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

Snyder, D. and Miller, M. (1991). *Random Point Processes in Time and Space*. Springer-Verlag, 2 edition.

Stiller, M., Baryshnikov, G., Bocherens, H., *et al.* (2010). Withering away-25,000 years of genetic decline preceded cave bear extinction. *Molecular Biology and Evolution*, **27**(5), 975–8.

Strimmer, K. and Pybus, O. (2001). Exploring the Demographic History of DNA Sequences using the Generalized Skyline Plot. *Mol. Biol. Evol*, **18**(12), 2298–305.

Thomas, J., Carvalho, G., Haile, J., *et al.* (2019). Demographic reconstruction from ancient dna supports rapid extinction of the great auk. *eLife*, **8**, e47509.

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

van Trees, H. (1968). *Detection, Estimation, and Modulation Theory, Part I*. John Wiley and Sons Inc.

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

Wakeley, J. (2008). *Coalescent Theory: An Introduction*. Roberts and Company Publishers.

Appendix

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 (Drummond *et al.*, 2005) 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 $\boldsymbol{\eta} = [\eta_1, \ldots, \eta_p] = [\log N_1, \ldots, \log N_p]$. We want the transformed prior $g(\boldsymbol{\eta})$. Applying the multivariate change of variables formula gives $g(\boldsymbol{\eta}) = f(e^{\boldsymbol{\eta}})|\det[\Delta\rho^{-1}]|$, with $\Delta\rho^{-1} = [e^{\eta_1}, \ldots, e^{\eta_p}]\,\mathrm{I}_p$ as the Jacobian of ρ^{-1}. This implies that $|\det[\Delta\rho^{-1}]| = e^{\sum_{j=1}^{p} \eta_j}$.

Substituting and expanding gives the SMP log-prior:

$$\log g(\boldsymbol{\eta}) = \eta_p - \eta_1 + \sum_{j=2}^{p} -e^{\eta_j - \eta_{j-1}}. \tag{A1}$$

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 $2 \leqslant j \leqslant p-1$, ${\partial^2 \boldsymbol{G}}/{\partial \eta_1^2} = -e^{\eta_2 - \eta_1}$ 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 $\boldsymbol{\mathcal{I}} = [m_1, \ldots, m_p]\, \mathrm{I}_p$ (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 $g(\boldsymbol{\eta}) = Z^{-1} \tau^{\frac{p-2}{2}} e^{-\frac{\tau}{2} \sum_{j=1}^{p-1} \delta_j^{-1} (\eta_{j+1} - \eta_j)^2}$ (Minin *et al.*, 2008) 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 $2 \leqslant j \leqslant p-1$, ${\partial^2 \boldsymbol{G}}/{\partial \eta_1^2} = -{\tau}/{\delta_1}$ 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 $\boldsymbol{\mathcal{J}}_{\text{GMRF}}$.

Further Smoothing Results

In the main text we asserted that the Ω computed at the robust point of $m_j = {m}/{p}$ (Parag and Pybus, 2019) 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. A1. 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. A1 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. A2. These show how larger values of the smoothing 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 the main text we explained how the Skyride implicitly implements the scaling $\tau \to \tau/p$. While this reduces some of the effect of p shown in Fig. 3, it still leads to decaying curves that can, for a given τ, be deceptively dependent on smoothing. Here we propose the key transformation $\tau \to \tau/2p(p-1)$, 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 derive in Eq. (A3) later in the Appendix. Its striking impact on the spread of curves from Fig. 3 is given in Fig. A3.

Further Model Selection Bounds

In the the main text we derived lower bounds on Ω^2, which led to the model rejection parameter, p^* (see Eq. (14)). Here we extend and support those results. In Fig. A4 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. A5 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

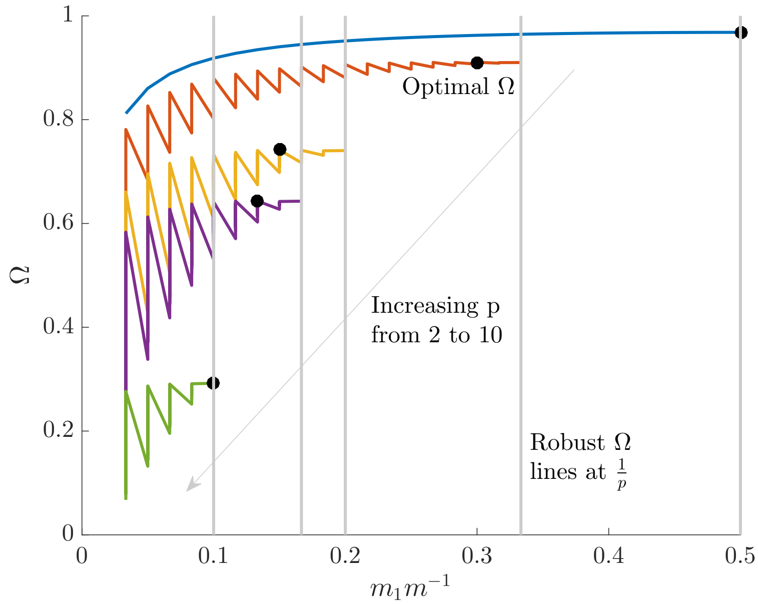

Fig. A1: **Robust and Ω optimal designs.** For the GMRF smoothing prior with $\delta_j = 1$ for all j and $\tau = 1$, we show that the optimal Ω design point is not always the same as the robust design point, at which $\frac{m_1}{m} = \frac{1}{p}$. The coloured Ω curves are (along the dashed arrow) for $p = [2, 3, 5, 6, 10]$ 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.

Eq. (A2), with $q = {}^{\mathrm{tr}[\boldsymbol{\mathcal{P}}]}/_{m}$ as follows

$$\Delta\mathbb{I} \leqslant \frac{1}{2} p \log\left(1+q\right) \leqslant \frac{1}{2} pq. \tag{A2}$$

This expression reveals that p is akin to a signal bandwidth (by comparison with standard Shannon-Hartley theory (Cover and Thomas, 2006)) and is therefore a key controlling factor in defining how much additional information the prior will introduce. This supports our proposed p^* rejection criterion.

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 (Huang and Zhang, 2018), which states that $\Delta\mathbb{I} \leqslant {}^{\zeta}/_{2}$, with $\zeta = \mathrm{tr}[\boldsymbol{\mathcal{I}}^{-\frac{1}{2}}\boldsymbol{\mathcal{P}}\boldsymbol{\mathcal{I}}^{-\frac{1}{2}}]$. At the robust point, we get $\zeta = \mathrm{tr}[\boldsymbol{\mathcal{I}}^{-1}\boldsymbol{\mathcal{P}}]$, which leads to the second inequality in Eq. (A2). Thus, our bound is tighter than that in (Huang and Zhang, 2018), and useful for broader, future mathematical analyses of $\Delta\mathbb{I}$. This

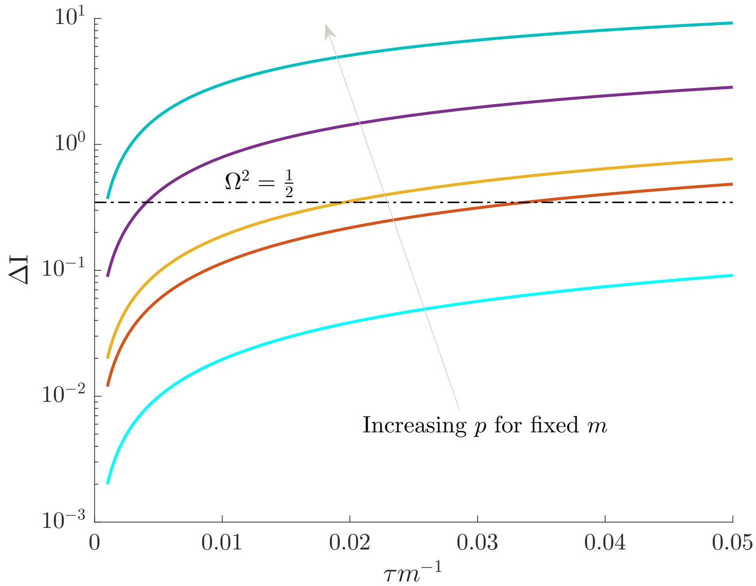

Fig. A2: **Prior mutual information increases with skyline complexity.** For the uniform GMRF, we show that under fixed smoothing (and hence τ/m), the additional mutual information introduced by the prior, $\Delta\mathbb{I} = \mathbb{E}_0[-\log \Omega]$, significantly increases with the complexity, p, 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). 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.

inequality also clarifies why m/p is often important for characterising performance here.

We can also use the bound of (Huang and Zhang, 2018) to derive alternate (but slacker) lower bounds on Ω^2. This gives the first inequality in Eq. (A3). Applying this to the uniform GMRF gives the second inequality:

$$\Omega^2 \geqslant e^{-pq} \implies \Omega^2 \geqslant e^{-\frac{2}{m}p(p-1)\tau}. \tag{A3}$$

Interestingly, Eq. (A3) shows that the dependence of Ω^2 on the smoothing 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 (high-dimensional) skyline plots.

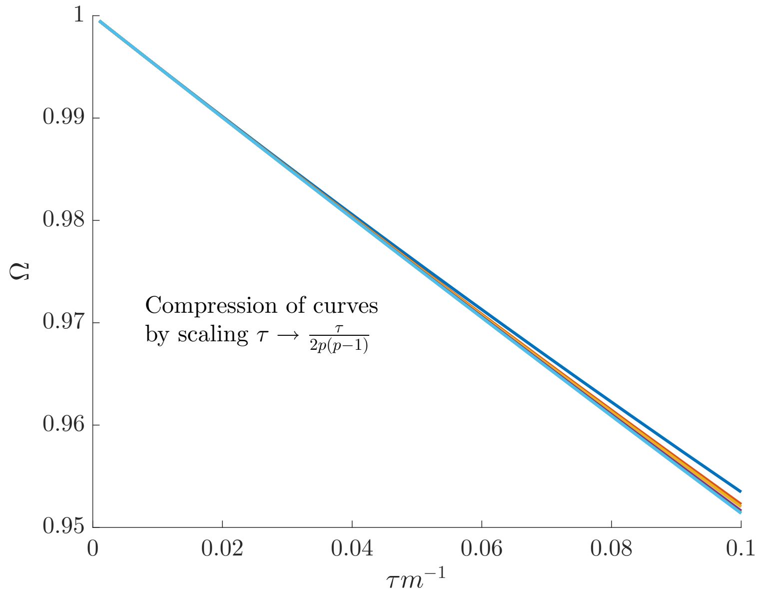

Fig. A3: **Negating the impact of skyline dimension.** We show how an appropriate quadratic scaling of the GMRF precision parameter, τ, can remove the complexity (p) induced smoothing contribution portrayed in Fig. 3 of the main text. This scaling significantly compresses the coloured Ω curves shown, which are for $p = [2, 4, 5, 10, 20]$ at $m = 100$ with $m_j = {}^{m}/_{p}$ (robust design point). The resulting Ω^2 values are now all comfortably above the $^1/_2$ threshold and justified by our information theoretic metrics.

Ancillary Uncertainty Statistics

In the Egyptian-HCV simulated example we defined two 95% HPD based ancillary statistics for characterising the visual uncertainty present in a skyline plot demographic estimate. In Fig. A6 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. A6 in the main text but observe here that trends between the more common (and more easily visualised) HPD based measures and our novel statistic are largely consistent.

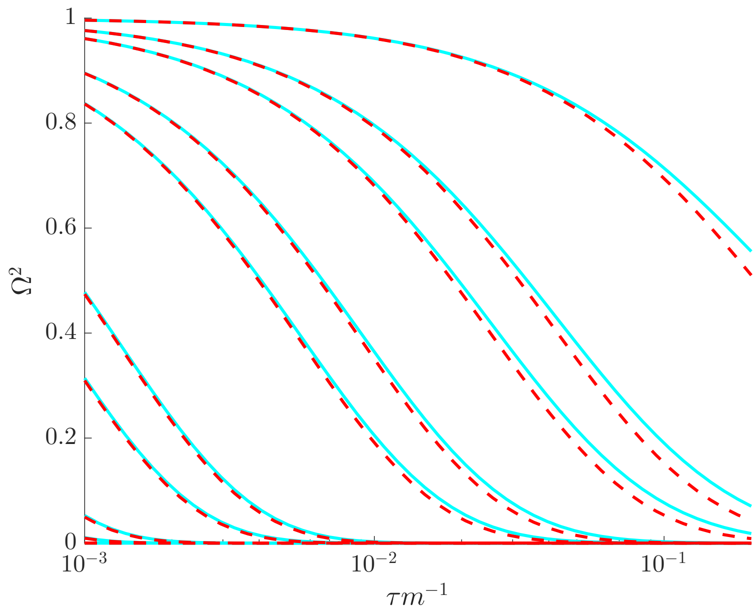

Fig. A4: **Lower bounds on Ω^2.** For the GMRF smoothing prior with $\delta_j = 1$ for all j and $m = 200$, we compare the lower bound on Ω^2 (red, dashed, see Eq. (14)) with the actual value of Ω^2 (cyan) at the robust design point of $m_j = {}^m\!/_p$. We examine all integer p values that are factors of m, and find that qualitatively similar comparisons hold for different τ and m settings. In general the lower bound (ω^2) is a good approximation to Ω^2.

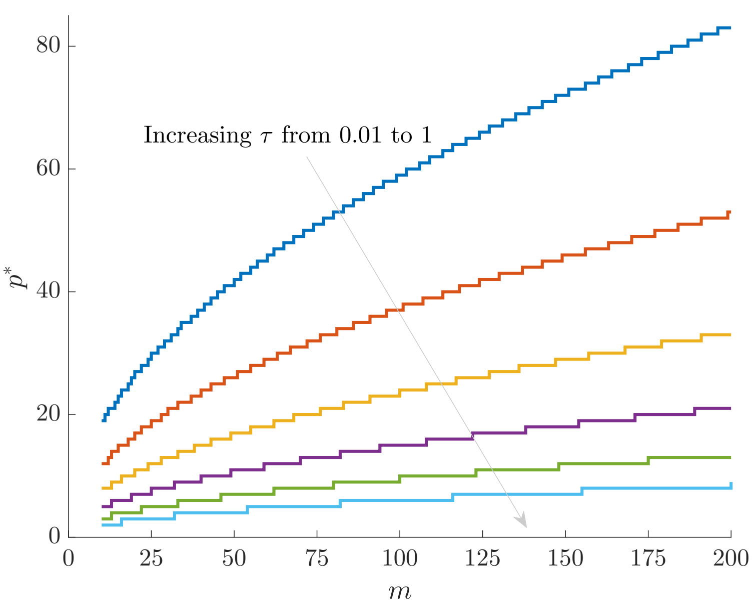

Fig. A5: **Maximum p model selection boundary.** For the GMRF smoothing prior with $\delta_j = 1$ for all j and at the robust point $m_j = {}^m\!/_p$, we compute the maximum allowed number of skyline segments, p^*, such that $\Omega^2 \geqslant {}^1\!/_2$. These curves increase with m and decrease with 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.

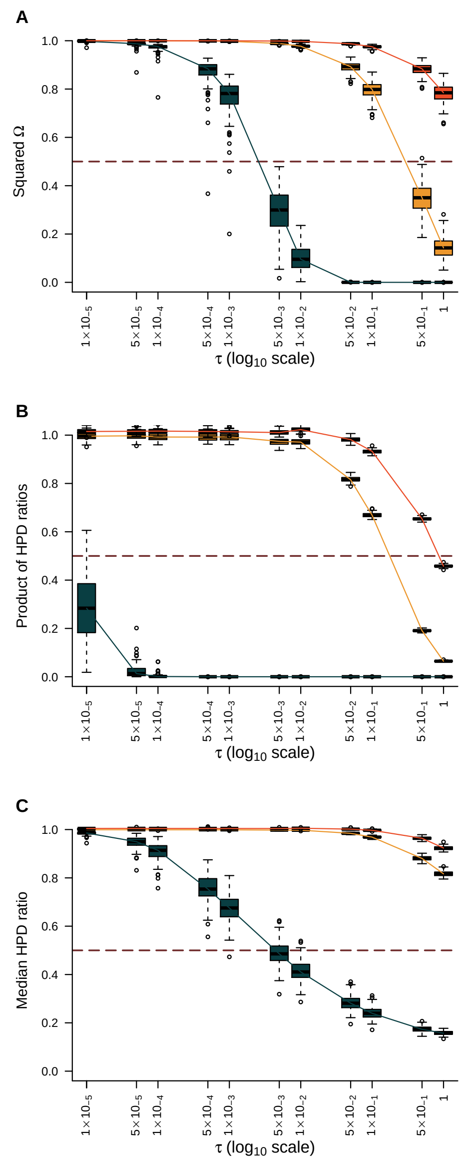

Fig. A6: **Trends in HPD-based statistics and Ω^2 under various time-aware GMRF settings.** The Ω^2 (panel A), median HPD ratio of $\log N_j$ (panel B) and HPD product (panel C) statistics are computed across $\log N_j$ over various combinations of m_j and τ. Box-plots summarise our results over 100 observed coalescent trees simulated from previously inferred demographic trends found for the Egyptian HCV dataset. Analyses with $m_j = 1$ are in dark green, $m_j = 4$ in yellow and $m_j = 8$ in orange. The solid lines link the median values across boxes for a given m_j value. The dashed line is positioned at the threshold $\Omega^2 = 1/2$.