Adaptive Estimation for Epidemic Renewal and Phylogenetic Skyline Models

KV Parag[*, 1] and CA Donnelly[1, 2]

[1]MRC Centre for Global Infectious Disease Analysis, Imperial College London, London, W2 1PG, UK
[2]Department of Statistics, University of Oxford, Oxford, OX1 3LB, UK
[*]Email: k.parag@imperial.ac.uk

Abstract—**The renewal model uses the observed incidence across an epidemic to estimate its underlying time-varying effective reproductive number, *R(t)*. The skyline model infers the time-varying effective population size, *N(t)*, responsible for the shape of an observed phylogeny of sequences sampled from an infected population. While both models solve different epidemiological problems, the bias and precision of their estimates depend on *p*-dimensional piecewise-constant descriptions of their variables of interest. At large *p* estimates can detect rapid changes but are noisy, while at small *p* inference, though precise, lacks temporal resolution. Surprisingly, no transparent, principled approach for optimally selecting *p*, for either model, exists. Usually, *p* is set heuristically, or obscurely controlled using complex algorithms. We present an easily computable and interpretable method for choosing *p* based on the minimum description length (MDL) formalism of information theory. Unlike many standard model selection techniques, MDL accounts for the additional statistical complexity induced by how parameters interact. As a result, our method optimises *p* so that *R(t)* and *N(t)* estimates properly adapt to the available data. It also outperforms comparable Akaike and Bayesian information criteria over several model classification problems. Our approach requires some knowledge of the parameter space, and exposes the similarities between renewal and skyline models.**

Key words: coalescent processes, renewal models, skyline plots, minimum description length, epidemiology, phylodynamics.

I. Introduction

Sampled phylogenies (or genealogies) and incidence curves (also known as epi-curves) are two related but distinct types of epidemiological data that are often used to learn about infectious epidemics. Phylogenies map the tree of ancestral relationships among sequences that were sampled from an infected population [5]. They provide a retrospective view of epidemic dynamics by allowing estimation of the effective size of that population. Incidence curves chart the number of infected individuals (infecteds) observed longitudinally across the epidemic [34], and provide ongoing insight into the rate of spread of that epidemic, by facilitating inference of its effective reproduction number. Minimalistic examples of each data type are given in the top-left and middle-right panels of Fig. 1.

The effective reproduction number, $R(t)$, is a key diagnostic of whether an epidemic is growing or under control. It defines how many secondary infections an infected will, on average, propagate [34]. The renewal model [6] [7] is a popular approach for inferring $R(t)$ from incidence curves, and has been used, for example, to predict Ebola virus disease case counts and assess the transmissibility of pandemic influenza [7] [3] [19]. The effective population size, $N(t)$, is essential for gauging how past fluctuations in the infected demography shaped epidemic spread. It measures the number of infecteds that contribute offspring (i.e. transmit the disease) to the next generation [10]. The skyline plot model [26] [5] is a prominent means of estimating $N(t)$ from phylogenies, and has provided insight into the historical growth and transmission of HIV and hepatitis C, among others [25] [15].

While renewal and skyline models depict very different aspects of an infectious disease, they possess some statistical similarities. Foremost is their approximation of $N(t)$ and $R(t)$ by p-dimensional piecewise-constant functions (see bottom panels of Fig. 1). Here p is the number of parameters to be inferred from the data and time t is regressive for phylogenies but progressive for incidence curves. The choice of p is critical to the quality of inference. Models with large p are better able to track rapid changes but are susceptible to noise and uncertainty [3]. Smaller p improves estimate precision but reduces flexibility, easily over-smoothing salient changes [16] [31]. Optimally selecting p, in a manner that is justified by the available data, is vital to deriving reliable conclusions from these models.

Surprisingly, no transparent, principled and easily computable p selection strategy exists. In renewal models, p is often set by trial and error, or defined using heuristic sliding windows [3] [6]. In skyline models, where this problem is more actively researched, smoothing priors or sophisticated algorithms are employed [16] [1]. The first assumes some type of autocorrelation between neighbouring piecewise parameters, while the second uses path sampling and power posterior distributions to solve more involved evolutionary model selection problems. In both cases p is implicitly or obscurely controlled, the rationale behind its choice is difficult to interrogate and the computational demand of the method is non-trivial [1] [22] [10]. Consequently, there is a need for new p selection metrics.

Here we attempt to answer this need by developing and validating a minimum description length (MDL)-based approach to renewal and skyline model selection. MDL is a formalism from information theory that treats model selection as equivalent to finding the best way of compressing observed data (i.e. its shortest description) [28]. MDL is advantageous because it includes both model dimensionality and parametric complexity within its definition of model complexity [29]. Parametric complexity describes how the functional relationship between parameters matters [17], and is usually ignored by standard selection criteria. However, in general MDL is intractable [8], which may explain why it has not penetrated the epidemiological literature.

We overcome this issue by deriving a simple Fisher information approximation (FIA) to MDL. This is achieved by recognising that sampled phylogenies and incidence curves both sit within a Poisson point process framework [30], and by capitalising on the piecewise-constant structure of skyline and renewal models [23]. The result is a pair of analogous FIA metrics that lead to adaptive estimates of $N(t)$ and $R(t)$ by selecting the p most justified by the observed Poisson data. Our FIA expressions decompose model complexity into clearly interpretable contributions and are as easy to compute as standard Akaike (AIC) and Bayesian information criteria (BIC). We find, over a range of selection problems, that the FIA generally outperforms the AIC and BIC, emphasizing the importance of including parametric complexity. This improvement comes at the expense of requiring some knowledge about the piecewise parameter space domain.

II. Materials and Methods

A. Phylogenetic Skyline and Epidemic Renewal Models

The phylogenetic skyline and epidemic renewal models are two distinct and popular approaches to solving epidemiological inference problems. The skyline model [26] [5] infers the hidden time-varying effective population size, $N(t)$, from a phylogeny of sequences sampled from that infected population; while the renewal model [6] [7] estimates the hidden time-varying effective reproduction number, $R(t)$, from the observed incidence of an infectious disease. Here t indicates time, which is progressive (moving from past to present) in the renewal model, but reversed (retrospective) in the skyline model. Although both models solve different problems, they approximate their variable of interest, $\theta(t)$, with a p-dimensional piecewise-constant function, and assume a Poisson point process relationship between it and the observed data, $y(t)$, as in Eq. (1).

$$\theta(t) = \sum_{j=1}^{p} \theta_j 1(\epsilon_{j-1} \leq t < \epsilon_j), \quad y(t) \sim \text{Poiss}\left(\mu(\theta(t))\right) \tag{1}$$

Here $\theta(t)$ is either $N(t)$ or $R(t)$ and $y(t)$ is either phylogenetic or incidence data, depending on the model of interest. The j^{th} piecewise component of $\theta(t)$, which is valid over the interval $[\epsilon_{j-1}, \epsilon_j)$, is θ_j. The rate function, $\mu(\theta(t))$ (or $\mu(t)$ for short), allows us to treat these usually distinct models within the same Poisson point process framework, where $\mathbb{P}(y(t) - y(0) = n) = {}^1\!/_{n!}\,(\int_0^t \mu(s)\,\mathrm{d}s)^n e^{-\int_0^t \mu(s)\,\mathrm{d}s}$. This implies that the time between y events $\sim \exp(\mu(\theta(t)))$. We want to estimate the parameter vector $\theta = [\theta_1, \ldots, \theta_p]$ from the observed data over $0 \leq t \leq T$, denoted y_0^T.

The skyline model is based on the coalescent approach to phylogenetics [13]. Here sampled genetic sequences (lineages) from an infected population elicit a reconstructed phylogeny, in which these lineages successively merge into their common ancestor. The observed branching or coalescent times of this phylogeny form a Poisson point process that contains information about the piecewise population parameters $N := [N_1, \ldots, N_p]$. This data type is known as a Poisson count record [30], and the key Eq. (1) relation between the observed data and the inferred parameters is $\mu(t) = \binom{n(t)}{2} N(t)^{-1}$ with $n(t)$ as the number of lineages in the phylogeny at time t. The left panels of Fig. 1 illustrate the skyline inference problem.

Let the count record of the coalescent event times over $0 \leq t \leq T$ be C_0^T. The log-likelihood, $l_p(N) = \log \mathbb{P}(C_0^T \mid N)$, which statistically describes how this count record informs on the p-dimensional $N(t)$ then follows as Eq. (2) [5] [23].

$$l_p(N) = \Gamma + \sum_{j=1}^{p} m_j \log \frac{1}{N_j} - \frac{\omega_j}{N_j} \tag{2}$$

Here m_j counts the number of coalescent event times within the duration of N_j (i.e. $[\epsilon_{j-1}, \epsilon_j)$), while $\omega_j = \int_{\epsilon_{j-1}}^{\epsilon_j} \binom{n(t)}{2}\, dt$ and $\Gamma = \sum_{i=1}^{m} \log \binom{n(c_i)}{2}$ are constants for a given phylogeny, with c_i as the i^{th} coalescent event time. Since $N(t)$ can have a large dynamic range (e.g. for exponentially growing epidemics) it is necessary to perform estimation under the robust log transform [23]. Note that the skyline models we use have piecewise segment change-point times, ϵ_j, that coincide with coalescent event times, as in [26] [5] [20].

We obtain the maximum likelihood estimate (MLE), $\log \hat{N}_j$, and Fisher information (FI), $\mathcal{I}(\log N_j)$ of this model by solving $\nabla_N l_p(N) = 0$ and $\mathbb{E}[-\nabla_N^2 l_p(N)]$ with $\nabla_N := \{{}^{\partial}\!/_{\partial N_j}\}$ as the vector derivative operator [14]. Log-transforming gives Eq. (3) [23].

$$\log \hat{N}_j = \log \omega_j - \log m_j, \quad \mathcal{I}(\log N_j) = m_j \tag{3}$$

Note that across the whole count record there are m coalescent events with $\sum_{j=1}^{p} m_j = m$. The MLE and FI are important measures for describing the population size estimates given m [14]. For a given p, the MLE controls the per segment bias because as m_j increases $\log N_j - \log \hat{N}_j$ decreases. The FI defines the precision i.e. the inverse of the variance around the MLEs, and also (directly) improves with m_j. We will find these two quantities to be vital, when formulating our approach to p-model selection. Thus, the FI and MLE control the per segment performance, while p determines how well the overall piecewise function adapts to the underlying generating process.

The renewal model is based on the classic renewal equation approach to epidemic transmission [33]. This states that the number of new infecteds depends on past incidence through the generation time distribution, and the effective reproduction number. As incidence is usually observed on a coarse time scale (e.g. days or weeks), exact infection times are not available. As a result, time is binned or discretised, with the number of infecteds observed in the t^{th} bin denoted $I(t)$. For simplicity we assume daily bins. The generation time distribution is specified by $w(s)$, the probability that an infected takes between $s-1$ and s days to transmit that infection [6].

The total infectiousness of the disease is defined by $\Lambda(t) := \sum_{s=1}^{t-1} I(t-s)w(s)$, with $\sum_{s=1}^{c} w(s) = 1$ for a generation time with maximum memory of c. We make the common assumptions that $w(s)$ is known (it is disease specific) and stationary (does not change with time) [3]. If an epidemic is observed for $T = m > c$ days then the historical incidence, I_1^T, informs on the piecewise parameters to be estimated, $R = [R_1, \ldots, R_p]$. This contrasts the skyline model, as information is now available from binned sums of events instead actual event times. This type of data is known as a histogram record [30]. The right panels of Fig. 1 explain the renewal inference problem and the relation between histogram and count record data.

We derive the renewal log-likelihood using Eq. (1), by noting that the log-likelihood of the relevant binned Poisson point process is $-\int_0^m \mu(u))\,\mathrm{d}u + \sum_{t=1}^{m} I(t) \log \left(\int_{u_{t-1}}^{u_t} \mu(u)\,\mathrm{d}u\right) - I(t)!$ [30]. Here u is continuous time and $u_t - u_{t-1} = 1$ defines the endpoints of the t^{th} day ($u_0 = 0$). The renewal equation asserts that $\mathbb{E}[I(t)] = \Lambda(t)R(t)$ [6]. Setting this equal to our binned Poisson mean gives $\int_{u_{t-1}}^{u_t} \mu(u)\,\mathrm{d}u = \Lambda(t)R(t)$ and recovers the standard renewal log-likelihood. Note that this is commonly derived by simply assuming discrete Poisson noise around $\Lambda(t)R(t)$ [7]. Our alternate derivation exposes the statistical similarities between renewal and skyline models and allows generalisation of standard renewal approaches to variable width histogram records (e.g. irregularly timed epi-curves) by choosing appropriate bin endpoints $[u_{j-1}, u_j)$.

The above log-likelihood is for the maximally flexible m-parameter renewal model. For a general p-dimensional model ($p \leq m$) the log-likelihood, $l_p(R) = \log \mathbb{P}(I_1^T \mid R)$, follows by grouping terms from the m-parameter case. This leads to Eq. (4), with constant $\Gamma = \sum_{t=1}^{m} -\log I(t)! + I(t) \log \Lambda(t)$ invariant to all p-groupings.

$$l_p(R) = \Gamma + \sum_{j=1}^{p} i_j \log R_j - \lambda_j R_j \tag{4}$$

If the j^{th} group contains m_j days or bins, depicted by the set $\kappa^{(j)} = \{m_{j-1}+1, \ldots, m_{j-1}+m_j\}$ with duration $[\epsilon_{j-1}, \epsilon_j)$, then we can define grouped sums $\lambda_j := \sum_{t \in \kappa^{(j)}} \Lambda(t)$, $i_j := \sum_{t \in \kappa^{(j)}} I(t)$. This leads to the MLE and FI of Eq. (5) with $\sum_{j=1}^{p} m_j = m$ [7] [23].

$$\hat{R}_j = i_j \lambda_j^{-1}, \qquad \mathcal{I}(2\sqrt{R_j}) = \lambda_j \tag{5}$$

As each m_j becomes large the per segment bias $R_j - \hat{R}_j$ will decrease. Using results from [23], we can show that the square root of the reproduction number is the most robust parametrisation for standard renewal models. We compute the FI under this parametrisation to reveal that the total infectiousness controls the precision around our MLEs. This will also improve as m_j increases, but with the caveat that the parameters underlying bigger epidemics (specified by larger historical incidence values, and controlled through $\Lambda(t)$) are easier to estimate than those of smaller ones.

In both models we find a clear piecewise separation of MLEs and FIs. Per segment bias and precision depend on the quantity of data apportioned to each parameter. This data division is controlled by p, which balances per segment performance against the overall fit of the model to its generating process. Thus, model dimensionality fundamentally controls inference quality. Large p means more segments, which can adapt to rapid $N(t)$ or $R(t)$ changes. However, this also rarefies the per segment data (grouped sums like λ_j or m_j decrease) with both models becoming unidentifiable if $p > m$. Small p improves segment inference, but stiffens the model. We next explore information theoretic approaches to p-selection that formally utilise both MLEs and FIs within their decision making algorithms.

B. Model and Parametric Complexity

Our proposed approach to model selection relies on the MDL framework of [28]. This treats modelling as an attempt to compress

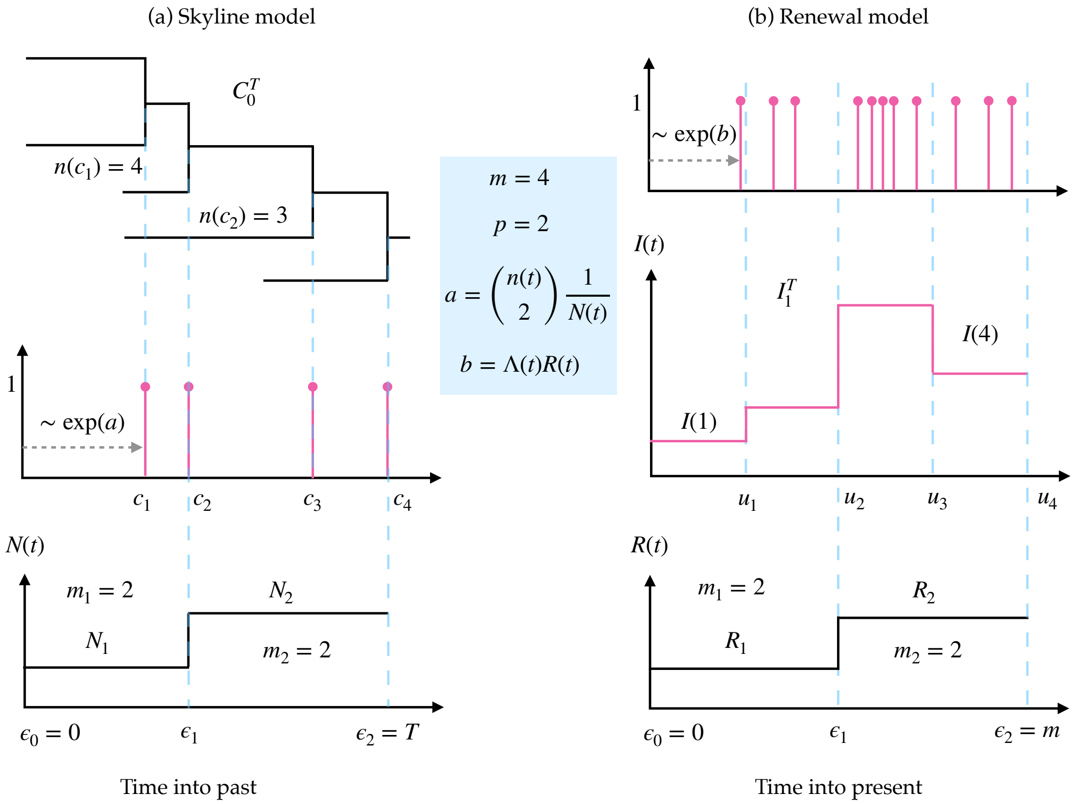

Fig. 1: **Poisson skyline and renewal inference problems.** The left panels show how the reconstructed phylogeny of infecteds (top) leads to (branching) coalescent events, which form a Poisson count record (middle). The timing of these observable events encodes information about the piecewise effective population size function to be inferred (bottom). The right panels indicate how infecteds, which naturally conform to a Poisson count record (top) are usually only observed at the resolution of days or weeks, leading to a Poisson histogram record (middle). The number of infecteds falling in any histogram bin is informative of the piecewise effective reproduction number (bottom). Both models feature data with $m = 4$ and involve $p = 2$ parameters to be estimated. The notation used is explained in Materials and Methods.

the regularities in the observed data, which is equivalent to learning about its statistical structure. MDL evaluates a p-parameter model, $\mathcal{M}_p$, in terms of its code length, $M_p = \phi(\mathcal{M}_p) + \phi(\mathcal{D} \,|\, \mathcal{M}_p)$ [8]. Here $\phi(x)$ computes the length to encode x (in e.g. nats or bits) and $\mathcal{D}$ is the observed data. M_p is the sum of the information required to describe $\mathcal{M}_p$ and the data given that $\mathcal{M}_p$ is chosen. More complex models have larger $\phi(\mathcal{M}_p)$ (more bits are needed to depict just the model), and smaller $\phi(\mathcal{D} \,|\, \mathcal{M}_p)$ (as complex models should better fit the data, there is less remaining information to detail). If n models are used to describe $\mathcal{D}$ then the model with $p^* = \arg\min_{1 \leq p \leq n} M_p$ best compresses or most succinctly represents the data.

The model with p^* is known to possess the desirable properties of generalisability and consistency [8]. The first means that $\mathcal{M}_{p^*}$ provides good predictions on newly observed data (i.e. it fits the underlying data generating process instead of a specific instance of data derived from that process), while the second indicates that the selected p^* will converge to the true model index (if one exists) as data increase [24] [8] [2]. However, in its exact form, MDL is very difficult to compute. We therefore use its well known FIA, from [29], which we denote FIA_p for $\mathcal{M}_p$ in Eq. (6).

$$\text{FIA}_p = -l_p(\hat{\theta}) + \frac{p}{2} \log \frac{m}{2\pi} + \log \int \det \left[m^{-1} \mathcal{I}(\theta) \right]^{\frac{1}{2}} \, \mathrm{d}\theta \qquad (6)$$

Here 'det' is the standard matrix determinant. The approximation of Eq. (6) is good, provided certain regularity conditions are met. These mostly relate to the FI being identifiable and continuous in θ, and are not issues for either skyline or renewal models [17]. While we will apply the FIA within a class of renewal or skyline models, this restriction is unnecessary. The FIA can be used to select among any variously parametrised and non-nested models [8].

Eq. (6) explicates the main advantages of our approach: interpretability, completeness and computability. The FIA assesses model complexity as the number of distinguishable distributions that a model can portray [8]. This is done via the last two terms in Eq. (6) (the first term is part of most model selection criteria and defines model fit). The $\frac{p}{2} \log \frac{m}{2\pi}$ term states that complexity increases with both the number of parameters and the data dimension. Higher p indicates a more flexible model, while larger m also increases the number of describable distributions by improving the resolution of inference. Standard and easily computable metrics, such as the AIC and BIC also, to varying extents, account for these (p, m) effects [12].

However, most standard metrics do not have an equivalent to the integral term in Eq. (6), which depicts parametric complexity [29] [17]. Parametric complexity measures the contribution of the functional form of a model to its overall complexity. It explains why two parameter sinusoidal and exponential models have nonidentical complexities, for example (the different functional relationships between the parameters lead to distinct sets of distinguishable distributions). This concept is detailed in [24] and [8], and is an important but often neglected aspect of model complexity. In general, the integral term can be intractable [29]. However, in Section III-B we show that the FIA for renewal and skyline models, as a consequence of their piecewise separable MLEs and FIs (Eq. (3) and Eq. (5)), is no more difficult to compute than the AIC or BIC.

This term therefore allows the FIA to more comprehensively assess model complexity than the AIC, BIC and other standard metrics, without any computational downsides. More sophisticated techniques such as Bayesian model selection (BMS) do account for parametric complexity, but at the expense of tractability, subjectivity (prior choices) and interpretability [12]. The FIA decomposes complexity into transparent and commensurable contributors, allowing relative determination of the main sources of model complexity. For example, when m becomes large the parametric complexity is comparatively unimportant and the FIA converges to the BIC [29]. Moreover, the FIA can approach the performance of these more sophisticated techniques without the computational cost. If an uninformative (Jeffrey's) prior is used, the FIA selects the same model as BMS [17] [8].

Importantly, model complexity is clearly not the same as model dimensionality. While parametric complexity is invariant to parameter transforms and independent of sample size and model fit, it does require knowledge of the parameter domain (the integral limits) [8]. In this work we will generally assume some arbitrary but sensible domain. However, when this is not possible another approximation to MDL, denoted QK_p and given in Eq. (7), can be used.

$$\text{QK}_p = -l_p(\hat{\theta}) + \frac{1}{2}\log\det[\mathcal{I}(\hat{\theta})] + \sum_{j=1}^{p}\log(|\,\hat{\theta}_j\,| + m^{-\frac{1}{4}}) \tag{7}$$

This Qian-Kunsch (QK) approximation was developed in [27], and trades some interpretability for the benefit of not having to demarcate the multidimensional domain of integration.

III. Results

A. The Insufficiency of Log-Likelihoods

The inference performance of both the renewal and skyline models, for a given dataset, strongly depends on the model dimensionality, p, that is chosen. As observed in Section I, current approaches to p-selection utilise ad-hoc rules or elaborate algorithms that are difficult to interrogate. Here we emphasise why finding an optimal p, denoted p^*, is important and illustrate the downfalls of inadequately balancing bias and precision. We start by proving that over-fitting and over-parametrisation are guaranteed consequences of depending solely on the log-likelihood for p-selection. We remove terms from Eq. (2) and Eq. (4) that are invariant to model dimensionality, and substitute the MLEs of Eq. (3) and Eq. (5) to obtain Eq. (8).

$$l_p(\hat{N}) = \sum_{j=1}^{p} m_j \log\frac{m_j}{\omega_j}, \quad l_p(\hat{R}) = \sum_{j=1}^{p} i_j \log\frac{i_j}{\lambda_j} \tag{8}$$

Both the renewal and skyline log-likelihoods take the form $l_p(\hat{\theta}) = \sum_{j=1}^{p} a_j \log\frac{a_j}{b_j}$, due to their inherent and dominant Poisson, piecewise constant structure. Here a_j and b_j as grouped variables that are directly computable from the observed data (C_0^T or I_1^T depending on the model). The most complex model supportable by the data is at $p = m$, with $l_m(\hat{\theta}) = \sum_{i=1}^{m} a_i \log\frac{a_i}{b_i}$. As the data size ($m$) is fixed, and if $\kappa^{(j)}$ is the set of i indices that must be clumped to obtain the j^{th} grouping then $a_j = \sum_{i\in\kappa^{(j)}} a_i$ and $b_j = \sum_{i\in\kappa^{(j)}} b_i$. The log-sum inequality from [4] states that $\sum_{i\in\kappa^{(j)}} a_i \log\frac{a_i}{b_i} \geq \left(\sum_{i\in\kappa^{(j)}} a_i\right) \log\left(\sum_{i\in\kappa^{(j)}} a_i\right)/\left(\sum_{i\in\kappa^{(j)}} b_i\right)$. Repeating this across all possible p groupings results in Eq. (9).

$$p^* = \min_{1\leq p\leq m} -l_p(\hat{\theta}) = m, \text{ for } \hat{\theta} = \hat{N} \text{ or } \hat{R} \tag{9}$$

Consequently, log-likelihood based model selection always chooses the highest dimensional renewal or skyline model. This result also holds when solving Eq. (9) over a subset of all possible p, provided smaller p models are obtained by taking non-overlapping groupings of larger p ones [9]. It is therefore necessary to penalise the log-likelihood with some term that increases with p.

The highest p model is most sensitive to changes in $\theta(t)$, but extremely noisy and likely to overfit the data. This noise is reflected in a poor FI. From Eq. (3) and Eq. (5) it is clear that grouping linearly increases the FI, hence smoothing noise. However, this improved precision comes with lower flexibility. At the extreme of $p = 1$, for example, $\theta(t)$ is approximated by a single, perennial parameter, and the log-likelihood $l_1(\hat{\theta}) = \left(\sum_{i=1}^{m} a_i\right) \log\left(\sum_{i=1}^{m} a_i\right)/\left(\sum_{i=1}^{m} b_i\right)$ is unchanged for all combinations of data that produce the same grouped sums. This oversmooths and underfits. We will always select $p^* = 1$ if our log-likelihood penalty is too sensitive to dimensionality.

Some concrete examples of bad model selection are now presented. Here we use deterministic groupings of size k to control p i.e. every $\kappa^{(j)}$ clumps k successive indices (the last index is m). In Fig. 2 we examine skyline models with periodic exponential fluctuations (top panels) and bottleneck variations (bottom panels), in $\log N(t)$. The periodic case describes seasonal epidemic oscillations in infecteds, while the bottleneck simulates the severe decline that results from a catastrophic event. In Fig. 3 we investigate renewal models featuring cyclical (top panels) and sigmoidal (bottom panels) $R(t)$ dynamics. The cyclical model represents the pattern of spread for a seasonal epidemic (e.g. influenza), while the sigmoidal one simulates a vaccination policy that quickly leads to outbreak control.

In both figures we observe underfitting at low p (left panels) and overfitting at high p (right panels). The detrimental effects of choosing the wrong model are not only dramatic, but also realistic. For example, in the skyline examples the underfitted case corresponds to the fundamental Kingman coalescent model [13], which is often used as a null model in phylogenetics. Alternatively, the classic skyline [26], which is at the core of many coalescent inference algorithms, is exactly as noisy as the overfitted case. Correctly penalising the log-likelihood is therefore essential for good estimation. Section III-B investigates several appropriate model selection penalties.

B. Minimum Description Length Selection

Having clarified the impact of non-adaptive estimation in Eq. (9), we develop and appraise various model selection metrics, in terms of how they penalise renewal and skyline log-likelihoods. The most common and readily computed metrics are the AIC and BIC [8] [12], which we reformulate in Eq. (10) and Eq. (11), with $(a_j, b_j) = (m_j, \omega_j)$ or (i_j, λ_j) for skyline and renewal models respectively.

$$\text{AIC}_p = \sum_{j=1}^{p} -a_j \log\frac{a_j}{b_j} + 1 \tag{10}$$

$$\text{BIC}_p = \sum_{j=1}^{p} -a_j \log\frac{a_j}{b_j} + \frac{1}{2}\log m \tag{11}$$

By decomposing the AIC and BIC on a per segment basis (for a model with p segments or dimensions), as in Eq. (10) and Eq. (11), we gain insight into exactly how they penalise the log-likelihood. Specifically, the AIC simply treats model dimensionality as a proxy for complexity, while the BIC also factors in the total dimension of the available data. Note that a small sample correction to the AIC, which adds a further ${}^{p+1}/_{m-p-1}$ to the penalty in Eq. (10), was used

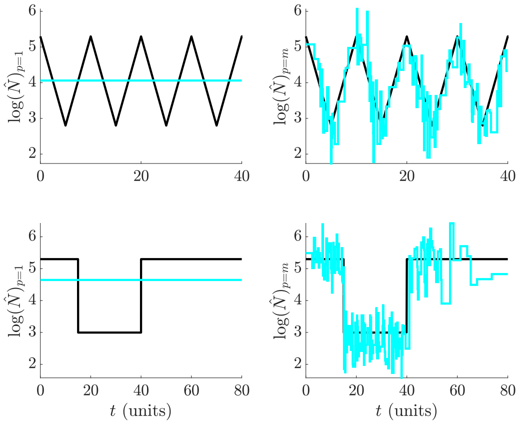

Fig. 2: **Skyline model under and overfitting.** Small p (large k) leads to smooth but biased estimates characteristic of underfitting (left panels). Large p (small k) results in noisy estimates that respond well to changes. This is symptomatic of overfitting (right panels). The MLEs $\log \hat{N}$ are in cyan, the true $\log N(t)$ in black, and $m = 800$.

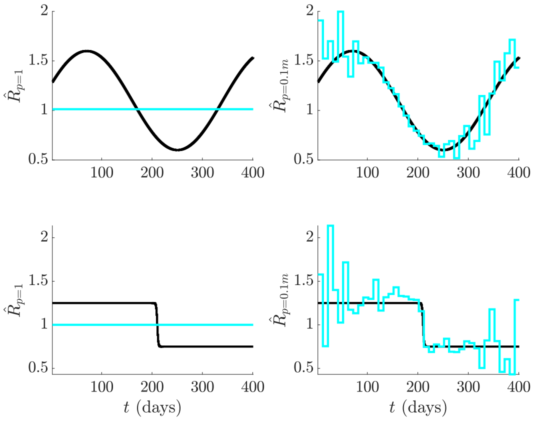

Fig. 3: **Renewal model under and overfitting.** Small p (large k) results in precise but underfitted and inflexible estimates (left panels). Large p (small k) leads to flexible, overfitted inferences with poor stability (right panels). The MLEs $\hat{R}$ are in cyan, the true $R(t)$ is in black and simulations are at $m = T = 400$.

in [31] for skyline models. We found this correction inconsequential to our later simulations and so used the standard AIC only.

As discussed in Section II-B, these metrics are insufficient descriptions because they ignore parametric complexity[1]. Consequently, we proposed the MDL approaches of Eq. (6) and Eq. (7). We now derive and specialise these expressions to skyline and renewal models. Adapting the FIA metric of Eq. (6) forms a key result of this work. Its integral term, $\Omega = \log \int \det[m^{-1}\mathcal{I}(\theta)]^{\frac{1}{2}}$, can, in general, be intractable [29]. However, the piecewise structure of both skyline and renewal models, which leads to orthogonal (diagonal) FI matrices, allows us to decompose $\det[m^{-1}\mathcal{I}(\theta)]^{\frac{1}{2}}$ as $\prod_{j=1}^{p} \sqrt{m^{-1}\mathcal{I}_j(\theta)}$ with $\mathcal{I}_j(\theta)$ as j^{th} diagonal element of $\mathcal{I}(\theta)$. Note that $\theta = N$ or R for the skyline and renewal model respectively.

Using this decomposition we can partition Ω across each piecewise segment as $-\frac{p}{2}\log m + \log \prod_{j=1}^{p} \int \mathcal{I}_j(\theta)^{\frac{1}{2}}\, \mathrm{d}\theta_j$. The $\int \mathcal{I}_j(\theta)^{\frac{1}{2}}\, \mathrm{d}\theta_j$ is invariant to parameter transformations [8]. Let η_j denote the robust transform of $\log \theta_j$ or $2\sqrt{\theta_j}$ for the skyline or renewal model, respectively [23]. Applying the FI change of variable formula [14] gives $\int \mathcal{I}_j(\theta)^{\frac{1}{2}}\, \mathrm{d}\theta_j = \int \mathcal{I}_j(\eta)^{\frac{1}{2}}\, \mathrm{d}\eta_j = \mathcal{I}_j(\eta)^{\frac{1}{2}} \int 1\, \mathrm{d}\eta_j$. The last equality follows by substituting the robust FI values from Eq. (3) ($\mathcal{I}_j(\eta) = m_j$) and Eq. (5) ($\mathcal{I}_j(\eta) = \lambda_j$). Consequently, $\Omega = -\frac{p}{2}\log m + \sum_{j=1}^{p} \frac{1}{2}\log \mathcal{I}_j(\eta) + \log \int 1\, \mathrm{d}\eta_j$. The domain of integration for each parameter is all that remains to be solved.

We make the reasonable assumption that each piecewise parameter, θ_j, has an identical domain. This is $N_j \in [1, v]$ and $R_j \in [0, v]$, with v as an unknown model-dependent maximum. The minima of 1 and 0 are sensible for these models. This gives $\int 1\, \mathrm{d}\eta_j = \log v$ or $2\sqrt{v}$ for the skyline or renewal model. Substituting into Ω and then Eq. (6) leads, after some algebra, to Eq. (12) and Eq. (13) below.

$$\mathrm{FIA}_p = \sum_{j=1}^{p} -m_j \log \frac{m_j}{\omega_j} + \frac{1}{2}\log m_j + \frac{1}{2}\log \frac{(\log v)^2}{2\pi} \tag{12}$$

$$\mathrm{FIA}_p = \sum_{j=1}^{p} -i_j \log \frac{i_j}{\lambda_j} + \frac{1}{2}\log \lambda_j + \frac{1}{2}\log \frac{2v}{\pi} \tag{13}$$

Eq. (12) and Eq. (13) present an interesting and complete view of piecewise model complexity. When compared to the BIC (Eq. (11)) we see that the FIA accounts for how the data are divided among segments, making explicit use of the robust FI of each model. This is an improvement over simply using the (clumped) data dimension m. Intriguingly, the maximum value of each parameter to be inferred, v, is also central to computing model complexity. This makes sense as models with larger parameter spaces can describe more types of statistical behaviours [8]. By comparing the second and third terms of these expressions we can get an idea of the relative contribution of the data and parameter spaces to complexity.

One weakness of the FIA is its dependence on the unknown v, which must be assumed finite. The QK MDL approximation of [27] resolves this issue. We obtain the QK criteria by simply substituting appropriate FIs and MLEs from Section II-A into Eq. (7). Expressions identical to Eq. (12) and Eq. (13) result, except for the parameter space term, which is replaced as shown in Eq. (14) and Eq. (15).

$$\mathrm{QK}_p : \frac{1}{2}\log \frac{(\log v)^2}{2\pi} \longrightarrow \log\left(\log \frac{\omega_j}{m_j} + m^{-\frac{1}{4}}\right) \tag{14}$$

$$\mathrm{QK}_p : \frac{1}{2}\log \frac{2v}{\pi} \longrightarrow \log\left(\frac{i_j}{\lambda_j} + m^{-\frac{1}{4}}\right) + \frac{1}{2}\log \frac{\lambda_j}{i_j} \tag{15}$$

These replacements require no knowledge of the parameter domain, but still approximate the parametric complexity of the model [27]. However, in gaining this domain independence we lose some performance (see Section III-D and Section III-D), and transparency, relative to the FIA criteria. Importantly, observe that both the FIA and QK are as easy to compute as the AIC or BIC. The similarity in the skyline and renewal model expressions reflects the significance of their piecewise Poisson formulations. We will next investigate the practical performance of our MDL approaches.

C. Adaptive Estimation: Epidemic Renewal Models

We validate our FIA approach on several renewal inference problems. We simulate epidemic incidence curves, $I(t)$, under some true $R(t)$, and with a gamma generation time distribution, $w(t)$, that closely mimics that used in [19] for depicting Ebola outbreaks. Simulations are initialised with 10 infecteds, as in [3], and we condition on the epidemic not dying out, in addition to removing initial sequences of zero incidence at start-up. These stipulations ensure identifiable inference. We consider an observation period of $T = 400$ days, and select from among the set of models with

[1] In some contexts, such as when the data sample size is asymptotically large, parametric complexity can be neglected without detriment [8].

$10 \leq k \leq T$ such that T is divisible by k. Here k determines the number of days that are grouped to form a piecewise segment, and model dimensionality, p, is bijective in k i.e. $pk = T = m$.

We apply the criteria developed in Section III-B to select among possible p-parameter (or equivalently k-grouped) renewal models. For the FIA approach we set $v = 100$ as a seemingly appropriate upper bound on the reproduction number domain. We start by highlighting how the FIA (i) regulates between the over and under-fitting extremes from Fig. 3, and (ii) updates its selected p^* as the data increase. These points are illustrated in Fig. 4 and Fig. 5.

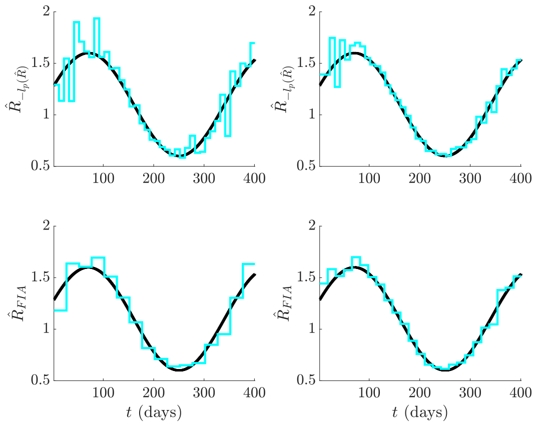

Fig. 4: **Adaptive cyclical estimation with FIA.** The top panels show the optimal log-likelihood based $R(t)$ MLEs for 1 (left) and 6 (right) incidence data streams simulated under cyclical reproduction numbers. The bottom panels provide the FIA adaptive estimates at the same settings with $v = 100$.

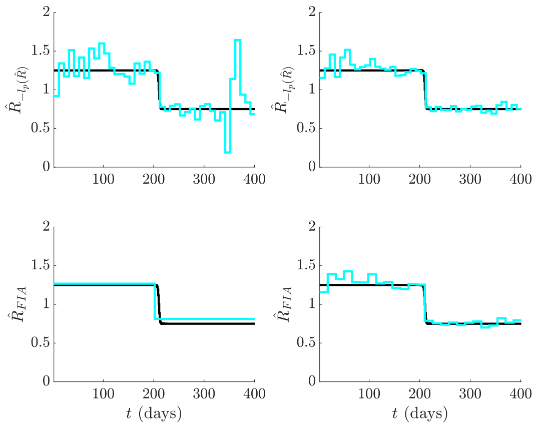

Fig. 5: **Adaptive sigmoidal estimation with FIA.** The top panels present optimal log-likelihood based $R(t)$ MLEs for 1 (left) and 6 (right) observed incidence data streams simulated under sharp, sigmoidal time-varying reproduction numbers. The bottom panels give the FIA adaptive estimates at the same settings with $v = 100$.

The left panels of both figures exemplify (i) as the FIA (bottom) reduces p from the maximum chosen by the log-likelihood (top), leading to trajectories that best balance noise against dimensionality. In particular, observe how the FIA chooses just two parameters to estimate the sigmoidal fall in Fig. 5, thus pinpointing its key characteristics. Interestingly, as the observed data are increased (right panels of Fig. 4 and Fig. 5) the FIA adapts its choice of p to reflect the improved resolution that is now justified by more data, hence demonstrating (ii). We obtained this increased dataset by appending 5 further independent incidence curves, conditional on $R(t)$.

While the above examples provide practical insight into the merits of the FIA, they cannot rigorously assess its performance, since non-piecewise $R(t)$ functions have no true $p = p^*$ or $k = k^* = {}^{T}\!/_{p^*}$. We therefore study two further problems in which a true p^* exists: a simple binary classification, and a more complex model search. In both, we benchmark the FIA against the AIC, BIC and QK criteria from Section III-B. While all subsequent simulations are performed under a fixed incidence curve, we note that, when $R(t)$ is piecewise-constant, increasing the number of conditionally independent curves improves the probability of selecting the true model.

For the first problem we set $T = 200$ days and use a constant null model (model 1) with $R(t) = 1.5$, to exemplify an uncontrolled epidemic. The alternative model 2 changes to $R(t \geq {}^{T}\!/_{2}) = 0.5$ to mimic the introduction of rapid control at ${}^{T}\!/_{2}$ (inset of Fig. 6). We randomly generate 10^3 epidemics with some null model probability and compute the frequentist probability that each criteria selects the correct model in Fig. 6. The FIA outperforms all other criteria, with the QK as its closest competitor. The AIC performs poorly, as does the log-likelihood (not shown), because they are biased towards the more complex model 2. Relative metric performance is unchanged if we instead set $R(t \geq {}^{T}\!/_{2}) = 2.5$ (an accelerating epidemic).

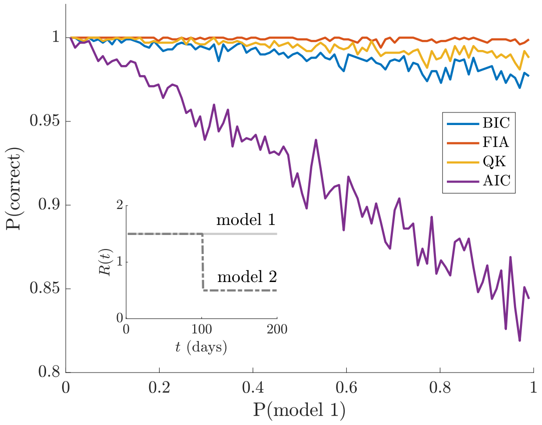

Fig. 6: **Binary hypothesis selection problem.** We test the ability of several model selection criteria by considering a binary classification problem in which the null model 1 has no change in $R(t)$ (solid, inset), while the alternative model 2 has a rapid decline (dashed, inset). We generate 10^3 independent epidemic curves randomly according to model 1 with probability $\mathbb{P}(\text{model 1})$, and compute the ability of each criteria to decipher the correct model, $\mathbb{P}(\text{correct})$. We find that the FIA approach outperforms all other metrics.

For the second, and more complicated problem, we consider models with piecewise-constant $R(t)$ changes after every k^* days, with k^* looping across the search space of $20 \leq k \leq T = m$. As before this space is restricted so that $pk = T$ with p and k as integers, with $T = 400$ days. For every possible k^* we generate 10^5 independent epidemics, allowing $R(t)$ to vary in each run, with magnitudes uniformly drawn from $[0.5, 5]$. Each run features a different random telegraph $R(t)$ function and we set $v = 100$.

Key results are shown in Fig. 7. The FIA attains the best accuracy, followed by the QK, BIC and AIC (main panel). The strong performance of both MDL-based criteria suggests that parametric complexity is important. However, the FIA does not always dominate, and can do worse than the BIC and QK when v is large compared to the actual space from which $R(t)$ is drawn (a similar effect occurs if the minimum $R(t)$ is notably above 0). We discuss these cases in Section III-E, explaining why $v = 1.5$ is used in the inset of Fig. 7.

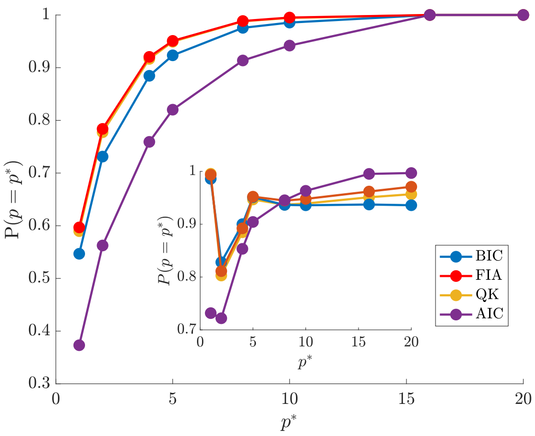

Fig. 7: **Renewal model selection consistency problem.** We simulate 10^5 epidemics from renewal models with $20 \leq k \leq T = 400$ and $p = {}^{T}\!/_{k}$. We test the ability of several model selection criteria to detect the true $p = p^*$ from among this set. Each epidemic has an independent, piecewise-constant $R(t)$. In the main panel these pieces are uniformly drawn from $[0.5, 5]$, $v = 100$, and we consider the probability of detection as a function of p^*. The FIA metric dominates all others at every p^*. In the inset the $R(t)$ range is $[0.75, 1.5]$ and now $v = 1.5$. Here the FIA has the best average performance but does not dominate at every p^*. Circles indicate data-points.

D. Adaptive Estimation: Phylogenetic Skyline Models

We verify the FIA performance on several skyline problems. We simulate uniformly sampled phylogenies (i.e. the sample times forming the tips of the phylogeny are spread evenly over some interval), with m coalescent events, using [11]. Increasing the sampling density within that interval leads to improved data and hence a larger m. We define our p segments as groups of k coalescent events. Skyline model selection is more involved because the end-points of the p segments coincide with coalescent events (see Section II-A). While this ensures statistical identifiability, it means that grouping is sensitive to phylogenetic noise [31], and that p changes for a given k if m varies ($m = pk$). This can result in MLEs, even at optimal groupings, appearing delayed or biased relative to $N(t)$, when $N(t)$ is not within the class of grouped piecewise functions.

Nevertheless, we start by examining how our FIA approach balances between the extremes in Fig. 2. We restrict our grouping parameter to $4 \leq k \leq 80$, set $v = 10^3$ $(\max N(t) = 300)$ and apply the FIA of Eq. (12) to obtain Fig. 8 and Fig. 9. Two points are immediately visible: (i) the FIA (right panels) regulates the noise from the log-likelihood (left panels), and (ii) the FIA supports higher p^* when the data are increased (bottom panels). Specifically, the FIA characterises the bottleneck of Fig. 9 using a minimum of segments but with a delay. As data accumulate, more groups can be justified and so the FIA is able to compensate for that bias. Note that the last 1-2 coalescent events are often truncated, as they can span half the time-scale, and bias all model selection criteria [18].

We now consider two model selection problems, in which $N(t)$ belongs to the piecewise-constant function class, to formally evaluate the FIA against the QK, BIC and AIC (see Section III-B). The first is a binary hypothesis test between a Kingman coalescent null model [13] with $N_1 = 1000$, and an alternative model featuring a single shift to $N_2 = 500$ at switch time $\tau = 250$ units. We set $v = 10^5$ and simulate 500 replicate phylogenies, with m controlling the quantity of data available per piecewise component (so the total number of coalescent events is $2m$). This is a slight abuse of previous definitions of m but is more useful here as we want to recover $p = 1$ for the null model and $p = 2$ for the alternative. Samples are introduced at 0 and τ time units only. Our model selection results are given in Fig. 10. The grouping parameter search space is $4 \leq k \leq m$ with the number of integer groups (now per component) defined as $p = {}^{m}\!/_{k}$.

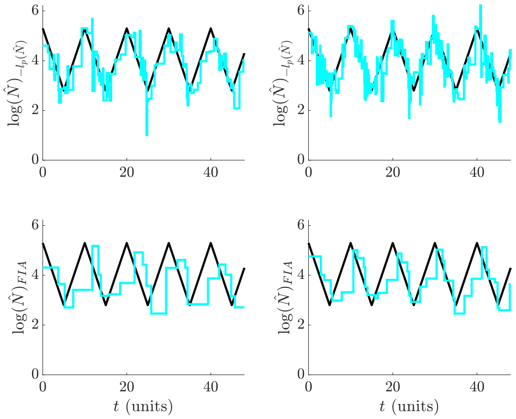

Fig. 8: **Adaptive periodic estimation with FIA.** Periodically exponential population size histories are inferred under optimal log-likelihood groupings (top panels) and FIA based selection at $v = 10^3$ (bottom panels). Phylogenies are sampled uniformly over $[0, 50]$ time units (with some extra initial samples) and data size increases from left ($m = 400$) to right ($m = 1000$).

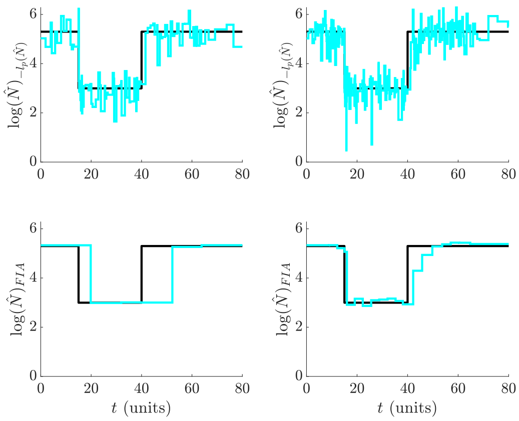

Fig. 9: **Adaptive bottleneck estimation with FIA.** Bottleneck population size histories are inferred under optimal log-likelihood groupings (top panels) and FIA based selection at $v = 10^3$ (bottom panels). Phylogenies are sampled uniformly over $[0, 60]$ time units (with some extra samples at change-points) and data size increases from left ($m = 400$) to right ($m = 1000$).

Overall, the FIA outperformed all other criteria, with the QK in close second (top panel). However, when the data sample size, m, is small, the BIC is best. Closer examination reveals that the FIA and QK have the best overall binary classification performance, achieving the highest true positive and lowest false positive rates (bottom panel). Observe that except for the AIC, which is known to not have the

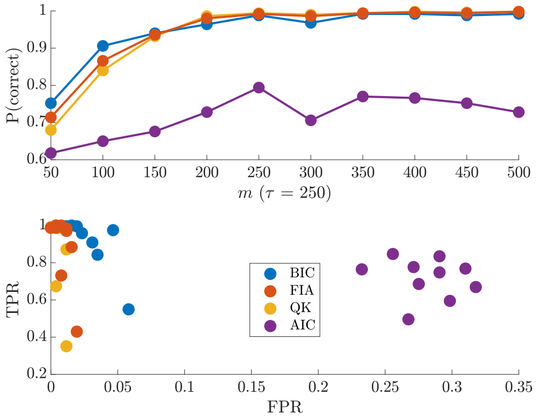

Fig. 10: **Binary hypothesis selection problem.** We simulate 500 conditionally independent phylogenies from skyline models with $4 \leq k \leq m$, and test the classification ability of model selection criteria. The null model is a Kingman coalescent with $N_1 = 1000$, and the alternative features a sharp fall to $N_2 = 500$ at $\tau = 250$ time units. The top panel give the probability of correct classification $\mathbb{P}$(correct) as a function of data size m. The FIA performs best, on average, but the BIC is better at small m. The bottom panel gives the true (TPR) and false positive rates (FPR) of the metrics. The FIA and QK have the best rates overall. Circles indicate data-points.

consistency property, the other methods all converge to near perfect detection with increasing data. We found performance to be relatively unchanged with v, and to hold if τ is doubled.

The second classification problem is more complex, and requires selection from among 5 possible square waves, with varying half-periods that are integer powers of 2. We define 15 change-point times at multiples of $\tau = 50$ time units (i.e. there are 16 components) and allow $N(t)$ to fluctuate between a maximum $N_{\max}$ and $1/2\, N_{\max}$. At each change-point and 0, equal numbers of samples are introduced, to allow approximately m coalescent events per component (there are $16m$ total coalescent events across the phylogeny). The possible models are depicted in Fig. 11. A similar problem, but for Gaussian MDL selection under binary trees, was investigated in [9].

We simulate 200 phylogenies according to each wave and compute the probability that each metric selects the correct model at $N_{\max} = 300$ and 600 in the left and right panels of Fig. 12. The grouping parameter (k) search space is set to m times the half-period of every wave and $v = 10^3$. The FIA has the best overall performance at both $N_{\max}$ settings, with the QK close behind. At $N_{\max} = 300$, there is a greater mismatch with v and so the FIA does not uniformly dominate (the AIC is better at small m). As $N_{\max} = 600$ gets closer to v this issue dissipates. We discuss the dependence of FIA on v for this problem in Section III-E. The probability of detection improves as the sample phylogeny data size (m) increases (consistency).

E. *Weaknesses of Piecewise Model Selection*

In Section III-C and Section III-D we found the FIA to be a viable and top performing model selection strategy, when compared to standard metrics of similar computability such as the AIC and BIC. However, the FIA does not always dominate, and can do worse if v is large relative to the actual space from which $R(t)$ or $N(t)$ is drawn. In such cases, the incorrect parameter bounds can lead to the FIA overestimating the complexity of the generating renewal or skyline models. While the QK criteria offers a more stable and reasonably performing MDL alternative, it is less interpretable. Here we examine the nature of this v dependence, and discuss some general issues limiting piecewise model selection.

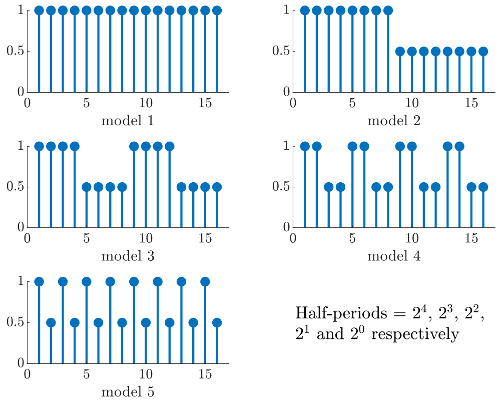

Fig. 11: **Possible square wave models.** The complex model selection problem must pick the correct model from 5 square waves given a phylogeny sampled at each change-point. Each square wave varies between $N_{\max}$ and $1/2\, N_{\max}$ (ratios shown on y axes), and occurs with varying half-periods over 16 segments (x axes) of duration τ.

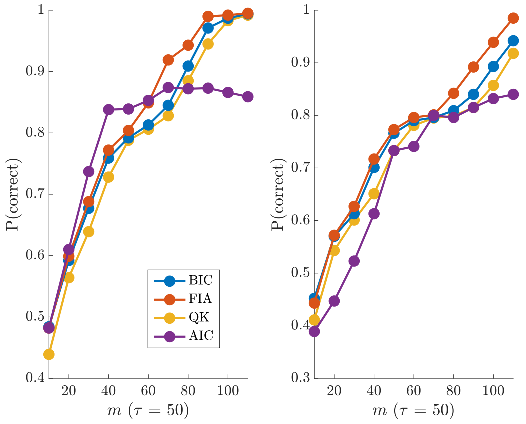

Fig. 12: **Skyline model selection consistency problem.** We simulate 200 sampled phylogenies from each square wave model of Fig. 11, with m coalescent events per component. The probability that several model selection criteria select the true (correct) model is shown at $v = 10^3$ for $N_{\max} = 300$ (left) and $N_{\max} = 600$ (right). The FIA is the most accurate criteria on average. Its performance improves with m and as v gets closer to the true $N_{\max}$. Circles are data-points.

In the inset of Fig. 7 we showed the FIA outperforming other metrics for a model selection problem over piecewise $R(t)$ functions drawn within the artificial range [0.75, 1.5] (the AIC was better at higher p due to its tendency to overfit). We achieved this by setting v to the true $R_{\max} = 1.5$. However, when there is a significant mismatch between v and $R_{\max}$ we find that the FIA is inferior to the QK and BIC. Fig. 13 illustrates, at $v = 100$ and 6, how the magnitude of this mismatch influences relative performance. However, this effect is not always important, as seen in the main panel of Fig. 7.

The skyline model also features this FIA v-dependence. We characterise this effect by re-examining the square wave model selection problem of Fig. 12, but over a range of v between 10^2 and 10^5. Fig. 14 investigates the resulting changes in the FIA detection probability, at $N_{\max} = 300$ (left) and 600 (right). There we observe,

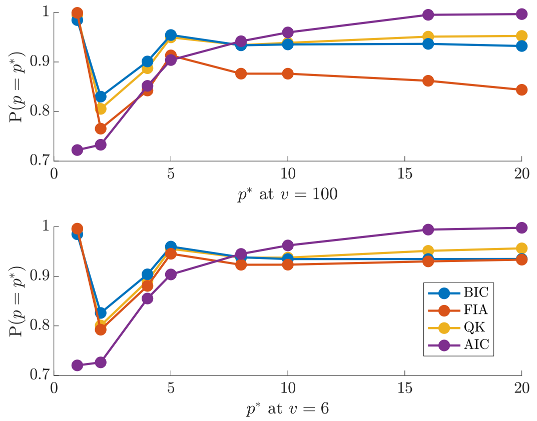

Fig. 13: **Renewal FIA parameter space sensitivity.** We repeat the simulations in the inset of Fig. 7 but at different v. The performance of FIA clearly depends on the discrepancy between v and $\max_t R(t) = 1.5$, and becomes inferior when v is dramatically above this maximum (top panel). Circles are data-points.

that while the FIA is sensitive to v, it still performs well over the entire range. Thus, sometimes, FIA can be a choice model selection metric, even in the absence of reasonable parameter space knowledge.

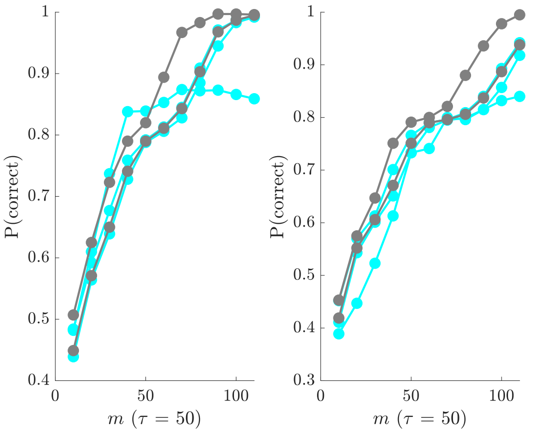

Fig. 14: **Skyline FIA parameter space sensitivity.** We revisit the simulations of Fig. 11, but vary v between 10^2 and 10^5. The AIC, BIC and QK from Fig. 11 are in cyan, while the best and worst case FIA values are in grey. While the FIA does depend on v, interestingly, its performance is still superior on average. The left and right panels are at $N_{\max} = 300$ and 600. Circles are data-points.

Lastly, we comment on some general issues that affect the model selection performance of any metric on renewal and skyline models. The MLEs and FIs of the renewal model depend on the λ_j and i_j groups. Consequently, epidemics with low observed incidence (i.e. likely to have $i_j = 0$) and diseases possessing sharp (low variance) generation time distributions (i.e. likely to achieve $\lambda_j = 0$) will be difficult to adaptively estimate. This is why we conditioned on the epidemic not dying out in Section II-A. Similarly, the MLEs and FIs of the skyline are sensitive to m_j, meaning that it is necessary to ensure each group has coalescent events falling within its duration. Forcing segment end-points to coincide with coalescent events, as in [5], guards against this identifiability problem [23]. However, skyline model selection remains difficult even after averting this issue. This follows from the random timing of coalescent events, which means that regular k groupings can miss change-points, and that long branches can bias analysis [20]. These are known skyline plot issues and evidence why we truncated the last few events in the non-piecewise $N(t)$ simulations of Section III-D. Furthermore, with both models there will always be a limit to the maximum temporal precision attainable by $R(t)$ and $N(t)$ estimates. Changes in $R(t)$ on a finer time scale than that of the observed incidence curve are impossible to recover, while it is not possible to ever get more $N(t)$ segments than the number of coalescent events [23]. This cautions against naively applying the criteria we have developed here. It is necessary to first understand and then prepare for these preconditions before sensible model selection results can be obtained.

IV. Discussion

Identifying fluctuations in effective population size, $N(t)$ and reproduction number, $R(t)$ is vital to understanding the retrospective and continuing behaviour of an epidemic, at the population level. A significant swing in $R(t)$ could, for example, inform on a key change to disease transmission, while a steep shift in $N(t)$ could evidence the historical impact of a vaccine [3] [31]. Piecewise-constant approaches, such as the skyline and renewal model, are a tractable way of separating insignificant fluctuations (the constant segments) from noteworthy ones (the change-points). However, the efficacy of these models requires principled and data-justified selection of their dimension, p. Failure to do so, as in Fig. 3 and Fig. 2, could result in salient changes being misidentified (i.e. underfitting) or random noise being over-interpreted (i.e. overfitting).

Existing approaches to p-selection are mostly either heuristic (e.g. by visual affirmation), obscure (i.e. p is implicitly set by complex algorithms) or computationally demanding [3] [1] [10]. Resolving these issues is our focus in this work. We started by proving that ascribing p solely on the evidence of the log-likelihood (i.e. the model fit) guarantees overfitting (see Eq. (9)). Consequently, it is absolutely necessary to penalise the log-likelihood with a measure of model complexity. Standard AIC and BIC, which are easy to compute, treat model complexity as either equivalent to p or p mediated by the observed data size (see Eq. (10) and Eq. (11)). However, this description is incomplete, and neglects parametric complexity.

Parametric complexity describes how the functional relationship among parameters matters. The general FIA of Eq. (6) defines this complexity as an integral across parameter space [17]. Unfortunately, this integral is often difficult to evaluate, rendering the FIA impractical. However, the piecewise-constant nature of renewal and skyline models, together with their Poisson data structures, allowed us to reduce this integral. This led to Eq. (12) and Eq. (13), which form our main results. These expressions are no more difficult to compute than AIC and BIC, and disaggregate model complexity as follows:

$$\underbrace{\underbrace{\sum\nolimits_{j=1}^{p}}_{\text{model dimension}} \underbrace{-i_j \log \frac{i_j}{\lambda_j}}_{\text{model fit}} + \underbrace{\frac{1}{2}\log \lambda_j}_{\text{data resolution}} + \underbrace{\frac{1}{2}\log \frac{2v}{\pi}}_{\text{parametric complexity}}}_{\text{model fit versus complexity}}$$

While the above breakdown is for Eq. (13), an analogous one exists for Eq. (12). Intriguingly, the parametric complexity is now a simple function of v, the unknown parameter domain maximum.

Knowledge of v is the main cost of our metric. This parameter limit requirement is not unusual and can often improve estimates. In [21] and [22], for example, this extra knowledge was shown to facilitate exact inference from sampled phylogenies. In Fig. 13 and Fig. 14 we explored the effect of incorrectly specifying v. While drastic mismatches between the true and assumed v can be detrimental, we found that in some cases poor knowledge of v can actually be inconsequential. We adapted the QK metric of [27] to obtain Eq. (14) and Eq. (15). These expressions, though less interpretable than the FIA, also somewhat account for parametric complexity and offer good performance should reasonable knowledge of v be unavailable.

The FIA approximates the MDL model selection strategy, which is known to have the desirable theoretical properties of generalisability (it balances overfitting and underfitting) and consistency (it selects the true model with increasing probability as data accumulate) [8]. In Fig. 4-Fig. 5 and Fig. 8-Fig. 9 we demonstrated that the FIA not only inherits the generalisability property, but also regulates its selections based on the available data. Higher data resolution supports larger p as both bias and variance can be simultaneously reduced under these conditions [32]. We then validated the consistency of FIA. Fig. 6, Fig. 7, Fig. 10 and Fig. 12 confirmed this property, in addition to benchmarking its performance against the comparable AIC and BIC. We found that the FIA consistently outperformed all other metrics, provided that v was not drastically misspecified.

Thus, we recommend FIA as a principled, transparent and computationally simple means of adaptively estimating informative changes in $N(t)$ and $R(t)$, and for diagnosing the relative contributions of different components of model complexity. Sampled phylogenies and incidence curves, and their associated skyline and renewal models, have often been treated distinctly within the epidemiological literature. While they do solve different problems, a key point of our work is the unification of both their models and data structures within the piecewise Poisson framework. This allowed us to analogously characterise their statistical and complexity properties, and achieve cross-model insights. Piecewise Poisson models abound in biology [23]. This may allow wider application of MDL and promote the cross-fertilisation of statistical insights between modelling approaches.

Acknowledgments

KVP and CAD acknowledge joint Centre funding from the UK Medical Research Council and Department for International Development under grant reference MR/R015600/1. CAD thanks the UK National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Modelling Methodology at Imperial College London in partnership with Public Health England (PHE) for funding (grant HPRU-2012–10080).

References

[1] G. Baele, P. Lemey, T. Bedford, et al. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Mol. Biol. Evol*, 29(9):2157–67, 2012.

[2] A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Trans. Info. Theo*, 44(6):2743–60, 1998.

[3] A. Cori, N. Ferguson, C. Fraser, et al. A new framework and software to estimate time-varying reproduction numbers during epidemics. *Am. J. Epidemiol*, 178(9):1505–12, 2013.

[4] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley and Sons, second edition, 2006.

[5] A. Drummond, A. Rambaut, B. Shapiro, and O. Pybus. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol*, 22(5):1185–92, 2005.

[6] C. Fraser. Estimating individual and household reproduction numbers in an emerging epidemic. *PLoS One*, 8:e758, 2007.

[7] C. Fraser, D. Cummings, D. Klinkenberg, et al. Influenza transmission in households during the 1918 pandemic. *Am. J. Epidemiol*, 174(5):505–14, 2011.

[8] P. Grunwald. *The Minimum Description Length Principle*. The MIT Press, 2007.

[9] A. Hanson and P. Fu. *Advances in Minimum Description Length: Theory and Applications*, chapter Applications of MDL to selected families of models. MIT Press, 2004.

[10] S. Ho and B. Shapiro. Skyline-plot methods for estimating demographic history from nucleotide sequences. *Mol. Ecol. Res*, 11:423–34, 2011.

[11] M. Karcher, J. Palacios, S. Lan, et al. PHYLODYN: an R package for phylodynamic simulation and inference. *Mol. Ecol. Res*, 17:96–100, 2017.

[12] R. Kass and A. Raftery. Bayes factors. *J. Amer. Stat. Assoc*, 90(430):773–95, 1995.

[13] J. Kingman. On the genealogy of large populations. *J. Appl. Prob*, 19:27–43, 1982.

[14] E. Lehmann and G. Casella. *Theory of Point Estimation*. Springer-Verlag, second edition, 1998.

[15] P. Lemey, O. Pybus, B. Wang, et al. Tracing the origin and history of the HIV-2 epidemic. *PNAS*, 100(11):6588–92, 2003.

[16] V. Minin, E. Bloomquist, and M. Suchard. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol. Biol. Evol*, 25(7):1459–71, 2008.

[17] J. Myung, D. Navarro, and M. Pitt. Model selection by normalized maximum likelihood. *J. Math. Psychol*, 50:167–79, 2006.

[18] M. Nordborg. *Handbook of Statistical Genetics: Coalescent Theory*. John Wiley and Sons, 2001.

[19] P. Nouvellet, A. Cori, T. Garske, et al. A simple approach to measure transmissibility and forecast incidence. *Epidemics*, 22:29–35, 2018.

[20] K. Parag, L. du Plessis, and O. Pybus. An integrated framework for the joint inference of demographic history and sampling intensity from genealogies or genetic sequences. *BioRxiv*, 686378, 2019.

[21] K. Parag and O. Pybus. Optimal point process filtering and estimation of the coalescent process. *J. Theor. Biol*, 421:153–67, 2017.

[22] K. Parag and O. Pybus. Exact bayesian inference for phylogenetic birth-death models. *Bioinformatics*, 34(21):3638–45, 2018.

[23] K. Parag and O. Pybus. Robust design for coalescent model inference. *Syst. Biol*, syz008, 2019.

[24] M. Pitt, I. Myung, and S. Zhang. Toward a method of selecting among computational models of cognition. *Psych. Rev*, 109(3):472–91, 2002.

[25] O. Pybus, M. Charleston, S. Gupta, et al. The epidemic behavior of the hepatitis C virus. *Science*, 292(5525):2323–5, 2001.

[26] O. Pybus, A. Rambaut, and P. Harvey. An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics*, 155:1429–37, 2000.

[27] G. Qian and H. Kunsch. Some notes on Rissanen's stochastic complexity. *IEEE Trans. Info. Theo*, 44(2):782–6, 1998.

[28] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–71, 1978.

[29] J. Rissanen. Fisher information and stochastic complexity. *IEEE Trans. Info. Theo*, 42(1):40–7, 1996.

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

[31] K. Strimmer and O. Pybus. Exploring the demographic history of DNA sequences using the generalized skyline plot. *Mol. Biol. Evol*, 18(12):2298–305, 2001.

[32] T. van Erven and P. Grunwald. Catching up faster by switching sooner: a predictive approach to adaptive estimation with an application to the AIC–BIC dilemma. *J. R. Statist. Soc. B*, 74(3):361–417, 2012.

[33] J. Wallinga and M. Lipsitch. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proc. R. Soc. B*, 274:599–604, 2007.

[34] J. Wallinga and P. Teunis. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *Am. J. Epidemiol*, 160(6):509–16, 2004.