

RESEARCH ARTICLE

An approximate Bayesian approach for estimation of the instantaneous reproduction number under misreported epidemic data

Oswaldo Gressani¹  | Christel Faes¹ | Niel Hens^{1,2}

¹Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat), Data Science Institute, Hasselt University, Hasselt, Belgium

²Centre for Health Economics Research and Modelling Infectious Diseases, Vaxinfectio, University of Antwerp, Antwerp, Belgium

Correspondence

Oswaldo Gressani, Data Science Institute, Hasselt University, Agoralaan, Campus Diepenbeek, Gebouw D E125, 3590 Belgium.

Email: oswaldo.gressani@uhasselt.be

Funding information

European Union's Research and Innovation Action EpiPose, Grant/Award Number: 101003688



This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

Abstract

In epidemic models, the effective reproduction number is of central importance to assess the transmission dynamics of an infectious disease and to orient health intervention strategies. Publicly shared data during an outbreak often suffers from two sources of misreporting (underreporting and delay in reporting) that should not be overlooked when estimating epidemiological parameters. The main statistical challenge in models that intrinsically account for a misreporting process lies in the joint estimation of the time-varying reproduction number and the delay/underreporting parameters. Existing Bayesian approaches typically rely on Markov chain Monte Carlo algorithms that are extremely costly from a computational perspective. We propose a much faster alternative based on Laplacian-P-splines (LPS) that combines Bayesian penalized B-splines for flexible and smooth estimation of the instantaneous reproduction number and Laplace approximations to selected posterior distributions for fast computation. Assuming a known generation interval distribution, the incidence at a given calendar time is governed by the epidemic renewal equation and the delay structure is specified through a composite link framework. Laplace approximations to the conditional posterior of the spline vector are obtained from analytical versions of the gradient and Hessian of the log-likelihood, implying a drastic speed-up in the computation of posterior estimates. Furthermore, the proposed LPS approach can be used to obtain point estimates and approximate credible intervals for the delay and reporting probabilities. Simulation of epidemics with different combinations for the underreporting rate and delay structure (one-day, two-day, and weekend delays) show that the proposed LPS methodology delivers fast and accurate estimates outperforming existing methods that do not take into account underreporting and delay patterns. Finally, LPS is illustrated in two real case studies of epidemic outbreaks.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2023 The Authors. *Biometrical Journal* published by Wiley-VCH GmbH.

KEYWORDS

Bayesian P-splines, epidemic renewal equation, Laplace approximation, misreported data

1 | INTRODUCTION

In the presence of an epidemic outbreak, it is of vital importance to gain insights into the transmissibility of a disease and have a clear understanding of the mechanisms driving the dynamics of infections over time. Real-time information on epidemiological parameters can have a determinant role in orienting public health policies and initiating proactive interventions for disease control and prevention. The effective reproduction number, R_t , defined as the average number of secondary infections generated by a primary infected individual in a susceptible population at a calendar time $t > 0$ (Bettencourt & Ribeiro, 2008; Hethcote, 2000) is probably among the most important parameters that permit to gain knowledge of time-dependent variations in the transmission potential (Nishiura & Chowell, 2009). As our interest lies in measuring transmission (under misreporting) at a specific time point, R_t denotes the instantaneous reproduction number rather than the case reproduction number that is used to quantify transmission from a cohort perspective (Cori et al., 2013; Gostic et al., 2020). During the last 20 years or so, serious efforts have been invested in the development of sophisticated inferential methods to estimate the time-varying reproduction number, a challenging task as recently recalled and beautifully summarized by Gostic et al. (2020). Among early contributors, one can cite Wallinga and Teunis (2004) who propose a likelihood-based estimation of the effective reproduction number solely based on information provided by the observed epidemic curve. This work was further extended and generalized by Cauchemez et al. (2006) who assume no prior knowledge of the generation interval, that is, the time elapsed between when a susceptible person becomes infected (infector) and when that individual infects another person (infected) (Svensson, 2007). Their model captures the pattern of R_t over time by using partial tracing information and Markov chain Monte Carlo (MCMC) for posterior inference.

Delay in reporting and underreporting of incidence data (Cui & Kaldor, 1998; Fraser et al., 2009; Lawless, 1994) adds a further layer of difficulty that cannot be ignored when designing a model to estimate the reproduction number, as misreported data alter the true underlying signal of an epidemic curve and hence introduce bias in estimates of R_t . The model of Hens et al. (2011) explicitly accounts for underreporting and provides estimates of R_t based on a frequentist likelihood approach that assumes a fixed serial interval distribution (the time elapsed between symptom onset in an infected and its infector). Azmon et al. (2014) go one step further and use a Bayesian semiparametric approach with penalized radial splines to model R_t accounting simultaneously for underreporting and delay in reporting. They use the renewal equation (Feller, 1941; Fraser, 2007; Nouvellet et al., 2018; Wallinga & Lipsitch, 2007) to establish a link between R_t and daily incidence counts to describe the evolutionary dynamics of an epidemic. When resorting to Bayesian methods for inference in epidemiological models, MCMC sampling practically imposes itself as a default option as it is a deeply routed and versatile tool that is made accessible and implementable by many computer software packages such as WinBUGS (Lunn et al., 2000) or JAGS (Plummer et al., 2003).

Notwithstanding the capacity of MCMC to explore virtually any posterior target distribution, there is often a large computational price to pay accompanied by eventual convergence problems and the systematic necessity to diagnose MCMC samples. To overcome these limitations and get rid of the computational hurdles imposed by MCMC, we propose a completely sampling-free approximate Bayesian inference approach for fast and flexible estimation of the reproduction number R_t in an epidemic model with misreported data. In particular, we revisit the model of Azmon et al. (2014) by using Bayesian P-splines (Eilers & Marx, 1996; Lang & Brezger, 2004) for flexible estimation of the time-varying reproduction number (Gressani et al., 2022) and Laplace approximations (Gressani & Lambert, 2018, 2021; Rue et al., 2009) to the conditional posterior of the latent spline vector related to R_t for fast computation. Our Laplacian-P-splines (LPS) model is based on the following three assumptions: (1) a closed susceptible population (i.e., no imported cases), (2) the generation interval distribution is assumed to be known, and (3) an informative prior on the reporting rate is available. A composite link model (Eilers, 2007; Thompson & Baker, 1981) is used to represent the delay process, for which we investigate three possible structures, one-day, two-day, and weekend delays. Moreover, we assume that the mean number of new contaminations is driven by the renewal equation, that is, the product of the effective reproduction number and a discrete convolution between past cases and generation probabilities. Several simulation scenarios show that the proposed

methodology gives accurate estimates of the reporting and delay probabilities and is also able to precisely capture the pattern of R_t over the course of an epidemic. Encouraging results are also observed when comparing LPS with the EpiEstim package of Cori et al. (2013), which is known for producing robust estimates of R_t . The key advantage of our approach is that even though we work from a completely Bayesian perspective, LPS delivers estimates of key epidemiological model parameters in seconds, while several minutes or hours would be needed with MCMC algorithms. This is partly due to the fact that Laplace approximations are based on analytically derived expressions for the gradient and Hessian of the log-likelihood of the model.

The presentation of the LPS methodology for fast inference of R_t under misreported data is structured as follows. In Section 2, the Laplacian-P-splines model is introduced and priors are imposed on the hyperparameters. After summarizing the Bayesian model, we show how the conditional posterior of the spline vector related to R_t is approached with Laplace approximations. Next, posterior inference on reporting and delay probabilities is presented along with the construction of credible intervals for the model parameters. Section 3 is devoted to a detailed numerical study that assesses the performance of LPS under various epidemic scenarios. In Section 4, we illustrate our new methodology on real datasets, and Section 5 concludes the paper with a discussion.

2 | THE BAYESIAN LAPLACIAN-P-SPLINES MODEL

2.1 | Misreported epidemic data

Let $T > 0$ denote the total number of days of an epidemic and $\mathcal{M} = \{M_1, \dots, M_T\}$ the latent set of contaminations with $M_t \in \mathbb{N}$ the (unobserved) number of new contaminations on day t . We write p_j for the probability that j days have passed until occurrence of infection in an infector–infectee pair and denote by $\mathbf{p} = \{p_1, \dots, p_k\}$ the generation interval distribution of maximum length k , assumed to be known here. Following Azmon et al. (2014), we assume that M_t is Poisson distributed with mean μ_t and probability mass function:

$$p(M_t | R_t, \mathcal{H}_t^k, \mathbf{p}) = \frac{\exp(-\mu_t) \mu_t^{M_t}}{M_t!}, \quad (1)$$

where R_t is the reproduction number at day t and $\mathcal{H}_t^k = \{M_{t-1}, \dots, M_{t-k}\}$ is the set of past values for the number of cases with history of length k . The relationship between the mean number of new cases at day t and past infections is governed by the epidemic renewal equation:

$$\mu_t = \begin{cases} \mu_1 & ; \text{ for } t = 1, \\ R_t \left(\sum_{s=1}^{\min(t-1, k)} p_s M_{t-s} \right); & \text{ for } t > 1. \end{cases} \quad (2)$$

Equation (2) suggests that for $t > 1$ the mean number of new contaminations on day t (namely μ_t) is a convex combination of the past number(s) in the set \mathcal{H}_t^k weighted by R_t , that is, the average number of secondary cases generated by a primary case at moment t . The observed set of disease counts subject to underreporting, and delay in reporting is denoted by $\mathcal{D} = \{O_1, \dots, O_T\}$. The daily reporting probability is given by $\rho \in (0, 1)$ and is considered to be time-homogeneous over the entire duration of the epidemic. The fraction of cases on day i reported on day t is written as $\delta_{i \rightarrow t} \in [0, 1]$ and represents a delay probability that can be embedded under various structures in the model via a composite link framework (Eilers, 2007). In this paper, we consider three delay structures proposed in Section 2.2 of Azmon et al. (2014), namely a one-day, two-day, and weekend delay pattern. The underreporting-delay process is reflected in the Poisson distributional assumption for O_t with mean $\rho \mu_t^d$:

$$p(O_t | \rho, \mu_t^d) = \frac{\exp(-\rho \mu_t^d) (\rho \mu_t^d)^{O_t}}{O_t!}, \quad (3)$$

where $\mu_t^d := \sum_{i=1}^t \delta_{i \rightarrow t} \mu_i$ is the average number of cases on day t subject to delays computed by aggregating the current and past (unobserved) mean number of cases weighted by their associated delay probability. Mathematically, the delay

pattern is determined by a square composition matrix C of dimension 7×7 , with rows and columns representing the days of a week. The link between $\boldsymbol{\mu} = (\mu_1, \dots, \mu_7)^\top$ and μ_t^d can thus be written compactly as $\mu_t^d = C_{,wd(t)}^\top \boldsymbol{\mu}$, where $C_{,wd(t)}$ is column $wd(t)$ of the composition matrix C and $wd(t) \in \{1 := \text{Monday}, \dots, 7 := \text{Sunday}\}$ is an index function returning an integer corresponding to the day of the week for time point t . By construction, adding the probabilities along each row of C yields unity; a constraint translated as $\sum_t \delta_{i \rightarrow t} = 1$. Appendix A contains the composition matrices for the three delay patterns considered in this paper.

2.2 | Bayesian model formulation

2.2.1 | Flexible specification of the effective reproduction number

P-splines (Eilers & Marx, 1996) are an interesting candidate to model the time-varying reproduction number dynamics over the considered epidemic period. Two main appealing features of this spline smoother are worth mentioning. First, the penalty matrix can be straightforwardly obtained with minimal numerical effort for any chosen penalty order. Second, P-splines are naturally adapted in a Bayesian framework by replacing the deterministic discrete difference penalty by random walks with Gaussian errors (Lang & Brezger, 2004), yielding Gaussian priors for the B-spline coefficients and thus translating frequentist P-splines to Bayesian P-splines. This motivates our choice for modeling the log of the effective reproduction number as a linear combination of B-splines, that is, $\log(R_t) = \sum_{k=1}^K \theta_k b_k(t) = \boldsymbol{\theta}^\top b(t)$, where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)^\top$ is the latent vector of B-spline coefficients and $b(\cdot) = (b_1(\cdot), \dots, b_K(\cdot))^\top$ is a basis of cubic B-splines on the domain $\mathcal{T} = [0, T]$ ranging from 0 to the last day of the epidemic. A “large” number K of B-spline basis functions is specified to ensure that the fitted curve for R_t is flexible enough, and a discrete penalty term $\lambda \boldsymbol{\theta}^\top P \boldsymbol{\theta}$ is introduced as a measure of roughness of the B-spline coefficients with $\lambda > 0$ as a tuning parameter. The penalty matrix is given by $P = D_r^\top D_r + \varepsilon I_K$ and equals the product of r th-order difference matrices D_r with a small perturbation on the main diagonal (here $\varepsilon = 10^{-5}$) to ensure full rankedness. The Gaussian prior for the spline vector is denoted by $\boldsymbol{\theta} | \lambda \sim \mathcal{N}_{\dim(\boldsymbol{\theta})}(0, Q_\lambda^{-1})$, with precision matrix $Q_\lambda = \lambda P$. The tuning parameter is assigned a Gamma prior $\lambda \sim \mathcal{G}(a_\lambda, b_\lambda)$ with mean a_λ/b_λ and variance a_λ/b_λ^2 . Choosing $a_\lambda = b_\lambda = 10^{-5}$ yields a dispersed (yet proper) prior for λ with a large variance (see, e.g., Lang & Brezger, 2004; Lambert & Eilers, 2005).

2.2.2 | Prior assumptions on reporting and delay probabilities

The tuning parameter and the reporting and delay probabilities are gathered in the hyperparameter vector $\boldsymbol{\eta} = (\lambda, \rho, \delta_{i \rightarrow j}; i, j = 1, \dots, 7)^\top$. A noninformative uniform prior is imposed on the delay probabilities, that is, $\delta_{i \rightarrow j} \sim \mathcal{U}(0, 1)$ for $i, j = 1:\text{Monday}, 2:\text{Tuesday}, \dots, 7:\text{Sunday}$. Typically, the reporting rate ρ cannot be obtained from real-time data (Heesterbeek et al., 2015) and thus needs to be estimated, adding an extra layer of difficulty in the inference process. As noted by Thompson et al. (2019), underreporting has already proved to be a burden for inference and forecasting in various infectious disease models. Without minimal prior knowledge of ρ , posterior estimates of key epidemiological quantities such as the effective reproduction number will likely be biased and accompanied by high uncertainty with wide credible intervals. We, therefore, assume that minimal prior information is available for the reporting probability translated by a uniform prior $\rho \sim \mathcal{U}(a_\rho, b_\rho)$ with bounds $0 \leq a_\rho < b_\rho \leq 1$ that encompass the true underlying ρ . Such informative priors can, for instance, be constructed using hierarchical models based on available historical data (Riou et al., 2018) or by using posterior distributions of ρ inferred from previous studies (Stocks et al., 2020). When serological surveillance data are available, prior knowledge of the reporting rate can be extrapolated by computing the ratio of reported cases over the number of seropositive individuals (Abrams et al., 2021; Zhao et al., 2020).

To approximate the (latent) number of cases on a given day M_t , we use a simple inflation factor approach as in Jandarov et al. (2014) and Stocks et al. (2020) based on our prior assumption for ρ . More specifically, we use the following approximation $\tilde{M}_t = (1/\tilde{\rho})O_t$ (rounded to the nearest integer), where $\tilde{\rho} = (a_\rho + b_\rho)/2$ is the midpoint in the prior domain $[a_\rho, b_\rho]$. Although more sophisticated methods exist to account for underreporting (see, e.g., Bracher & Held, 2021), the main rationale for using a simple multiplication rule is that the focus of this paper is on the methodological approach for fast approximate Bayesian inference of R_t taking reporting/delay into account, rather than on an explicit modeling of the

(under)reporting process in itself. To summarize, the full Bayesian model is given by

$$\begin{aligned}
 (O_t | \rho, \mu_t^d) &\sim \text{Poisson}(\rho \mu_t^d), \\
 \mu_t^d &= \sum_{i=1}^t \delta_{i \rightarrow t} R_i \left(\sum_s p_s \tilde{M}_{i-s} \right), \\
 \log(R_t) &= \sum_{k=1}^K \theta_k b_k(t), \\
 \theta | \lambda &\sim \mathcal{N}_{\dim(\theta)}(0, Q_\lambda^{-1}), \\
 \lambda &\sim \mathcal{G}(a_\lambda = 10^{-5}, b_\lambda = 10^{-5}), \\
 \delta_{i \rightarrow j} &\sim \mathcal{U}(0, 1) \text{ for } i, j = 1, \dots, 7, \\
 \rho &\sim \mathcal{U}(a_\rho, b_\rho) \text{ with } 0 \leq a_\rho < b_\rho \leq 1.
 \end{aligned} \tag{4}$$

2.3 | Approximation of the conditional posterior spline vector

The log-likelihood function of the Poisson model for the observed number of cases is

$$\ell(\theta, \eta; D) \doteq \sum_{t=1}^T O_t \log(\rho \mu_t^d) - \rho \mu_t^d, \tag{5}$$

where \doteq denotes equality up to an additive constant. Replacing μ_t^d by its extensive form in terms of the epidemic renewal equation and the spline specification of the effective reproduction number, the mean of O_t , namely $\rho \mu_t^d$, is written as the following function:

$$s_t(\theta, \eta) := \rho \sum_{i=1}^t \delta_{i \rightarrow t} \exp \left(\sum_{k=1}^K \theta_k b_k(i) \right) \left(\sum_s p_s \tilde{M}_{i-s} \right). \tag{6}$$

The above equation is used to write the log-likelihood as follows:

$$\ell(\theta, \eta; D) \doteq \sum_{t=1}^T (O_t \log(s_t(\theta, \eta)) - s_t(\theta, \eta)). \tag{7}$$

Using (7) and Bayes' rule, the conditional posterior of the spline vector is

$$\begin{aligned}
 p(\theta | \eta, D) &\propto \exp(\ell(\theta, \eta; D)) p(\theta | \lambda) \\
 &\propto \exp \left(\sum_{t=1}^T (O_t \log(s_t(\theta, \eta)) - s_t(\theta, \eta)) - \frac{\lambda}{2} \theta^\top P \theta \right).
 \end{aligned} \tag{8}$$

Let $g_t(\theta, \eta) := O_t \log(s_t(\theta, \eta)) - s_t(\theta, \eta)$ denote the contribution of observables at day t to the log-likelihood. A Laplace approximation to $p(\theta | \eta, D)$ is obtained by iteratively computing a second-order Taylor expansion of $g_t(\theta, \eta)$ in terms of θ by starting from an initial guess $\theta^{(0)}$. The Taylor expansion to $g_t(\theta, \eta)$ yields a quadratic form in θ and plugging the latter into (8), one recovers (up to a multiplicative constant) a multivariate Gaussian density. The iterative Laplace approximation scheme is implemented in a Newton–Raphson type algorithm for which the gradient and Hessian of $g_t(\theta, \eta)$ and hence of the log-likelihood are analytically derived in Appendix B for maximum numerical efficiency. The Laplace approximated conditional posterior of the B-spline vector after convergence of the Newton–Raphson algorithm is denoted by $\tilde{p}_G(\theta | \eta, D) = \mathcal{N}_{\dim(\theta)}(\theta^*(\eta), \Sigma^*(\eta))$, where $\theta^*(\eta)$ is the mean (mode) and $\Sigma^*(\eta)$ the variance–covariance matrix, for a given value of the hyperparameter vector η .

2.4 | Posterior inference on hyperparameters

The posterior of the hyperparameter vector η can be written in terms of the conditional posterior of the B-spline vector derived in Section 2.3, namely

$$p(\eta|D) = \frac{p(\theta, \eta|D)}{p(\theta|\eta, D)} \propto \frac{\exp(\ell(\theta, \eta; D))p(\theta|\lambda)p(\lambda)p(\rho) \prod_{i,j} p(\delta_{i \rightarrow j})}{p(\theta|\eta, D)}. \tag{9}$$

Following Tierney and Kadane (1986) and Rue et al. (2009), the above hyperparameter posterior can be approximated by replacing the denominator in (9) with its Laplace approximation and by substituting θ by the modal value $\theta^*(\eta)$ of the latter Laplace approximation. The resulting approximated posterior is then solely a function of η :

$$\tilde{p}(\eta|D) \propto \frac{\exp(\ell(\theta, \eta; D))p(\theta|\lambda)p(\lambda)p(\rho) \prod_{i,j} p(\delta_{i \rightarrow j})}{\tilde{p}_G(\theta|\eta, D)} \Big|_{\theta=\theta^*(\eta)}. \tag{10}$$

The uniform priors on the reporting and delay probabilities vanish into the proportionality constant and so the approximated hyperparameter posterior (10) is written extensively as

$$\begin{aligned} \tilde{p}(\eta|D) \propto \exp \left(\sum_{t=1}^T (O_t \log(s_t(\theta^*(\eta), \eta)) - s_t(\theta^*(\eta), \eta)) - \frac{\lambda}{2} \theta^*(\eta)^\top P \theta^*(\eta) \right) \\ \times \lambda^{\frac{K}{2} + a_\lambda - 1} \exp(-b_\lambda \lambda) |\Sigma^*(\eta)|^{\frac{1}{2}}. \end{aligned} \tag{11}$$

As the hyperparameters live in different domains, for example, $\lambda > 0$ and $\rho \in (0, 1)$, we propose a transformation to ensure that all variables are unbounded with values in \mathbb{R} . This transformation is crucial to ensure numerical stability when using algorithms to explore $\tilde{p}(\eta|D)$. Let us define $v = \log(\lambda)$, $\check{\rho} = \log(-\log(\rho))$ and $\check{\delta}_{i \rightarrow j} = \log(-\log(\delta_{i \rightarrow j}))$ for $i, j = 1, \dots, 7$ and denote our transformed hyperparameter vector as $\check{\eta} = (v, \check{\rho}, \check{\delta}_{i \rightarrow j}; i, j = 1, \dots, 7)^\top$. Using the multivariate transformation method, the hyperparameter posterior becomes

$$\begin{aligned} \tilde{p}(\check{\eta}|D) \propto \exp \left(\sum_{t=1}^T (O_t \log(s_t(\theta^*(\check{\eta}), \check{\eta})) - s_t(\theta^*(\check{\eta}), \check{\eta})) - \frac{\exp(v)}{2} \theta^{*T}(\check{\eta}) P \theta^*(\check{\eta}) \right) \\ \times \exp(v)^{\frac{K}{2} + a_\lambda - 1} \exp(-b_\lambda \exp(v)) |\Sigma^*(\check{\eta})|^{\frac{1}{2}} \\ \times \left(\exp(v)(-\exp(-\exp(\check{\rho}) + \check{\rho})) \prod_{i,j} (-\exp(-\exp(\check{\delta}_{i \rightarrow j}) + \check{\delta}_{i \rightarrow j})) \right), \end{aligned} \tag{12}$$

where the last line equals the absolute value of the Jacobian from the multivariate transformation. The approximate posterior of the transformed hyperparameter vector in (12) is the main ingredient for posterior inference on η . At this stage, MCMC methods could be used to explore the above (approximate) target density (see, e.g., Gómez-Rubio & Rue, 2018; Vanhatalo et al., 2013). As the philosophy of our approach is to rely on a completely sampling free methodology, we decide to compute the maximum a posteriori (MAP) estimate of $\check{\eta}$ via a Newton–Raphson algorithm. Even though MAP approaches ignore the uncertainty surrounding the estimate contrary to grid-based strategies or MCMC samplers, they have the advantage of being less costly to implement from a computational perspective and still have good statistical properties as will be shown later in the simulation study.

When designing a Newton–Raphson algorithm to explore a complex posterior as in (12), great care needs to be taken to avoid convergence problems. For maximization, it is important to ensure that an ascent direction is taken at every iteration. This can be achieved by proposing a modified positive definite version of the negative Hessian whenever the latter fails to be positive definite (Goldfeld et al., 1966; Marquardt, 1963; Levenberg, 1944) combined with a backtracking strategy (e.g. step-halving) to ensure heading uphill. Taking this into account, our Newton–Raphson algorithm did not encounter any convergence issues and reached a maximum (at least a local one) in most cases. Divergence of the Newton–

Raphson algorithm can arise when iterations get absorbed into flat regions of the posterior distribution. In that case, a simple remedial measure consists of restarting the algorithm at another initial condition. Also, a too small number of B-splines in the basis might lead to an ill-behaved posterior distribution for η , so that computation of the MAP is often prone to numerical errors. We, therefore, recommend to use at least $K = 10$ B-splines (or more) to avoid such problems.

2.5 | Approximate posterior estimates and credible intervals

Let us denote by $\hat{\eta}$ the MAP estimate of η obtained from the Newton–Raphson algorithm. Plugging the latter into the Laplace approximation scheme, one recovers $\theta^*(\hat{\eta})$ and $\Sigma^*(\hat{\eta}) := \Sigma^*$, that is, the point estimate of the B-spline vector and its associated estimated variance–covariance matrix. Using the latter quantities, a point estimate of the effective reproduction number is taken to be $\hat{R}_t = \exp(\theta^{*\top} b(t))$. Let $h(\theta|t) = \log(R_t) = \theta^\top b(t)$ be the log of the reproduction number at a given day t seen as a function of the spline vector θ and consider the first-order Taylor expansion of $h(\theta|t)$ around $\theta^* = \theta^*(\hat{\eta})$:

$$h(\theta|t) \approx h(\theta^*|t) + (\theta - \theta^*)^\top \nabla h(\theta|t)|_{\theta=\theta^*}, \quad (13)$$

with gradient $\nabla h(\theta|t)|_{\theta=\theta^*} = b(t)$. Note that (13) is a linear combination of the random vector θ and that the latter has a Gaussian (conditional) posterior due to the Laplace approximation scheme. It follows that a posteriori $h(\theta|t)$ is also approximately Gaussian with mean $\mathbb{E}(h(\theta|t)) \approx h(\theta^*|t)$ and variance $\mathbb{V}(h(\theta|t)) \approx \nabla^\top h(\theta|t)|_{\theta=\theta^*} \Sigma^*(\hat{\eta}) \nabla h(\theta|t)|_{\theta=\theta^*}$. Accordingly, a $(1 - \alpha) \times 100\%$ (approximate) quantile-based credible interval for $\log(R_t)$ on day t is

$$CI_{h(\theta|t)}^{1-\alpha} = h(\theta^*|t) \pm z_{\alpha/2} \sqrt{b(t)^\top \Sigma^* b(t)}, \quad (14)$$

where $z_{\alpha/2}$ is the $\alpha/2$ -upper quantile of a standard normal distribution. Applying the $\exp(\cdot)$ transform on (14) yields the desired credible interval for R_t . While the Gaussian prior on the spline vector θ helps ensure that its conditional posterior does not substantially deviate from a Gaussian distribution, the non-Gaussian priors on the hyperparameters in η might contribute in shaping a posterior distribution that has non-Gaussian features such as heavier tails. For this reason, we advise the use of a heavier tailed distribution for components of η and assume that the marginal posterior of a hyperparameter variable η_l has a Student- t distribution (see, e.g., Martins & Rue, 2014) with $\nu = \dim(\eta) - 5$ degrees of freedom, namely $(\eta_l|D) \sim t_\nu(\check{\eta}_{l,\text{MAP}}, \sigma_{l,\text{MAP}}^2)$, where the mean $\check{\eta}_{l,\text{MAP}}$ is the MAP estimate from the Newton–Raphson algorithm and $\sigma_{l,\text{MAP}}^2$ is the appropriate diagonal entry of the inverse of the negative Hessian of the log of (12) evaluated at the MAP estimate. The resulting $(1 - \alpha) \times 100\%$ credible interval for η_l is then $\check{\eta}_l \pm t_{\nu,\alpha/2} \sqrt{(\nu/(\nu - 2))\sigma_{l,\text{MAP}}^2}$.

3 | RESULTS

3.1 | Simulation study

The performance of our approach (with 10 cubic B-splines and a penalty of order 3) is assessed in six different epidemic scenarios with a duration of $T = 30$ days. In Scenarios 1–3, the incidence data are governed by a decaying effective reproduction number $R_t = \exp(\cos(t/13))$ as in Azmon et al. (2014) with a reporting rate fixed at $\rho = 0.2$. Scenarios 4 and 5 assume more complex structures for R_t to see whether the Laplacian-P-splines model is able to capture the true dynamics of the reproduction number, and Scenario 6 is characterized by a sharp drop in R_t at day $t = 15$. Table 1 summarizes the functional form of the reproduction number, the delay structure, the reporting rate, and the assumed prior information on ρ used in each scenario. We simulate $S = 500$ replications in each scenario assuming an epidemic starting with $M_1 = 10$ index cases on day $t = 1$ and an influenza-like generation interval distribution with a mean of 2.6 days and standard deviation of 1.5 days (Cori et al., 2013; Ferguson et al., 2005). The discretized version of the generation interval is obtained with the `discr_si()` routine of the EpiEstim package. The inflated latent number of cases for weekend days (Saturday and Sunday) in Scenario 3 is obtained by multiplying $\tilde{\rho}^{-1}$ by a simple average of the cases in the preceding days of the corresponding week. The simulated data ensure that at least one case is observed on the first day, that is, $O_1 \geq 1$. Figure 1 shows the latent (M_t in red) and observed (O_t in orange) incidence data for Scenarios 1–3.

TABLE 1 Reproduction number, delay structure, reporting rate and prior on ρ for Scenarios 1-6

Scenario	Reproduction number	Delay	ρ	Prior for ρ
1	$R_t = \exp(\cos(t/13))$	One-day	$\rho = 0.20$	$\rho \sim \mathcal{U}(0, 0.40)$
2	$R_t = \exp(\cos(t/13))$	Two-day	$\rho = 0.20$	$\rho \sim \mathcal{U}(0, 0.40)$
3	$R_t = \exp(\cos(t/13))$	Weekend	$\rho = 0.20$	$\rho \sim \mathcal{U}(0, 0.40)$
4	$R_t = \exp(0.40 + 0.30 \cos(t/10) + 0.80 \sin((t\pi)/6))$	One-day	$\rho = 0.60$	$\rho \sim \mathcal{U}(0.25, 0.85)$
5	$R_t = \exp(0.98 + \sin((1.75 + 0.022t)^2)) - 0.10$	Two-day	$\rho = 0.70$	$\rho \sim \mathcal{U}(0.34, 0.94)$
6	$R_t = 1.8 \mathbb{1}(t < 15) + 0.7 \mathbb{1}(t \geq 15)$	Weekend	$\rho = 0.40$	$\rho \sim \mathcal{U}(0.20, 0.60)$

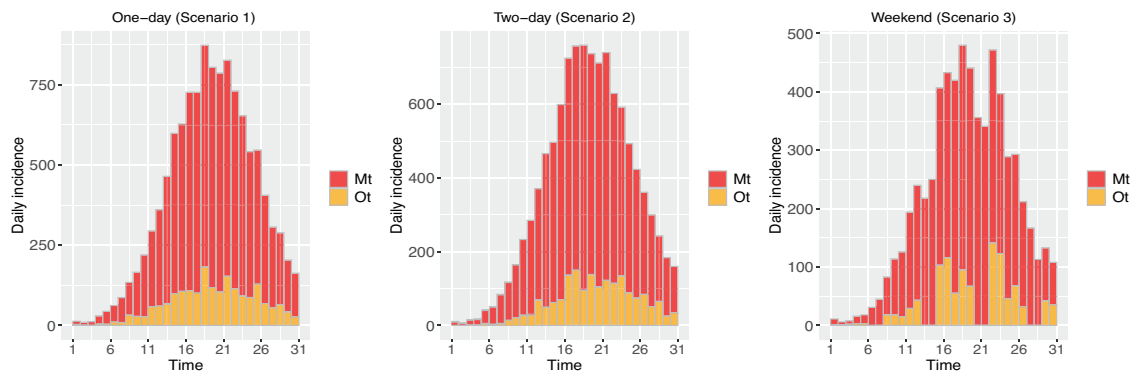


FIGURE 1 Stacked histograms of daily incidence for different delay patterns (Scenarios 1-3)

For each scenario, we report the mean estimate, empirical standard error (ESE), root mean square error (RMSE), and coverage probability for a 95% credible interval (CP95%) on the hyperparameters. The performance metric for R_t is taken to be the mean absolute error (MAE) defined as $MAE_{R_t} = S^{-1} \sum_{s=1}^S |\hat{R}_t^{(s)} - R_t|$. We also compare how LPS performs against the `estimate_R()` function of the EpiEstim package (Cori et al., 2013). In particular, we use the syntax `estimate_R(incid = Mtestim, method = "non_`

`parametric_si", config = make_config(list(si_distr = c(0,p))))`, where `Mtestim` corresponds to the (inflated) number of contaminations. Similarly, we use `incid = Observed` to obtain estimates based on the observed number of cases without an inflation factor, which is most commonly used in the literature. We choose the method `"non_parametric_si"` to specify the distribution of the serial interval, or as is the case here, the generation interval distribution denoted by p and inject the latter in the `make_config` option. Moreover, we keep the default option that estimates R_t on weekly sliding windows and uses the posterior mean as a point estimate. Table 2 summarizes the results related to the LPS hyperparameter estimates for Scenarios 1-3, and the left column of Figure 2 shows the estimated curves for R_t (gray) and the pointwise median of the $S = 500$ estimated curves obtained with LPS (dashed) and EpiEstim (dotted-dashed), respectively, using the simple inflation factor to estimate M_t . Figure 2 right column shows the MAE of R_t for days $t = 8, \dots, 30$ obtained with LPS (green) and the EpiEstim package with inflation factor on contaminations (light blue) and without inflation factor (dark blue). Estimates of the delay probabilities are relatively close to their true value with a coverage probability slightly above the 95% nominal value in most cases. As underreported data only convey limited information about the underlying transmission process, the credible intervals tend to be wide and conservative (i.e., overcoverage will be observed) especially when ρ is low. Mean estimates of the reporting probability are close to the true value in all scenarios with a more pronounced undercoverage under a two-day delay pattern (as in Azmon et al., 2014). It is also worth mentioning that the downward trend of R_t is well captured under the three considered delay patterns in Scenarios 1-3. Furthermore, Figure 2 (right column) shows that LPS is competitive against EpiEstim regarding the estimation of R_t . Simulation results for the hyperparameters in Scenarios 4-6 with LPS are given in Table 3. Again, the estimates are relatively close to their true value and the coverages are all reasonable. The undercoverage of certain delay probabilities in Scenario 4 can be explained by the complex shape of R_t and the fact that using a simple inflation factor to recover the latent number of contaminations may be too simplistic here. Figure 3 shows that the estimated R_t obtained with LPS captures the real underlying trend even with more complex structures such as in Scenario 4. In the latter scenario, the MAE of R_t is quite high with EpiEstim, while it remains reasonably low with LPS. In Scenario 6, neither EpiEstim nor LPS is able to capture

TABLE 2 Simulation results for the reporting and delay probabilities under Scenarios 1–3 with $S = 500$

Delay pattern	Parameter	True value	Mean	ESE	RMSE	CI.low	CI.up	CP95%
One-day (Scenario 1)	$\delta_{Mo \rightarrow Tu}$	0.4	0.283	0.089	0.147	0.029	0.660	100.0
	$\delta_{Tu \rightarrow We}$	0.5	0.383	0.081	0.142	0.072	0.710	100.0
	$\delta_{We \rightarrow Th}$	0.7	0.600	0.080	0.129	0.171	0.857	99.4
	$\delta_{Th \rightarrow Fr}$	0.3	0.169	0.068	0.148	0.004	0.615	100.0
	$\delta_{Fr \rightarrow Sa}$	0.4	0.270	0.074	0.149	0.030	0.628	100.0
	$\delta_{Sa \rightarrow Su}$	0.6	0.504	0.081	0.125	0.146	0.780	99.8
	$\delta_{Su \rightarrow Mo}$	0.5	0.421	0.092	0.121	0.078	0.748	100.0
	ρ	0.2	0.196	0.008	0.009	0.176	0.217	98.8
Two-day (Scenario 2)	$\delta_{Mo \rightarrow Tu}$	0.3	0.309	0.109	0.109	0.026	0.701	100.0
	$\delta_{Mo \rightarrow We}$	0.4	0.295	0.113	0.154	0.039	0.656	99.2
	$\delta_{Tu \rightarrow We}$	0.4	0.290	0.108	0.154	0.016	0.700	100.0
	$\delta_{Tu \rightarrow Th}$	0.3	0.216	0.089	0.123	0.018	0.582	100.0
	$\delta_{We \rightarrow Th}$	0.3	0.224	0.082	0.112	0.007	0.647	100.0
	$\delta_{We \rightarrow Fr}$	0.4	0.251	0.097	0.178	0.021	0.628	100.0
	$\delta_{Th \rightarrow Fr}$	0.5	0.415	0.096	0.128	0.052	0.773	100.0
	$\delta_{Th \rightarrow Sa}$	0.3	0.242	0.088	0.106	0.026	0.597	100.0
	$\delta_{Fr \rightarrow Sa}$	0.3	0.208	0.071	0.117	0.004	0.646	100.0
	$\delta_{Fr \rightarrow Su}$	0.4	0.299	0.108	0.148	0.041	0.647	99.4
	$\delta_{Sa \rightarrow Su}$	0.5	0.400	0.112	0.151	0.037	0.778	100.0
	$\delta_{Sa \rightarrow Mo}$	0.3	0.268	0.086	0.092	0.028	0.632	100.0
	$\delta_{Su \rightarrow Mo}$	0.3	0.257	0.092	0.101	0.011	0.677	100.0
	$\delta_{Su \rightarrow Tu}$	0.6	0.444	0.114	0.193	0.105	0.748	95.4
ρ	0.2	0.215	0.011	0.019	0.201	0.229	47.4	
Weekend (Scenario 3)	$\delta_{Mo \rightarrow Tu}$	0.4	0.403	0.138	0.138	0.000	0.923	100.0
	$\delta_{Tu \rightarrow We}$	0.5	0.187	0.076	0.322	0.002	0.676	99.6
	$\delta_{We \rightarrow Th}$	0.7	0.483	0.066	0.227	0.144	0.758	80.2
	$\delta_{Th \rightarrow Fr}$	0.3	0.288	0.086	0.087	0.035	0.653	100.0
	$\delta_{Fr \rightarrow Mo}$	0.4	0.447	0.108	0.118	0.074	0.780	100.0
	$\delta_{Sa \rightarrow Mo}$	0.6	0.534	0.093	0.114	0.000	0.971	100.0
	$\delta_{Su \rightarrow Mo}$	0.5	0.437	0.100	0.118	0.000	0.961	100.0
	ρ	0.2	0.183	0.001	0.017	0.165	0.203	76.8

the step function of R_t . This is as expected, since the case incidence data are subject to underreporting and delay and therefore cannot convey enough information to capture sharp drops in the reproduction number.

In terms of computational speed, the LPS approach is extremely fast. With an Intel Xeon E-2186M CPU running at 2.90 GHz, the elapsed real time for fitting the model with misreported data was recorded to be approximately 6 s for a one-day or weekend delay and a little bit more (around 10 s) for a two-day delay as the hyperparameter dimension is larger in the latter case. This is substantially less than any existing MCMC algorithm that typically takes minutes (or hours) to fit such complex epidemic models. Furthermore, as LPS does not rely on any sampling scheme, the additional burden of diagnostic checks (e.g., trace plots, Geweke statistics) is also avoided.

It is also worth mentioning that specifying the B-splines basis on the entire domain of the epidemic curve, that is, $\mathcal{T} = [0, T]$ implies that LPS provides a global smooth fit to the reproduction number in that domain, where the degree of smoothness is controlled by the smoothing parameter $\hat{\lambda}$. Applying LPS on a larger domain, say, $\tilde{\mathcal{T}} = [0, T + 7]$ (i.e., with an additional week of incidence data) will affect the estimated smoothing parameter, so that $\hat{\lambda}$ will typically differ from $\hat{\lambda}$, the smoothing parameter resulting from the fit in $\tilde{\mathcal{T}}$. As such, the estimates of R_t in the domain $[0, T]$ will also differ under \mathcal{T} and $\tilde{\mathcal{T}}$. The simulation study has shown that the trend of R_t is on average well captured by LPS (as confirmed by

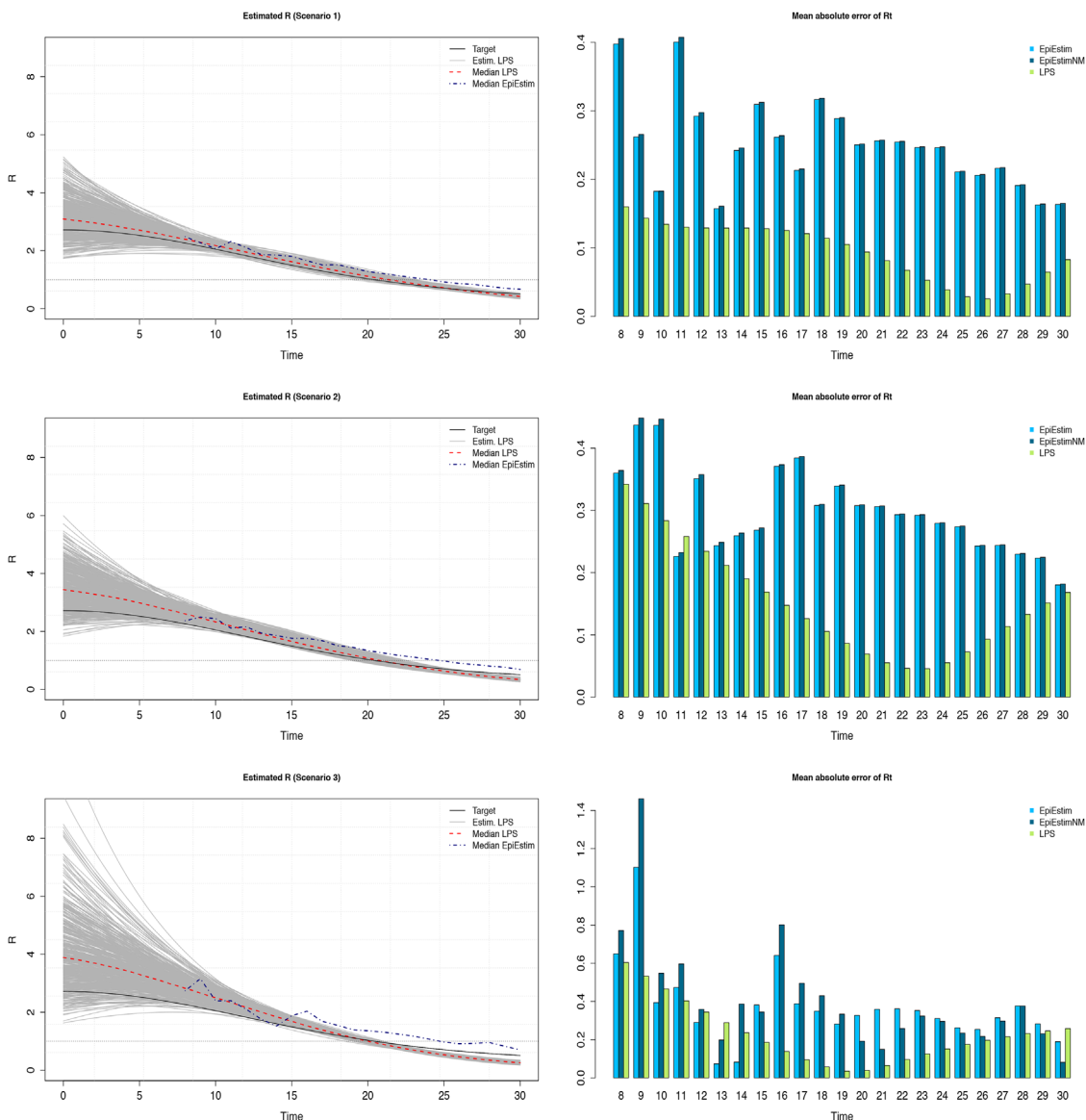


FIGURE 2 (Left column) Estimation of the time-varying reproduction number (gray curves) and pointwise median with LPS (dashed) and EpiEstim (dotted-dashed) for Scenarios 1–3. (Right column) MAE of R_t for days $t = 8, \dots, 30$ with LPS (green), EpiEstim (light blue) with multiplication factor, and EpiEstimNM (dark blue) ignoring the multiplication factor on contaminations.

the MAE values). Hence, even though LPS keeps changing estimates in the past, these changes will be characterized by fluctuations in a “close” neighborhood of the target and should therefore not be considered a limitation of our approach.

3.2 | Importance of prior information on ρ

When case incidence data is affected by underreporting and delay, prior information on the reporting rate ρ is of crucial importance. The simple multiplication rule given in Section 2.2.2 reflects our prior knowledge of the reporting rate and is used to approximate the daily latent number of cases M_t by multiplying the observed disease counts O_t by $\tilde{\rho}^{-1}$. We analyze the effect on the R_t estimate obtained with LPS resulting from a potential mismatch between ρ and $\tilde{\rho}$ by computing the absolute value of the average bias of R_t over days $t = 1, \dots, T$ under Scenario 1 and Scenario 3, respectively (cf. Section 3.1). The average bias of the reproduction number for a given couple $(\rho, \tilde{\rho})$ is computed as

$$\mathbb{B}_{(\rho, \tilde{\rho})}^R = \left| \frac{1}{T} \sum_{t=1}^T \left\{ \frac{1}{S} \sum_{s=1}^S \left(\hat{R}_t^{(\rho, \tilde{\rho})} - R_t \right) \right\} \right|, \tag{15}$$

TABLE 3 Simulation results for the reporting and delay probabilities under Scenarios 4–6 with $S = 500$

Delay pattern	Parameter	True value	Mean	ESE	RMSE	CI.low	CI.up	CP95%
One-day (Scenario 4)	$\delta_{Mo \rightarrow Tu}$	0.6	0.570	0.064	0.071	0.425	0.691	95.4
	$\delta_{Tu \rightarrow We}$	0.8	0.715	0.061	0.105	0.509	0.843	78.6
	$\delta_{We \rightarrow Th}$	0.3	0.097	0.047	0.209	0.011	0.348	89.6
	$\delta_{Th \rightarrow Fr}$	0.2	0.101	0.040	0.107	0.021	0.283	99.0
	$\delta_{Fr \rightarrow Sa}$	0.5	0.526	0.043	0.050	0.405	0.633	99.0
	$\delta_{Sa \rightarrow Su}$	0.7	0.817	0.050	0.128	0.651	0.904	85.8
	$\delta_{Su \rightarrow Mo}$	0.4	0.449	0.065	0.082	0.299	0.589	91.2
	ρ	0.6	0.606	0.012	0.014	0.578	0.633	98.0
Two-day (Scenario 5)	$\delta_{Mo \rightarrow Tu}$	0.4	0.287	0.117	0.163	0.046	0.615	98.2
	$\delta_{Mo \rightarrow We}$	0.5	0.539	0.122	0.128	0.155	0.820	99.8
	$\delta_{Tu \rightarrow We}$	0.7	0.541	0.122	0.201	0.074	0.864	98.6
	$\delta_{Tu \rightarrow Th}$	0.3	0.413	0.117	0.162	0.056	0.768	100.0
	$\delta_{We \rightarrow Th}$	0.2	0.252	0.091	0.105	0.002	0.735	99.8
	$\delta_{We \rightarrow Fr}$	0.4	0.272	0.118	0.174	0.017	0.681	99.2
	$\delta_{Th \rightarrow Fr}$	0.2	0.229	0.071	0.077	0.003	0.689	100.0
	$\delta_{Th \rightarrow Sa}$	0.2	0.287	0.096	0.129	0.010	0.718	100.0
	$\delta_{Fr \rightarrow Sa}$	0.6	0.432	0.072	0.183	0.018	0.840	99.8
	$\delta_{Fr \rightarrow Su}$	0.5	0.532	0.092	0.098	0.041	0.881	100.0
	$\delta_{Sa \rightarrow Su}$	0.4	0.435	0.086	0.093	0.009	0.864	100.0
	$\delta_{Sa \rightarrow Mo}$	0.3	0.244	0.074	0.093	0.011	0.655	99.8
	$\delta_{Su \rightarrow Mo}$	0.2	0.196	0.058	0.058	0.003	0.638	100.0
	$\delta_{Su \rightarrow Tu}$	0.1	0.178	0.089	0.118	0.006	0.576	99.6
ρ	0.7	0.767	0.038	0.077	0.661	0.841	93.2	
Weekend (Scenario 6)	$\delta_{Mo \rightarrow Tu}$	0.4	0.412	0.148	0.149	0.001	0.903	100.0
	$\delta_{Tu \rightarrow We}$	0.5	0.121	0.066	0.384	0.000	0.661	96.0
	$\delta_{We \rightarrow Th}$	0.7	0.483	0.101	0.239	0.094	0.803	93.2
	$\delta_{Th \rightarrow Fr}$	0.3	0.350	0.131	0.140	0.026	0.775	100.0
	$\delta_{Fr \rightarrow Mo}$	0.4	0.511	0.167	0.200	0.025	0.882	100.0
	$\delta_{Sa \rightarrow Mo}$	0.6	0.482	0.113	0.164	0.000	0.964	100.0
	$\delta_{Su \rightarrow Mo}$	0.5	0.451	0.100	0.112	0.000	0.966	100.0
	ρ	0.4	0.430	0.030	0.042	0.341	0.516	99.6

where $\hat{R}_t^{(\rho, \tilde{\rho})}$ denotes the estimated reproduction number with LPS at time point t under $\tilde{\rho}$ when the true underlying reporting rate is ρ . In Figure 4, each cell of the matrix corresponds to the (normalized) average bias of the reproduction number (i.e., we divided $\mathbb{B}_{(\rho, \tilde{\rho})}^R$ by the largest observed bias among all considered couples of ρ and $\tilde{\rho}$ in order to have values between 0 and 1) computed with the above formula and $S = 20$ replications. Results are intuitive and confirm the importance of prior information on ρ . The smallest biases are reached on (and alongside) the main diagonal, where the midpoint of the prior domain of the reporting rate, that is, $\tilde{\rho}$ is equal to (or close to) the true reporting rate from the data-generating process. The larger the discrepancy between $\tilde{\rho}$ and ρ , the smaller the precision with which LPS is able to estimate the instantaneous reproduction number in a setting accounting for misreporting. In addition, the bias is largest under the couple $\rho = 0.2$ and $\tilde{\rho} = 1$, that is, when the true reporting probability is small and prior information on ρ assumes that all cases are reported.

3.3 | Limitations of the LPS model for misreported data

A potential limitation of our LPS approach is that the daily reporting probability ρ is assumed time-homogeneous. In practice, during an epidemic outbreak, the reporting process is dynamic and changes over time. To highlight the limitation

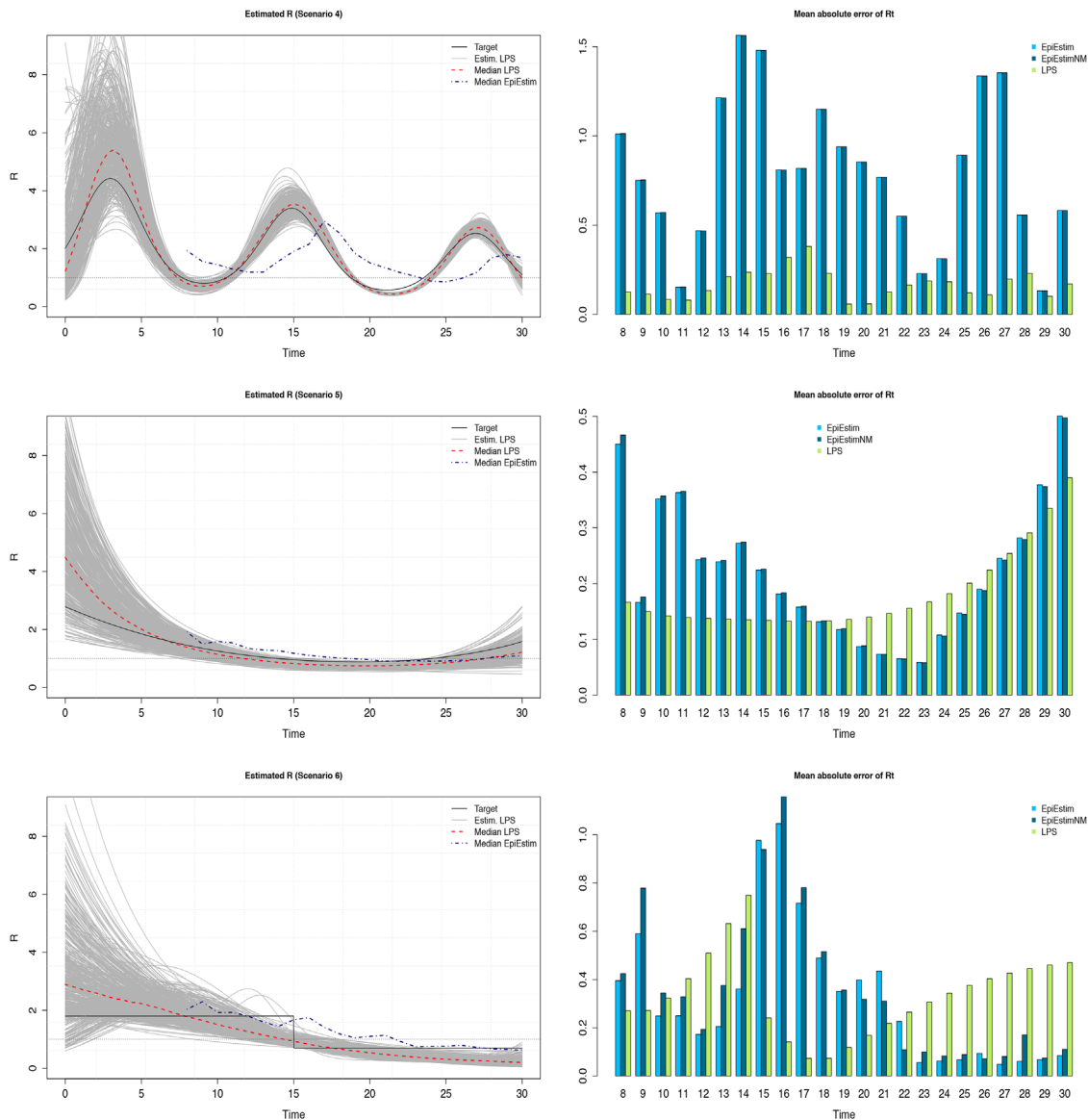


FIGURE 3 (Left column) Estimation of the time-varying reproduction number (gray curves) and pointwise median with LPS (dashed) and EpiEstim (dotted-dashed) for Scenarios 4–6. (Right column) MAE of R_t for days $t = 8, \dots, 30$ with LPS (green), EpiEstim (light blue) with multiplication factor, and EpiEstimNM (dark blue) ignoring the multiplication factor on contaminations.

of a constant reporting rate, we run the setting of Scenario 4 (cf. Section 3.1) with a data-generating process assuming a time-dependent reporting probability governed by an increasing step function $\rho_t = 0.2 \mathbb{1}(t < 8) + 0.4 \mathbb{1}(8 \leq t < 20) + 0.6 \mathbb{1}(t \geq 20)$ as represented in the left panel of Figure 5. The right panel of Figure 5 shows the median trajectory of the R_t curve computed over $S = 100$ replicated datasets. The nonnegligible bias is explained by a poor approximation of the daily latent number of cases M_t by the naive inflation factor approach of Section 2.2.2. The same overestimation phenomenon is observed in Azmon et al. (2014), where the effect of an increasing reporting parameter on estimation of R_t is assessed; yet the bias appears to be less pronounced with their MCMC approach. This reduced bias can be explained by the fact that MCMC takes into account the uncertainty surrounding the hyperparameter values by exploring the posterior space, while LPS simply relies on a MAP estimation scheme for η .

Another potential limitation is that LPS assumes a known and fixed generation interval distribution. The generation interval is less easily observed than the serial interval (the time elapsed between the commencement of symptoms in an infector–infectee pair) as it is hard to gather information on times of infection. Using serial intervals as a proxy for generation times is a potential source of bias when it comes to estimate the reproduction number (Britton & Scalia Tomba, 2019). Furthermore, even if information on generation times would be available, the generation interval is also time depen-

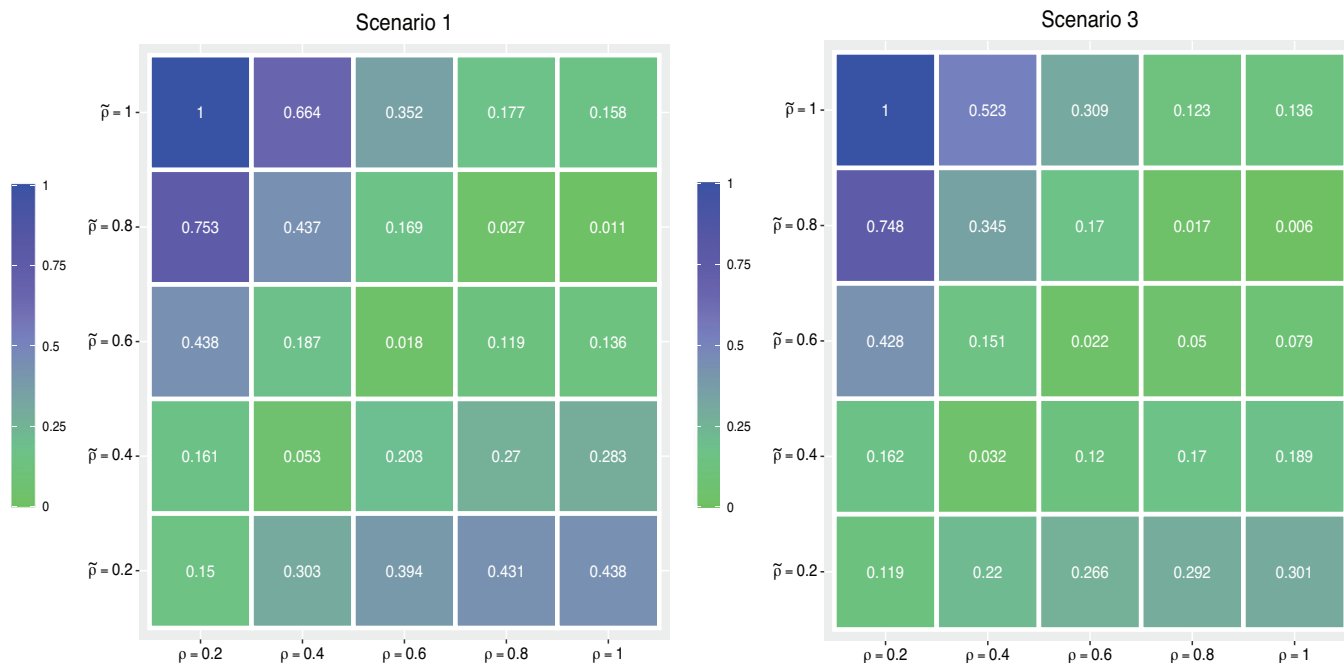


FIGURE 4 Normalized average bias of the reproduction number with LPS for different $\tilde{\rho}$ and ρ couples in Scenario 1 (left panel) and Scenario 3 (right panel). Smaller values on the main diagonal confirm the importance of prior information on ρ . A uniform prior on the reporting rate with midpoint $\tilde{\rho}$ too far from the true ρ will lead to biased estimates of R_t .

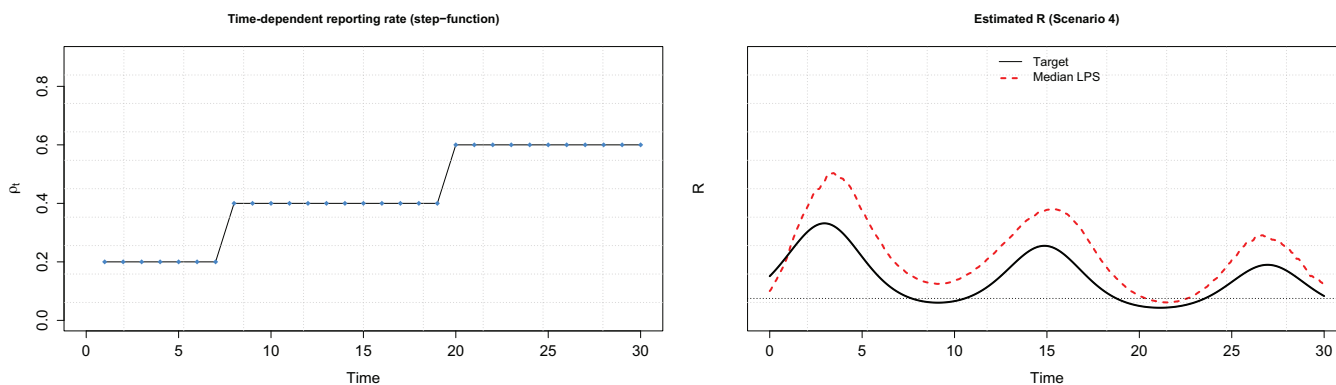


FIGURE 5 Illustration of the restrictive assumption of a constant reporting rate under LPS. A data-generating process governed by an increasing reporting rate (left panel) yields biased estimates of R_t .

dent and its dynamic is certainly influenced by the contagiousness of the variants of a given pathogen. Introducing such sources of heterogeneity and temporal dimensions within our LPS model for misreported data is of course challenging as the composite link structure and the renewal equation process already make the model complex.

4 | REAL DATA APPLICATIONS

4.1 | The 1918 influenza pandemic in Baltimore

The LPS methodology is first illustrated in the context of the 1918 H1N1 influenza pandemic in Baltimore, USA, with data obtained from the EpiEstim package (Cori et al., 2013). The dataset contains daily incidence of the onset of disease for a period of 92 days and a discrete daily distribution of the serial interval for influenza. We use the serial interval as a proxy for the generation interval. The serial interval for influenza given by EpiEstim is $\mathbf{p} =$

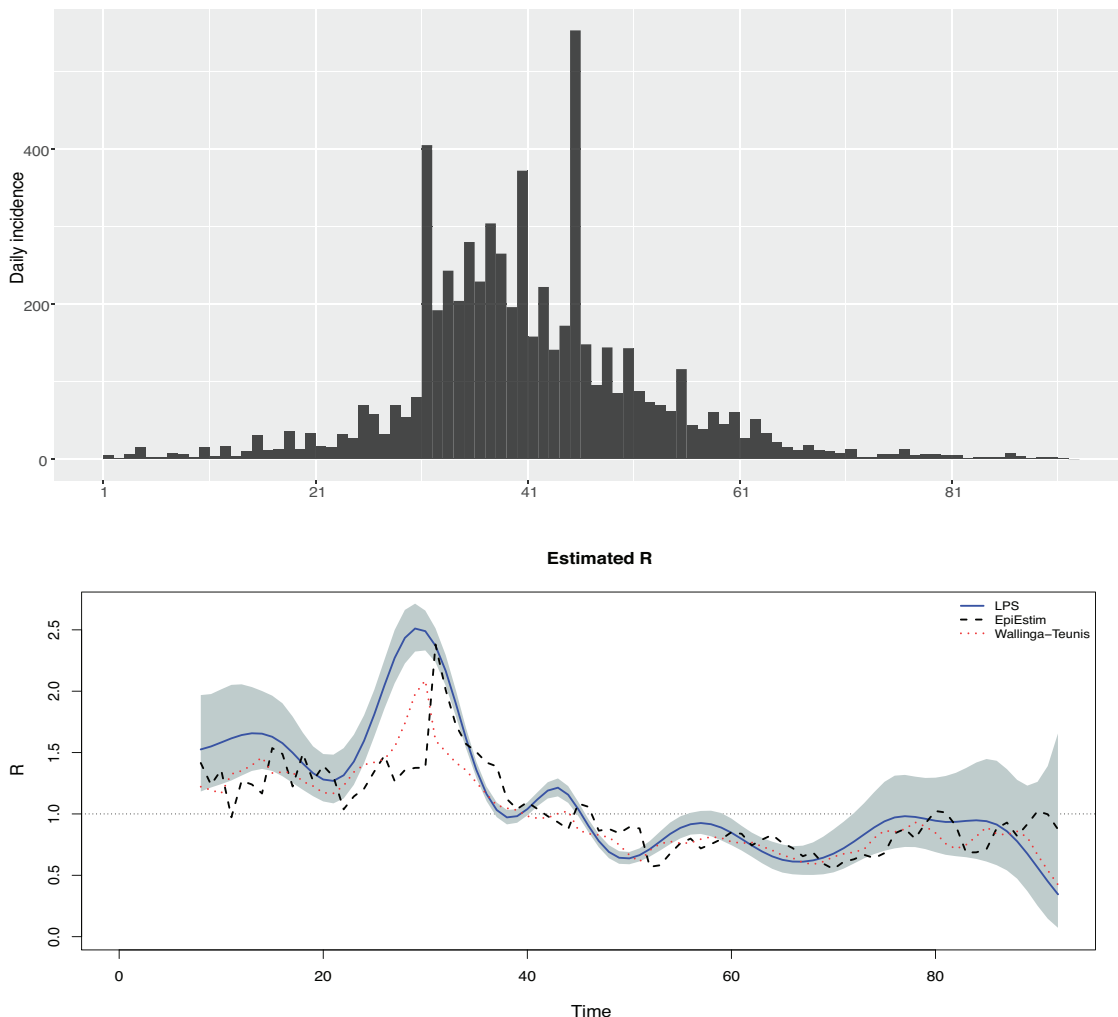


FIGURE 6 Daily incidence (top) and estimated R_t for the 1918 Influenza data in Baltimore with the EpiEstim (dashed), LPS (solid), and Wallinga–Teunis (dotted) method.

$\{0.233, 0.359, 0.198, 0.103, 0.053, 0.027, 0.014, 0.007, 0.003, 0.002, 0.001\}$. We assume that there is no underreporting (i.e., our prior assumption is such that $\tilde{\rho} = 1$) and that the delay process is governed by a one-day delay pattern. For a smooth estimation of the reproduction number, we use $K = 20$ (cubic) B-splines in $[0,92]$ and a third-order penalty to counterbalance the flexibility of the fitted curve. Figure 6 shows the daily incidence of the 1918 H1N1 data (top) and the estimated time-varying reproduction number (bottom) with LPS (solid), the `estimate_R()` routine of the EpiEstim package (dashed), and the Wallinga and Teunis (2004) method (dotted). The gray surface is the (approximate) 95% pointwise credible interval for R_t obtained with LPS. Around day $t = 30$, the estimated R_t reaches a peak before gradually decaying towards one, a pattern also observed in White and Pagano (2008). The reporting rate is estimated to be $\hat{\rho} = 0.886$ with a 95% credible interval $[0.832; 0.923]$.

4.2 | COVID-19 data for Australia

In a second application, we use LPS to estimate the time-varying reproduction number of COVID-19 hospitalizations in Australia between May and September 2020. The data are obtained from the **COVID19** package (Guidotti & Ardia, 2020). We use a discrete generation interval with a mean of 4 days and standard deviation of 2 days, namely $\mathbf{p} = \{0.053, 0.249, 0.297, 0.238, 0.163\}$ and estimate the model under a two-day delay structure with prior $\rho \sim \mathcal{U}(0.7, 0.8)$ and hence an inflation factor of $1/0.75$ on the observed number of cases. Figure 7 shows the daily incidence (top) and the estimated R_t under the two-day delay structure. There is a strong similarity between the estimated pattern of R_t for the three

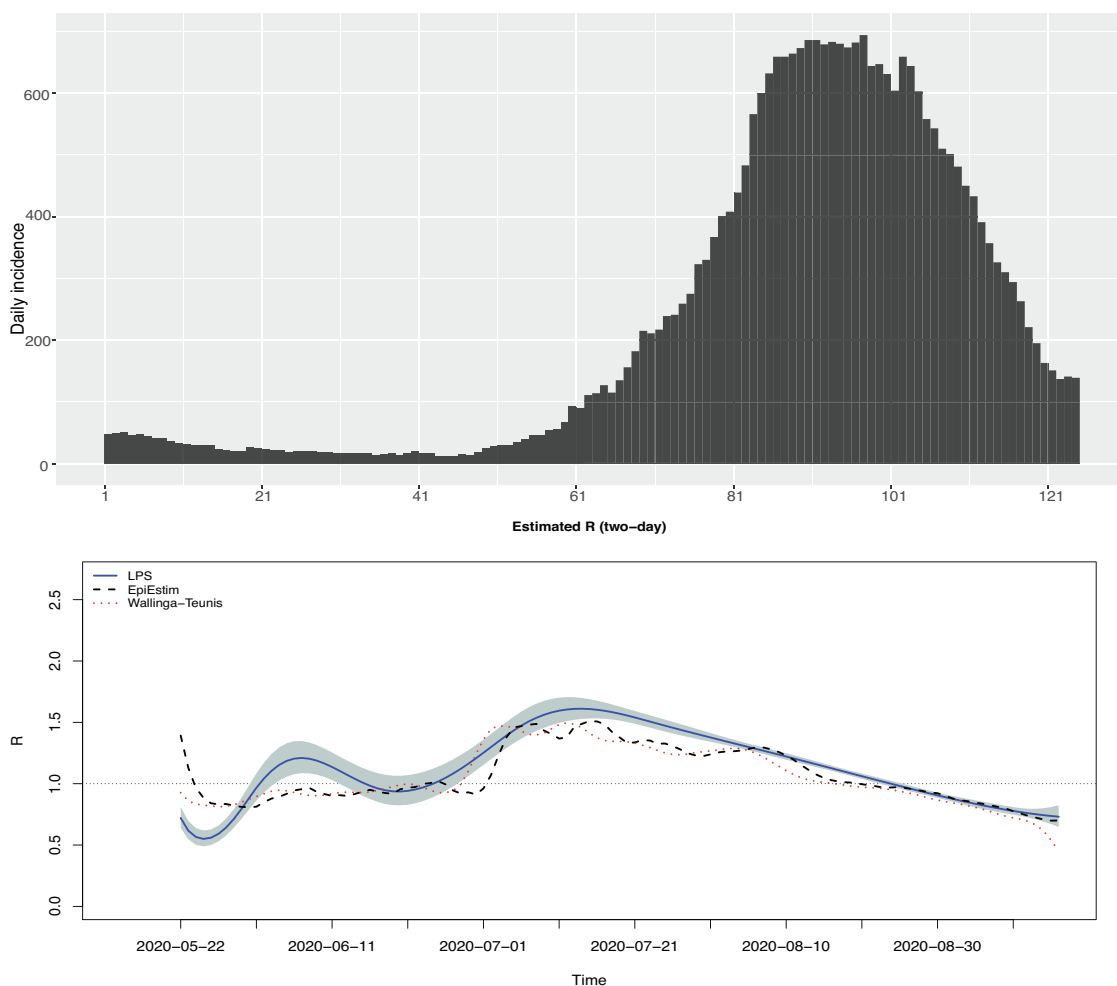


FIGURE 7 Daily incidence (top) and estimated R_t of COVID-19 for Australia between May and September 2020 with the EpiEstim (dashed), LPS (solid), and Wallinga–Teunis (dotted) method.

considered methods (LPS, EpiEstim, and Wallinga–Teunis). At the end of May, R_t is below one and increases during the next 2 months to reach a peak in July 2020. Then it slowly decreases to reach a value below one around the end of August.

5 | CONCLUDING REMARKS

The Laplacian-P-splines methodology presented in this paper combines Laplace approximation and Bayesian penalized B-splines for fast and flexible estimation of the time-varying reproduction number in an epidemic model with misreported data. The key benefit of our approach is its computational speed. While classic MCMC methods may take hours to deliver posterior estimates of key epidemiological parameters, estimation with LPS typically requires a couple of seconds. Provided minimal prior knowledge is available for the reporting probability (based, for instance, on historical data or serological studies), our results show that working with a simple multiplication rule on the observed set of disease counts provides satisfying estimates of the reproduction number.

This article shows that LPS performs at least as good as existing methods for estimation of R_t such as EpiEstim or the Wallinga–Teunis approach. Moreover, it allows for different specifications of the delay pattern (one-day, two-day, or weekend delays), covering practical scenarios arising in the real world during epidemic outbreaks. From here, several directions can be explored in the future to further improve the LPS methodology in the framework of epidemic modeling. First, it would be important to go beyond a naive multiplication factor approach to approximate the latent number of daily cases M_t . This will probably improve the accuracy of posterior estimates for R_t and for the reporting and delay probabilities. Second, instead of using the MAP estimator for the hyperparameters, an alternative (and also more costly) strategy would

be to use grid-based or MCMC approaches that would capture the uncertainty of posterior estimates more precisely and less locally than the MAP method considered here. Third, it would be interesting to refine the delay patterns and work for instance with ad hoc reporting structures that take into account public holidays. Finally, since the Poisson assumption imposed on the number of new (latent and observed) contaminations might underestimate variability in transmission (Imai et al., 2015), it would be relevant to extend our approach to account for underreporting under a negative binomial model to reflect a more flexible relationship between the mean and variance in infectious disease counts (Lloyd-Smith, 2007).

ACKNOWLEDGMENT

This project is funded by the European Union's Research and Innovation Action under the H2020 work programme, EpiPose (grant number 101003688).


CONFLICT OF INTEREST

The authors have declared no conflict of interest.

DATA AVAILABILITY STATEMENT

The data and simulation results that support the findings of this study are openly available at: https://github.com/oswaldogressani/RLPS_misreported_data

OPEN RESEARCH BADGES

 This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the [Supporting Information](#) section.

This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

ORCID

Oswaldo Gressani  <https://orcid.org/0000-0003-4152-6159>

REFERENCES

- Abrams, S., Wambua, J., Santermans, E., Willem, L., Kuylen, E., Coletti, P., Libin, P., Faes, C., Petrof, O., Herzog, S. A., Beutels, P., & Hens, N. (2021). Modelling the early phase of the Belgian COVID-19 epidemic using a stochastic compartmental model and studying its implied future trajectories. *Epidemics*, 35, 100449.
- Azmon, A., Faes, C., & Hens, N. (2014). On the estimation of the reproduction number based on misreported epidemic data. *Statistics in Medicine*, 33(7), 1176–1192. <https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.6015>
- Bettencourt, L. M., & Ribeiro, R. M. (2008). Real time Bayesian estimation of the epidemic potential of emerging infectious diseases. *PLoS ONE*, 3(5), e2185.
- Bracher, J., & Held, L. (2021). A marginal moment matching approach for fitting endemic-epidemic models to underreported disease surveillance counts. *Biometrics*, 77, 1202–1214.
- Britton, T., & Scalia Tomba, G. (2019). Estimation in emerging epidemics: Biases and remedies. *Journal of the Royal Society Interface*, 16(150), 20180670.
- Cauchemez, S., Boëlle, P.-Y., Thomas, G., & Valleron, A.-J. (2006). Estimating in real time the efficacy of measures to control emerging communicable diseases. *American Journal of Epidemiology*, 164(6), 591–597.
- Cori, A., Ferguson, N. M., Fraser, C., & Cauchemez, S. (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics. *American Journal of Epidemiology*, 178(9), 1505–1512.
- Cui, J., & Kaldor, J. (1998). Changing pattern of delays in reporting AIDS diagnoses in Australia. *Australian and New Zealand Journal of Public Health*, 22(4), 432–435.
- Eilers, P. H. C. (2007). Ill-posed problems with counts, the composite link model and penalized likelihood. *Statistical Modelling*, 7(3), 239–254.
- Eilers, P. H. C., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2), 89–102.
- Feller, W. (1941). On the integral equation of renewal theory. *The Annals of Mathematical Statistics*, 12(3), 243–267.
- Ferguson, N. M., Cummings, D. A., Cauchemez, S., Fraser, C., Riley, S., Meeyai, A., Iamsirithaworn, S., & Burke, D. S. (2005). Strategies for containing an emerging influenza pandemic in southeast Asia. *Nature*, 437(7056), 209–214.
- Fraser, C. (2007). Estimating individual and household reproduction numbers in an emerging epidemic. *PLoS ONE*, 2(8), e758.
- Fraser, C., Donnelly, C. A., Cauchemez, S., Hanage, W. P., Van Kerkhove, M. D., Hollingsworth, T. D., Griffin, J., Baggaley, R. F., Jenkins, H. E., Lyons, E. J., Jombart, T., Hinsley, W. R., Grassly, N. C., Balloux, F., Ghani, A. C., Ferguson, N. M., Rambaut, A., Pybus, O. G., Lopez-Gatell, H., ... Hugonnet, S. (2009). Pandemic potential of a strain of influenza A (H1N1): Early findings. *Science*, 324(5934), 1557–1561.

- Goldfeld, S. M., Quandt, R. E., & Trotter, H. F. (1966). Maximization by quadratic hill-climbing. *Econometrica*, *34*(3), 541–551.
- Gómez-Rubio, V., & Rue, H. (2018). Markov chain Monte Carlo with the integrated nested Laplace approximation. *Statistics and Computing*, *28*, 1033–1051.
- Gostic, K. M., McGough, L., Baskerville, E. B., Abbott, S., Joshi, K., Tedijanto, C., Kahn, R., Niehus, R., Hay, J. A., De Salazar, P. M., Hellewell, J., Meakin, S., Munday, J. D., Bosse, N. I., Sherratt, K., Thompson, R. N., White, L. F., Huisman, J. S., Scire, J., ... Cobey, S. (2020). Practical considerations for measuring the effective reproductive number. *PLOS Computational Biology*, *16*(12), e1008409.
- Gressani, O., & Lambert, P. (2018). Fast Bayesian inference using Laplace approximations in a flexible promotion time cure model based on P-splines. *Computational Statistics & Data Analysis*, *124*, 151–167.
- Gressani, O., & Lambert, P. (2021). Laplace approximations for fast Bayesian inference in generalized additive models based on P-splines. *Computational Statistics & Data Analysis*, *154*, 107088.
- Gressani, O., Wallinga, J., Althaus, C. L., Hens, N., & Faes, C. (2022). Epilps: A fast and flexible Bayesian tool for estimation of the time-varying reproduction number. *PLoS Computational Biology*, *18*(10), e1010618.
- Guidotti, E., & Ardia, D. (2020). Covid-19 data hub. *Journal of Open Source Software*, *5*(51), 2376.
- Heesterbeek, H., Anderson, R. M., Andreasen, V., Bansal, S., De Angelis, D., Dye, C., Eames, K. T. D., Edmunds, W. J., Frost, S. D. W., Funk, S., Hollingsworth, T. D., House, T., Isham, V., Klepac, P., Lessler, J., Lloyd-Smith, J. O., Metcalf, C. J. E., Mollison, D., Pellis, L., ... Viboud, C. (2015). Modeling infectious disease dynamics in the complex landscape of global health. *Science*, *347*(6227), aaa4339.
- Hens, N., Van Ranst, M., Aerts, M., Robesyn, E., Van Damme, P., & Beutels, P. (2011). Estimating the effective reproduction number for pandemic influenza from notification data made publicly available in real time: A multi-country analysis for influenza A/H1N1v 2009. *Vaccine*, *29*(5), 896–904.
- Hethcote, H. W. (2000). The mathematics of infectious diseases. *SIAM Review*, *42*(4), 599–653.
- Imai, C., Armstrong, B., Chalabi, Z., Mangtani, P., & Hashizume, M. (2015). Time series regression model for infectious disease and weather. *Environmental Research*, *142*, 319–327.
- Jandarov, R., Haran, M., Bjørnstad, O., & Grenfell, B. (2014). Emulating a gravity model to infer the spatiotemporal dynamics of an infectious disease. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *63*(3), 423–444.
- Lambert, P., & Eilers, P. H. C. (2005). Bayesian proportional hazards model with time-varying regression coefficients: A penalized Poisson regression approach. *Statistics in Medicine*, *24*(24), 3977–3989.
- Lang, S., & Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, *13*(1), 183–212.
- Lawless, J. (1994). Adjustments for reporting delays and the prediction of occurred but not reported events. *Canadian Journal of Statistics*, *22*(1), 15–31.
- Levenberg, K. (1944). A method for the solution of certain non-linear problems in least squares. *Quarterly of Applied Mathematics*, *2*(2), 164–168.
- Lloyd-Smith, J. O. (2007). Maximum likelihood estimation of the negative binomial dispersion parameter for highly overdispersed data, with applications to infectious diseases. *PLoS ONE*, *2*(2), e180.
- Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*(4), 325–337.
- Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, *11*(2), 431–441.
- Martins, T. G., & Rue, H. (2014). Extending integrated nested Laplace approximation to a class of near-Gaussian latent models. *Scandinavian Journal of Statistics*, *41*(4), 893–912.
- Nishiura, H., & Chowell, G. (2009). The effective reproduction number as a prelude to statistical estimation of time-dependent epidemic trends. In *Mathematical and statistical estimation approaches in epidemiology* (pp. 103–121). Springer.
- Nouvellet, P., Cori, A., Garske, T., Blake, I. M., Dorigatti, I., Hinsley, W., Jombart, T., Mills, H. L., Nedjati-Gilani, G., Van Kerkhove, M. D., Fraser, C., Donnelly, C. A., Ferguson, N. M., & Riley, S. (2018). A simple approach to measure transmissibility and forecast incidence. *Epidemics*, *22*, 29–35.
- Plummer, M., et al. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*. Vienna, Austria.
- Riou, J., Poletto, C., & Boëlle, P.-Y. (2018). Improving early epidemiological assessment of emerging Aedes-transmitted epidemics using historical data. *PLOS Neglected Tropical Diseases*, *12*(6), e0006526.
- Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *71*(2), 319–392.
- Stocks, T., Britton, T., & Höhle, M. (2020). Model selection and parameter estimation for dynamic epidemic models via iterated filtering: Application to rotavirus in Germany. *Biostatistics*, *21*(3), 400–416.
- Svensson, Å. (2007). A note on generation times in epidemic models. *Mathematical Biosciences*, *208*(1), 300–311.
- Thompson, R., & Baker, R. (1981). Composite link functions in generalized linear models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *30*(2), 125–131.
- Thompson, R., Stockwin, J., van Gaalen, R., Polonsky, J., Kamvar, Z., Demarsh, P., Dahlgvist, E., Li, S., Miguel, E., Jombart, T., Lessler, J., Cauchemez, S., & Cori, A. (2019). Improved inference of time-varying reproduction numbers during infectious disease outbreaks. *Epidemics*, *29*, 100356.
- Tierney, L., & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, *81*(393), 82–86.

Vanhatalo, J., Riihimäki, J., Hartikainen, J., Jylänki, P., Tolvanen, V., & Vehtari, A. (2013). Gpstuff: Bayesian modeling with Gaussian processes. *Journal of Machine Learning Research*, 14, 1175–1179.

Wallinga, J., & Lipsitch, M. (2007). How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B: Biological Sciences*, 274(1609), 599–604.

Wallinga, J., & Teunis, P. (2004). Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of Epidemiology*, 160(6), 509–516.

White, L. F., & Pagano, M. (2008). Transmissibility of the influenza virus in the 1918 pandemic. *PLOS ONE*, 3(1), e1498.

Zhao, S., Musa, S. S., Lin, Q., Ran, J., Yang, G., Wang, W., Lou, Y., Yang, L., Gao, D., He, D., & Wang, M. H. (2020). Estimating the unreported number of novel coronavirus (2019-ncov) cases in China in the first half of January 2020: A data-driven modelling analysis of the early outbreak. *Journal of Clinical Medicine*, 9(2), 388.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Gressani, O., Faes, C., & Hens, N. (2023). An approximate Bayesian approach for estimation of the instantaneous reproduction number under misreported epidemic data. *Biometrical Journal*, 65, 2200024. <https://doi.org/10.1002/bimj.202200024>

APPENDIX A

Composition matrix for a one-day delay

$$C = \begin{pmatrix} 1 - \delta_{Mo \rightarrow Tu} & \delta_{Mo \rightarrow Tu} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 - \delta_{Tu \rightarrow We} & \delta_{Tu \rightarrow We} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 - \delta_{We \rightarrow Th} & \delta_{We \rightarrow Th} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 - \delta_{Th \rightarrow Fr} & \delta_{Th \rightarrow Fr} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 - \delta_{Fr \rightarrow Sa} & \delta_{Fr \rightarrow Sa} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 - \delta_{Sa \rightarrow Su} & \delta_{Sa \rightarrow Su} & 0 \\ \delta_{Su \rightarrow Mo} & 0 & 0 & 0 & 0 & 0 & 0 & 1 - \delta_{Su \rightarrow Mo} \end{pmatrix}. \tag{A.1}$$

Composition matrix for a two-day delay

$$C = \begin{pmatrix} 1 - \delta_{Mo \rightarrow Tu} - \delta_{Mo \rightarrow We} & \delta_{Mo \rightarrow Tu} & \delta_{Mo \rightarrow We} & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 - \delta_{Tu \rightarrow We} - \delta_{Tu \rightarrow Th} & \delta_{Tu \rightarrow We} & \delta_{Tu \rightarrow Th} & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \delta_{Su \rightarrow Mo} & \delta_{Su \rightarrow Tu} & \vdots & \vdots & \vdots & \vdots & \vdots & 1 - \delta_{Su \rightarrow Mo} - \delta_{Su \rightarrow Tu} \end{pmatrix}. \tag{A.2}$$

Composition matrix for a weekend delay

$$C = \begin{pmatrix} 1 - \delta_{Mo \rightarrow Tu} & \delta_{Mo \rightarrow Tu} & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 - \delta_{Tu \rightarrow We} & \delta_{Tu \rightarrow We} & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 - \delta_{We \rightarrow Th} & \delta_{We \rightarrow Th} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 - \delta_{Th \rightarrow Fr} & \delta_{Th \rightarrow Fr} & 0 & 0 \\ \delta_{Fr \rightarrow Mo} & 0 & 0 & 0 & 1 - \delta_{Fr \rightarrow Mo} & 0 & 0 \\ 1 - \delta_{Sa \rightarrow Mo} & \delta_{Sa \rightarrow Mo} & 0 & 0 & 0 & 0 & 0 \\ 1 - \delta_{Su \rightarrow Mo} & \delta_{Su \rightarrow Mo} & 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{A.3}$$

15214036, 2023, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bimj.202200024 by Universiteit Hasselt, Wiley Online Library on [25/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

APPENDIX B

The second-order Taylor expansion of $g_t(\theta, \eta)$ around an initial vector $\theta^{(0)}$ (e.g., a vector of ones) is written as

$$\begin{aligned} g_t(\theta, \eta) &\approx g_t(\theta^{(0)}, \eta) + (\theta - \theta^{(0)})^T \nabla g_t(\theta, \eta)|_{\theta=\theta^{(0)}} + \frac{1}{2}(\theta - \theta^{(0)})^T \nabla^2 g_t(\theta, \eta)|_{\theta=\theta^{(0)}} (\theta - \theta^{(0)}) \\ &\approx c + \theta^T \left(\nabla g_t(\theta, \eta)|_{\theta=\theta^{(0)}} - \nabla^2 g_t(\theta, \eta)|_{\theta=\theta^{(0)}} \theta^{(0)} \right) + \frac{1}{2} \theta^T \nabla^2 g_t(\theta, \eta)|_{\theta=\theta^{(0)}} \theta, \end{aligned} \quad (\text{B.1})$$

where c is a constant that does not depend on θ . An analytical version of (B.1) is obtained by computing the following gradient and Hessian matrix:

$$\begin{aligned} \nabla g_t(\theta, \eta)|_{\theta=\theta^{(0)}} &= \left(\frac{\partial g_t(\theta, \eta)}{\partial \theta_1}, \dots, \frac{\partial g_t(\theta, \eta)}{\partial \theta_K} \right)^T_{\theta=\theta^{(0)}}, \\ \nabla^2 g_t(\theta, \eta)|_{\theta=\theta^{(0)}} &= \frac{\partial^2 g_t(\theta, \eta)}{\partial \theta \partial \theta^T} \Big|_{\theta=\theta^{(0)}} = \begin{pmatrix} \frac{\partial^2 g_t(\theta, \eta)}{\partial \theta_1^2} & \frac{\partial^2 g_t(\theta, \eta)}{\partial \theta_1 \partial \theta_2} & \dots & \frac{\partial^2 g_t(\theta, \eta)}{\partial \theta_1 \partial \theta_K} \\ \vdots & \ddots & & \vdots \\ \frac{\partial^2 g_t(\theta, \eta)}{\partial \theta_K \partial \theta_1} & \frac{\partial^2 g_t(\theta, \eta)}{\partial \theta_K \partial \theta_2} & \dots & \frac{\partial^2 g_t(\theta, \eta)}{\partial \theta_K^2} \end{pmatrix}_{\theta=\theta^{(0)}}. \end{aligned} \quad (\text{B.2})$$

Gradient: Recall that $g_t(\theta, \eta) = O_t \log(s_t(\theta, \eta)) - s_t(\theta, \eta)$, so the derivative with respect to the k th B-spline coefficient is

$$\begin{aligned} \frac{\partial g_t(\theta, \eta)}{\partial \theta_k} &= O_t \frac{\partial s_t(\theta, \eta)}{\partial \theta_k} s_t(\theta, \eta)^{-1} - \frac{\partial s_t(\theta, \eta)}{\partial \theta_k}, \quad k = 1, \dots, K, \\ \text{with } \frac{\partial s_t(\theta, \eta)}{\partial \theta_k} &= \rho \sum_{i=1}^t \delta_{i \rightarrow t} \exp \left(\sum_{k=1}^K \theta_k b_k(i) \right) \left(\sum_s p_s \tilde{M}_{i-s} \right) b_k(i), \quad k = 1, \dots, K. \end{aligned} \quad (\text{B.3})$$

Hessian: To obtain the $K \times K$ Hessian matrix, the following second-order partial derivatives for $k, l = 1, \dots, K$ are computed:

$$\begin{aligned} \frac{\partial^2 g_t(\theta, \eta)}{\partial \theta_k \partial \theta_l} &= O_t \left(\frac{\partial^2 s_t(\theta, \eta)}{\partial \theta_k \partial \theta_l} s_t(\theta, \eta) - \frac{\partial s_t(\theta, \eta)}{\partial \theta_k} \frac{\partial s_t(\theta, \eta)}{\partial \theta_l} \right) s_t(\theta, \eta)^{-2} - \frac{\partial^2 s_t(\theta, \eta)}{\partial \theta_k \partial \theta_l}, \\ \text{with } \frac{\partial^2 s_t(\theta, \eta)}{\partial \theta_k \partial \theta_l} &= \rho \sum_{i=1}^t \delta_{i \rightarrow t} \exp \left(\sum_{k=1}^K \theta_k b_k(i) \right) \left(\sum_s p_s \tilde{M}_{i-s} \right) b_k(i) b_l(i). \end{aligned} \quad (\text{B.4})$$

Having computed the gradient and Hessian for all day indexes of the epidemic $t = 1, \dots, T$, the results are summed up to compute the gradient and Hessian of the log-likelihood (across all observations), namely $\nabla g(\theta, \eta)|_{\theta=\theta^{(0)}} := \sum_{t=1}^T \nabla g_t(\theta, \eta)|_{\theta=\theta^{(0)}}$ and $\nabla^2 g(\theta, \eta)|_{\theta=\theta^{(0)}} := \sum_{t=1}^T \nabla^2 g_t(\theta, \eta)|_{\theta=\theta^{(0)}}$. Using the second-order Taylor expansion in (B.1) (omitting the constant) and the log-likelihood function, we find

$$\begin{aligned} \ell(\theta, \eta; \mathcal{D}) &= \sum_{t=1}^T g_t(\theta, \eta) \\ &\approx \theta^T \left(\sum_{t=1}^T \nabla g_t(\theta, \eta)|_{\theta=\theta^{(0)}} - \sum_{t=1}^T \nabla^2 g_t(\theta, \eta)|_{\theta=\theta^{(0)}} \theta^{(0)} \right) + \frac{1}{2} \theta^T \sum_{t=1}^T \nabla^2 g_t(\theta, \eta)|_{\theta=\theta^{(0)}} \theta \\ &\approx \theta^T \left(\nabla g(\theta, \eta)|_{\theta=\theta^{(0)}} - \nabla^2 g(\theta, \eta)|_{\theta=\theta^{(0)}} \theta^{(0)} \right) + \frac{1}{2} \theta^T \nabla^2 g(\theta, \eta)|_{\theta=\theta^{(0)}} \theta. \end{aligned} \quad (\text{B.5})$$

Plugging (B.5) in (8) and rearranging the terms, one gets the Laplace approximation to the conditional posterior of the vector of B-spline parameters:

$$\begin{aligned} \tilde{p}_G(\boldsymbol{\theta}|\boldsymbol{\eta}, D) \propto \exp\left(-\frac{1}{2}\boldsymbol{\theta}^T\left(\lambda P - \nabla^2 g(\boldsymbol{\theta}, \boldsymbol{\eta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)}}\right)\boldsymbol{\theta}\right. \\ \left. + \boldsymbol{\theta}^T\left(\nabla g(\boldsymbol{\theta}, \boldsymbol{\eta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)}} - \nabla^2 g(\boldsymbol{\theta}, \boldsymbol{\eta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)}}\boldsymbol{\theta}^{(0)}\right)\right). \end{aligned} \quad (\text{B.6})$$

Note that (B.6) is (up to a multiplicative constant) a Gaussian density with mean (mode) and variance–covariance matrix equal to

$$\begin{aligned} \boldsymbol{\theta}^{(1)} &= (\lambda P - \nabla^2 g(\boldsymbol{\theta}, \boldsymbol{\eta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)}})^{-1} (\nabla g(\boldsymbol{\theta}, \boldsymbol{\eta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)}} - \nabla^2 g(\boldsymbol{\theta}, \boldsymbol{\eta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)}}\boldsymbol{\theta}^{(0)}), \\ \boldsymbol{\Sigma}^{(1)} &= (\lambda P - \nabla^2 g(\boldsymbol{\theta}, \boldsymbol{\eta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)}})^{-1}, \end{aligned} \quad (\text{B.7})$$

where the mean (mode) is obtained by solving the equation $\nabla_{\boldsymbol{\theta}} \log \tilde{p}_G(\boldsymbol{\theta}|\boldsymbol{\eta}, D) = 0$ for $\boldsymbol{\theta}$ and the variance–covariance matrix is $(-\nabla_{\boldsymbol{\theta}}^2 \log \tilde{p}_G(\boldsymbol{\theta}|\boldsymbol{\eta}, D))^{-1}$. Let $\boldsymbol{\theta}^*(\boldsymbol{\eta})$ and $\boldsymbol{\Sigma}^*(\boldsymbol{\eta})$ denote the mode and variance–covariance matrix towards which the iterative Laplace approximation scheme for $p(\boldsymbol{\theta}|\boldsymbol{\eta}, D)$ has converged for a given vector of hyperparameters $\boldsymbol{\eta}$. The final Laplace approximation is written (by abuse of notation) as

$$\tilde{p}_G(\boldsymbol{\theta}|\boldsymbol{\eta}, D) = \mathcal{N}_{\dim(\boldsymbol{\theta})}(\boldsymbol{\theta}^*(\boldsymbol{\eta}), \boldsymbol{\Sigma}^*(\boldsymbol{\eta})). \quad (\text{B.8})$$