Supplementary Materials S1: Supplementary Text

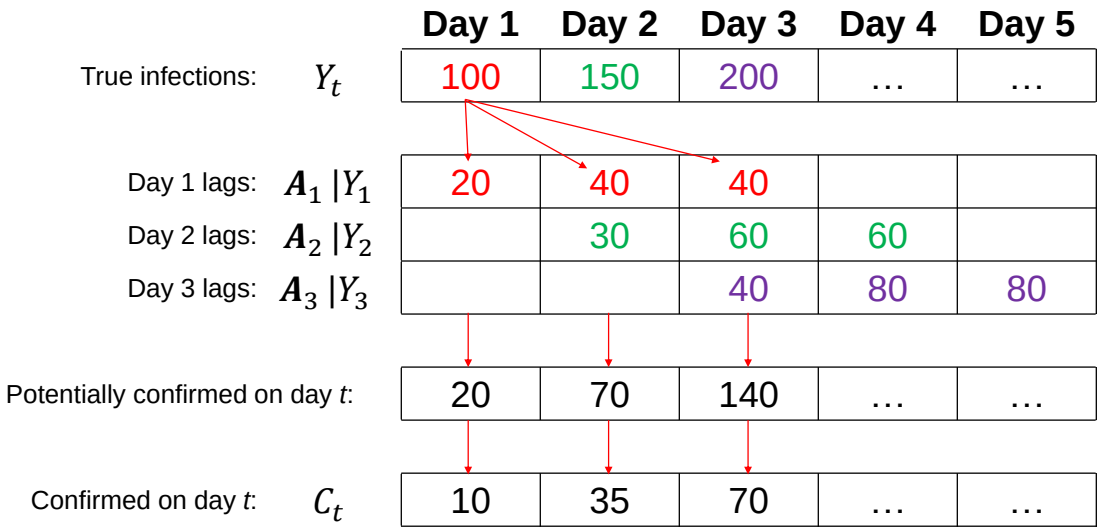
Regression Models for COVID-19 Epidemic Dynamics with Incomplete Data

Corbin Quick, Rounak Dey, and Xihong Lin
Harvard University

October 20, 2021

Contents

A Supplementary Table 1: Illustration of key variables



Above, the first row shows numbers of new infections Y_t on each day t. The second, third, and fourth rows show infections stratified by the day on which they are potentially confirmed, denoted by $\boldsymbol{A}_t$. The fifth row shows the total number of potentially confirmed infections on each day, denoted by M_t. The final row shows the number of infections actually confirmed on each day, denoted C_t, where $C_t < M_t$ indicates under-ascertainment.

Numbers above were chosen for illustration; under our assumed model, the observed fractions delayed and ascertained vary stochastically as described in the main text.

B Laplace approximations for the E step

B.1 Definitions and notations

Here we derive the Laplace approximations used in the E step of the EM algorithm. First, we review the notations used in the main text for a single region i with population size n_i observed across days $t = 1, 2, ..., T_i$.

- $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iT_i})$ denotes the numbers of new infections.
- $\boldsymbol{\Lambda}_i = (\Lambda_{i1}, \ldots, \Lambda_{iT_i})$ denotes the infection potentials.

- $\boldsymbol{R}_i = (R_{i1}, \ldots, R_{iT_i})$ denotes the effective reproduction numbers.
- $A_{i,t,k}$ is the number of individuals that were infected on day t and potentially confirmed on day $t+k$, and we denote $\boldsymbol{A}_i = (A_{i10}, \ldots, A_{iT_i0}, \ldots, A_{i1m_A}, \ldots, A_{iT_im_A})$.
- $\boldsymbol{M}_i = (M_{i1}, \ldots, M_{iT_i})$ denotes the numbers of individuals potentially confirmed.
- $\boldsymbol{C}_i = (C_{i1}, \ldots, C_{iT_i})$ denotes the numbers of individuals actually confirmed.
- $\boldsymbol{N}_i = (N_{i1}, \ldots, N_{iJ_i})$ denote the number of individuals given antibody tests (seroprevalence sample size) at survey periods $j = 1, \cdots, J_i$.
- $\boldsymbol{K}_i = (K_{i1}, \ldots, K_{iJ_i})$ denotes the numbers of individuals who tested positive in each seroprevalence survey period.

The vectors $\boldsymbol{Y}_i$ and $\boldsymbol{M}_i$ can be related to $\boldsymbol{A}_i$ by the equations,

$$\boldsymbol{Y}_i = \mathbf{P}_i \boldsymbol{A}_i, \quad \boldsymbol{M}_i = \mathbf{L}_i \boldsymbol{A}_i,$$

where $\mathbf{P}_i = \mathbf{1}_{m_A+1}^\top \otimes I_{T_i}$ is a $T_i \times T_i(m_A+1)$ matrix, $\mathbf{1}_{m_A+1}$ is the $(m_A+1)\times 1$ vector of all ones, I_{T_i} is the identity matrix of order T_i, and $\otimes$ is the Kronecker product. The $T_i \times T_i(m_A+1)$ matrix $\mathbf{L}_i$ is given by the first T_i rows of the $(T_i+m_A)\times T_i(m_A+1)$ matrix $\mathbf{L}_i^* = \begin{bmatrix} \mathbf{L}_{i1}^* & \mathbf{L}_{i2}^* & \ldots & \mathbf{L}_{i(m_A+1)}^* \end{bmatrix}$, where $\mathbf{L}_{ik}^*$-s are matrices of the following structure,

$$\mathbf{L}_{ik}^* = \begin{bmatrix} \mathbf{0}_{(k-1)\times T_i} \\ I_{T_i} \\ \mathbf{0}_{(m_A-k+1)\times T_i} \end{bmatrix}_{(T_i+m_A)\times T_i}.$$

We express the infection potentials as $\boldsymbol{\Lambda}_i = \mathbf{W}_i \boldsymbol{Y}_i + \boldsymbol{\lambda}_{i0}$, where $\boldsymbol{\lambda}_{i0} = y_{i\emptyset}(\mathbf{I} - \mathbf{W}_i)\mathbf{1}_{T_i\times 1}$ captures unobserved initial cases and $\mathbf{W}_i$ is a lower-triangular matrix given by,

$$\mathbf{W}_i = \begin{bmatrix} 0 & 0 & 0 & \ldots & \ldots & 0 \\ w_1 & 0 & 0 & \ldots & \ldots & 0 \\ w_2 & w_1 & 0 & \ldots & \ldots & 0 \\ & & \vdots & & & \\ w_{T_i-1} & w_{T_i-2} & w_{T_i-3} & \ldots & w_1 & 0 \end{bmatrix}.$$

B.2 Derivation of Laplace approximation

To derive the Laplace approximation required in the E step, we first note that the regions are independent, and thus we can perform the Laplace approximations of the moments of $\boldsymbol{A}_i$ given $\boldsymbol{C}_i, \boldsymbol{K}_i$ independently across the regions $i = 1, \ldots, \mathcal{R}$. To simplify notation, we drop the subscripts i indexing region for the rest of this section.

Since the complete-data score vector $\frac{\partial}{\partial \boldsymbol{\theta}}\ell(\boldsymbol{\theta})$ is linear with respect to the unobserved variables $\boldsymbol{A}$, the first-order Laplace approximation of the expected complete-data score vector $\mathcal{S}(\boldsymbol{\theta}|\boldsymbol{\theta'}) = \frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{\mathcal{Q}}(\boldsymbol{\theta}|\boldsymbol{\theta'}) = \mathbb{E}_{\boldsymbol{\theta'}}\left\{\frac{\partial}{\partial \boldsymbol{\theta}}\ell(\boldsymbol{\theta})\middle|\,\mathcal{D}_{obs}\right\}$ requires the first two conditional moments of $\boldsymbol{A}|\boldsymbol{C},\boldsymbol{K}$. Explicitly, the expected complete-data score vector can be written as

$$\mathcal{S}(\boldsymbol{\theta}|\boldsymbol{\theta'}) = \mathbb{E}_{\boldsymbol{\theta'}}\left\{\mathbf{M}_{\boldsymbol{\theta}}\boldsymbol{A}|\,\mathcal{D}_{obs}\right\} = \mathbf{M}_{\boldsymbol{\theta}}\mathbb{E}_{\boldsymbol{\theta'}}\left(\boldsymbol{A}\,|\mathcal{D}_{obs}\,\right)$$

for a suitably defined matrix $\mathbf{M}_{\boldsymbol{\theta}}$. We approximate these moments using a Laplace approximation, which can be expressed as a quadratic expansion of the log probability function $P_{\boldsymbol{\theta'}}(\mathcal{D}_{mis}|\mathcal{D}_{obs})$ about its conditional mode $\boldsymbol{a}^*$,

$$\log P_{\boldsymbol{\theta'}}(\mathcal{D}_{mis} = \boldsymbol{A}\,|\mathcal{D}_{obs}\,) \approx \log P_{\boldsymbol{\theta'}}(\mathcal{D}_{mis} = \boldsymbol{a}^*\,|\mathcal{D}_{obs}\,) + \frac{1}{2}(\boldsymbol{A}-\boldsymbol{a}^*)^\top \mathbf{H}_A(\boldsymbol{a}^*)(\boldsymbol{A}-\boldsymbol{a}^*). \tag{S1.1}$$

where $\mathbf{H}_A = \frac{\partial^2}{\partial \boldsymbol{A}\partial \boldsymbol{A}^\top}\log P_{\boldsymbol{\theta'}}(\mathcal{D}_{mis} = \boldsymbol{A}\,|\mathcal{D}_{obs}\,)$ is the Hessian with respect to $\boldsymbol{A}$, and the gradient (not shown) vanishes when evaluated at the mode. Note that $\log P_{\boldsymbol{\theta'}}(\mathcal{D}_{mis} = \boldsymbol{A}\,|\mathcal{D}_{obs}\,) = \ell(\boldsymbol{\theta'}|\mathcal{D}_{mis} = \boldsymbol{A}, \mathcal{D}_{obs}) - \log P_{\boldsymbol{\theta'}}(\mathcal{D}_{obs})$, where ℓ denotes the complete-data log-likelihood, and the second term $\log P_{\boldsymbol{\theta'}}(\mathcal{D}_{obs})$ does not depend on $\boldsymbol{A}$. Therefore, we can write the gradient as $\boldsymbol{G}_A = \frac{\partial}{\partial \boldsymbol{A}}\log P_{\boldsymbol{\theta'}}(\mathcal{D}_{mis} = \boldsymbol{A}\,|\mathcal{D}_{obs}\,) = \frac{\partial \ell}{\partial \boldsymbol{A}}$ and the Hessian as $\mathbf{H}_A = \frac{\partial^2 \ell}{\partial \boldsymbol{A}\partial \boldsymbol{A}^\top}$.

To calculate the conditional mode $\boldsymbol{a}^*$ (and the Hessian $\mathbf{H}_A(\boldsymbol{a}^*)$ evaluated under the mode) using the Newton-Raphson, we first require the gradient $\boldsymbol{G}_A$ and Hessian $\mathbf{H}_A$. Below, we derive these quantities sequentially by first finding the derivatives with respect to $\boldsymbol{Y}$ and $\boldsymbol{M}$, and then applying the chain rule.

The gradient and Hessian with respect to $\boldsymbol{Y}$ are given by,

$$\boldsymbol{G}_Y = \frac{\partial \ell}{\partial \boldsymbol{Y}} = \frac{\partial \ell^{Sero}}{\partial \boldsymbol{Y}} + [\log(R_1\Lambda_1), ..., \log(R_T\Lambda_T)] + \mathbf{W}^\top \mathrm{diag}(\boldsymbol{\Lambda})^{-1}\boldsymbol{Y} - \mathbf{W}^\top \boldsymbol{R},$$
$$\mathbf{H}_Y = \frac{\partial^2 \ell}{\partial \boldsymbol{Y}\partial \boldsymbol{Y}^\top} = \frac{\partial^2 \ell^{Sero}}{\partial \boldsymbol{Y}\partial \boldsymbol{Y}^\top} + \mathrm{diag}(\boldsymbol{\Lambda})^{-1}\mathbf{W} + \mathbf{W}^\top \mathrm{diag}(\boldsymbol{\Lambda})^{-1} - \mathbf{W}^\top \mathrm{diag}(\boldsymbol{Y})\mathrm{diag}(\boldsymbol{\Lambda})^{-2}\mathbf{W},$$

where ℓ^{Sero} is the log likelihood for the seroprevalence component of the model. In turn, these are given by

$$\frac{\partial \ell^{Sero}}{\partial \boldsymbol{Y}} = \mathbf{C}_S^\top \boldsymbol{G}_S, \quad \frac{\partial^2 \ell^{Sero}}{\partial \boldsymbol{Y}\partial \boldsymbol{Y}^\top} = \mathbf{C}_S^\top \mathbf{H}_S \mathbf{C}_S,$$

where $\mathbf{C}_S$ is a $p \times T$ matrix with elements $[C_S]_{tk} = I(t \leq \tau_k)$ such that $\mathbb{E}(\boldsymbol{K}|\boldsymbol{Y}) = \frac{1}{n}\mathrm{diag}(\boldsymbol{N})\mathbf{C}_S^\top \boldsymbol{Y}$. The elements of the gradient $\boldsymbol{G}_S$ and Hessian $\mathbf{H}_S$ (a diagonal matrix) of ℓ^{Sero} with respect to $\boldsymbol{S} = \mathbf{C}_S^\top \boldsymbol{Y}$ are

$$[\boldsymbol{G}_S]_j = \psi(S_j+1) - \psi(S_j - K_j + 1) - \psi(n - S_j + 1) + \psi(n - S_j - N_j + K_j + 1),$$
$$[\mathbf{H}_S]_{jj} = \psi_1(S_j+1) - \psi_1(S_j - K_j + 1) + \psi_1(n - S_j + 1) - \psi_1(n - S_j - N_j + K_j + 1)$$

where $\psi(x) = \frac{\partial}{\partial x}\log\Gamma(x)$ and $\psi_1(x) = \frac{\partial^2}{\partial x^2}\log\Gamma(x)$ are digamma and trigamma functions, respectively. Note that these quantities arise only in the E step of the EM algorithm, and in the correction factor for the information matrix.

Under random sampling with replacement (binomial rather than hypergeometric), these quantities instead have the forms,

$$[\boldsymbol{G}_S]_j = \frac{K_j}{S_j} - \frac{N_j - K_j}{n - S_j},$$
$$[\mathbf{H}_S]_{jj} = -\frac{K_j}{(S_j)^2} - \frac{(N_j - K_j)}{(n - S_j)^2}.$$

The gradient and Hessian with respect to $\boldsymbol{M}$ are given by,

$$\boldsymbol{G}_M = \frac{\partial \ell}{\partial \boldsymbol{M}} = \boldsymbol{\zeta}_0 + \mathbf{h}(\boldsymbol{M}) - \mathbf{h}(\boldsymbol{M} - \boldsymbol{C}),$$
$$\mathbf{H}_M = \frac{\partial^2 \ell}{\partial \boldsymbol{M} \partial \boldsymbol{M}^\top} = \mathbf{h}'(\boldsymbol{M}) - \mathbf{h}'(\boldsymbol{M} - \boldsymbol{C}),$$

where $\mathbf{h}(\boldsymbol{x}) = (\psi(x_1 + 1), ..., \psi(x_n + 1))^\top$ and $\mathbf{h}'(\boldsymbol{x}) = \mathrm{diag}(\psi_1(x_1 + 1), ..., \psi_n(x_n + 1))$.

Finally, the required gradient and Hessian with respect to $\boldsymbol{A}$ can be expressed as

$$\boldsymbol{G}_A = \frac{\partial \ell}{\partial \boldsymbol{A}} = \mathbf{P}^\top \boldsymbol{G}_Y + \mathbf{L}^\top \boldsymbol{G}_M + \boldsymbol{\eta} - \mathbf{h}(\boldsymbol{A}),$$
$$\mathbf{H}_A = \frac{\partial^2 \ell}{\partial \boldsymbol{A} \partial \boldsymbol{A}^\top} = \mathbf{P}^\top \mathbf{H}_Y \mathbf{P} + \mathbf{L}^\top \mathbf{H}_M \mathbf{L} - \mathbf{h}'(\boldsymbol{A}).$$

To calculate the conditional mode $\boldsymbol{a}_*$ using the Newton-Raphson, and to calculate the information matrix, we must invert the Hessian $\mathbf{H}_A$. While $\mathbf{H}_A$ is a $T_i(m_A + 1) \times T_i(m_A + 1)$ matrix, its inverse can be calculated efficiently by applying the Woodbury identity twice:

$$\mathbf{H}_A^{-1} = \mathbf{Q}^{-1} - \mathbf{Q}^{-1}\mathbf{J}^\top \left(\mathbf{H}_Y^{-1} + \mathbf{P}\mathbf{Q}^{-1}\mathbf{P}^\top\right)^{-1} \mathbf{P}\mathbf{Q}^{-1}$$

where

$$\begin{aligned}
\mathbf{Q}^{-1} &= \left(\mathbf{L}^\top \mathbf{H}_M \mathbf{L} - \mathbf{h}'(\boldsymbol{A})\right)^{-1} \\
&= -\mathbf{h}'(\boldsymbol{A})^{-1} - \mathbf{h}'(\boldsymbol{A})^{-1}\mathbf{L}^\top \left(\mathbf{H}_M^{-1} - \mathbf{L}\mathbf{h}'(\boldsymbol{A})^{-1}\mathbf{L}^\top\right)^{-1} \mathbf{L}\mathbf{h}'(\boldsymbol{A})^{-1}.
\end{aligned}$$

Also, it is helpful to observe that $\mathbf{P}\mathrm{diag}(\boldsymbol{x})\mathbf{P}^\top = \mathrm{diag}(\mathbf{P}\boldsymbol{x})$ and $\mathbf{L}\mathrm{diag}(\boldsymbol{x})\mathbf{L}^\top = \mathrm{diag}(\mathbf{L}\boldsymbol{x})$.

C Derivation of the observed data Fisher information matrix

To obtain standard errors for parameter estimates, we use the corrected information matrix,

$$\boldsymbol{\mathcal{J}}_{\mathcal{D}_{obs}}(\boldsymbol{\theta}) = \boldsymbol{\mathcal{J}}_{\mathcal{D}}(\boldsymbol{\theta}|\boldsymbol{\theta}') - \boldsymbol{\mathcal{J}}_{\mathcal{D}_{mis}|\mathcal{D}_{obs}}(\boldsymbol{\theta}|\boldsymbol{\theta}'),$$

where,

$$\boldsymbol{\mathcal{J}}_{\mathcal{D}}(\boldsymbol{\theta}|\boldsymbol{\theta}') = \mathbb{E}_{\boldsymbol{\theta}'}\left\{ -\frac{\partial^2 \ell(\boldsymbol{\theta}|\mathcal{D})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\middle| \mathcal{D}_{obs}\right\},$$
$$\boldsymbol{\mathcal{J}}_{\mathcal{D}_{mis}|\mathcal{D}_{obs}}(\boldsymbol{\theta}|\boldsymbol{\theta}') = \mathrm{Var}_{\boldsymbol{\theta}'}\left\{ \frac{\partial \ell(\boldsymbol{\theta}|\mathcal{D})}{\partial\boldsymbol{\theta}}\middle| \mathcal{D}_{obs}\right\}.$$

The complete data information matrix $\boldsymbol{\mathcal{J}}_{\mathcal{D}}$ is block diagonal, with separate blocks for transmission parameters $\boldsymbol{\beta}^{(R)}$ and ascertainment parameters $\boldsymbol{\beta}^{(\pi)}$. The correction term $\boldsymbol{\mathcal{J}}_{\mathcal{D}_{mis}|\mathcal{D}_{obs}}$ can be written as $\mathrm{Var}_{\boldsymbol{\theta}'}\left\{ \frac{\partial}{\partial\boldsymbol{\theta}}\ell(\boldsymbol{\theta}|\mathcal{D})\middle| \mathcal{D}_{obs}\right\} = \sum_{i=1}^{\mathcal{R}} \mathbf{Q}_i \mathrm{Var}_{\boldsymbol{\theta}'}\left(\boldsymbol{A}_i|\,\boldsymbol{C}_i, \boldsymbol{K}_i\right)\mathbf{Q}_i^\top$, where $\mathrm{Var}_{\boldsymbol{\theta}'}\left(\boldsymbol{A}_i|\,\boldsymbol{C}_i, \boldsymbol{K}_i\right) \approx -\mathbf{H}(\boldsymbol{a}_i^*)^{-1}$ is obtained using the Laplace approximation at the conditional mode $\boldsymbol{A}_i = \boldsymbol{a}_i^*$. We derive the explicit algebraic forms of $\mathcal{J}_{\mathcal{D}_{obs}}$ and $\mathbf{Q}_i$ next.

First, the complete data score vector $\tilde{\boldsymbol{\mathcal{S}}} = \frac{\partial}{\partial\boldsymbol{\theta}}\ell\left(\boldsymbol{\theta}|\mathcal{D}\right)$ can be expressed as $\tilde{\boldsymbol{\mathcal{S}}} = (\tilde{\boldsymbol{\mathcal{S}}}_R, \tilde{\boldsymbol{\mathcal{S}}}_\pi)$, where,

$$\tilde{\boldsymbol{\mathcal{S}}}_R = \sum_{i=1}^{\mathcal{R}} \mathbf{X}_i^\top\left(\boldsymbol{Y}_i - \mathrm{diag}(\boldsymbol{R}_i)\boldsymbol{\Lambda}_i\right) = \sum_{i=1}^{\mathcal{R}} \mathbf{X}_i^\top\left[\{\mathbf{I}_{T_i} - \mathrm{diag}(\boldsymbol{R}_i)\mathbf{W}_i\}\,\mathbf{P}_i\boldsymbol{A}_i - \mathrm{diag}(\boldsymbol{R}_i)\boldsymbol{\lambda}_{i0}\right],$$
$$\tilde{\boldsymbol{\mathcal{S}}}_\pi = \sum_{i=1}^{\mathcal{R}} \mathbf{Z}_i^\top\left(\boldsymbol{C}_i - \mathrm{diag}(\boldsymbol{\pi}_i)\boldsymbol{M}_i\right) = \sum_{i=1}^{\mathcal{R}} \mathbf{Z}_i^\top\left[\boldsymbol{C}_i - \mathrm{diag}(\boldsymbol{\pi}_i)\mathbf{L}_i\mathbf{A}_i\right].$$

Here, $\mathrm{diag}(\mathbf{x})$ is the diagonal matrix with diagonal elements given by the vector $\mathbf{x}$. The second set of equalities are observed by substituting $\boldsymbol{\Lambda}_i = \mathbf{W}_i\boldsymbol{Y}_i + \boldsymbol{\lambda}_{i0}$, $\boldsymbol{Y}_i = \mathbf{P}_i\boldsymbol{A}_i$, and $\boldsymbol{M}_i = \mathbf{L}_i\boldsymbol{A}_i$. Therefore,

$$\mathrm{Var}_{\boldsymbol{\theta}'}\left(\tilde{\boldsymbol{\mathcal{S}}}|\boldsymbol{C}, \boldsymbol{K}\right) = \mathrm{Var}_{\boldsymbol{\theta}'}\left(\begin{bmatrix}\tilde{\boldsymbol{\mathcal{S}}}_R \\ \tilde{\boldsymbol{\mathcal{S}}}_\pi\end{bmatrix}\middle| \boldsymbol{C}, \boldsymbol{K}\right) = \sum_{i=1}^{\mathcal{R}} \mathbf{Q}_i \mathrm{Var}_{\boldsymbol{\theta}'}\left(\boldsymbol{A}_i|\,\boldsymbol{C}_i, \boldsymbol{K}_i\right)\mathbf{Q}_i^\top,$$

where,

$$\mathbf{Q}_i = \begin{bmatrix}\mathbf{X}_i^\top\{\mathbf{I}_{T_i} - \mathrm{diag}(\boldsymbol{R}_i)\mathbf{W}_i\} & \mathbf{0} \\ \mathbf{0} & -\mathbf{Z}_i^\top\mathrm{diag}(\boldsymbol{\pi}_i)\end{bmatrix}\begin{bmatrix}\mathbf{P}_i \\ \mathbf{L}_i\end{bmatrix}$$

Therefore, corrected information matrix is,

$$\boldsymbol{\mathcal{J}}_{\mathcal{D}_{obs}} = \mathbb{E}\left(\begin{bmatrix}\boldsymbol{\mathcal{J}}_R & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\mathcal{J}}_\pi\end{bmatrix}\middle| \boldsymbol{C}, \boldsymbol{K}\right) - \mathrm{Var}\left(\tilde{\boldsymbol{\mathcal{S}}}\middle|\,\boldsymbol{C}, \boldsymbol{K}\right),$$

where,

$$\boldsymbol{\mathcal{J}}_R = -\frac{\partial^2\ell(\boldsymbol{\theta}|\mathcal{D})}{\partial\boldsymbol{\beta}^{(R)}\partial\boldsymbol{\beta}^{(R)\top}} = \sum_{i=1}^{\mathcal{R}} \mathbf{X}_i^\top\mathrm{diag}(R_{i1}\Lambda_{i1}, ..., R_{iT_i}\Lambda_{iT_i})\mathbf{X}_i$$
$$\boldsymbol{\mathcal{J}}_\pi = -\frac{\partial^2\ell(\boldsymbol{\theta}|\mathcal{D})}{\partial\boldsymbol{\beta}^{(\pi)}\partial\boldsymbol{\beta}^{(\pi)\top}} = \sum_{i=1}^{\mathcal{R}} \mathbf{Z}_i^\top\mathrm{diag}(M_{i1}\pi_{i1}(1-\pi_{i1}), ..., M_{iT_i}\pi_{iT_i}(1-\pi_{iT_i}))\mathbf{Z}_i.$$

D Two-stage EM procedure for estimating containment policy effects

As discussed in Section 5.3, to estimate the effects of containment policies on the effective reproductive numbers R_{it} across US states $i = 1, 2, ...\mathcal{R}$ and time-points $t = 1, 2, ..., J$, we fitted the R_t regression model (8) in the paper, which is reproduced below

$$\log R_{it} = \beta_{0i} + \beta_B(t) + \sum_k \boldsymbol{X}_{ik}(t)^\top \boldsymbol{\beta}_k, \tag{S1.2}$$

where the baseline function $\beta_B(t)$ is specified using a regression spline with B-spline basis as $\beta_B(t) = \boldsymbol{B}(t)^\top \boldsymbol{\beta}_B$, and $\boldsymbol{X}_{ik}(t)$ is a time-varying containment policy index for policy category k in state i.

We used a 2-stage estimation procedure to fit model (S1.2), where the first stage uses the EM algorithm proposed in Section 3 to estimate state-specific $\log R_{it}$ curves, and the second stage maximizes the expected complete-data likelihood under model (S1.2). This approach is more computationally efficient than directly fitting (S1.2) using the EM algorithm, as we can parallelize across regions in the first stage, and avoid computationally intensive E-step updates across states in the second stage. Below, we describe this procedure in greater detail, and examine properties of the resulting 2-stage estimator.

Recall that MERMAID estimates the effective reproductive numbers $\log R_{it}$ and ascertainment probabilities π_{it} over time in each region i. Here, we will partition the full set of MERMAID parameters into two components as $\boldsymbol{\theta} = (\boldsymbol{\beta}^{(\pi)}, \boldsymbol{\beta}^{(R)})$. In the M step of the EM algorithm, we maximize the expected complete-data log-likelihood, given by

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}') = \int P_{\boldsymbol{\theta}'}(\mathcal{D}_{mis}|\mathcal{D}_{obs}) \log P(\mathcal{D}_{mis}, \mathcal{D}_{obs}|\boldsymbol{\theta}) d\mathcal{D}_{mis},$$

where $\boldsymbol{\theta}'$ is evaluated at the parameter values from the previous iteration. Under the MERMAID model, we can expand $\log P(\mathcal{D}_{mis}, \mathcal{D}_{obs}|\boldsymbol{\theta})$ to write

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}') = Q_{(R)}(\boldsymbol{\beta}^{(R)}|\boldsymbol{\theta}') + Q_{(\pi)}(\boldsymbol{\beta}^{(\pi)}|\boldsymbol{\theta}'),$$

where $Q_{(R)}(\boldsymbol{\beta}^{(R)}|\boldsymbol{\theta}') = \mathbb{E}_{\boldsymbol{\theta}'}\left\{g(\mathcal{D}_{mis}, \mathcal{D}_{obs}|\boldsymbol{\beta}^{(R)})|\mathcal{D}_{obs}\right\}$ and $Q_{(\pi)}(\boldsymbol{\beta}^{(\pi)}|\boldsymbol{\theta}') = \mathbb{E}_{\boldsymbol{\theta}'}\left\{h(\mathcal{D}_{mis}, \mathcal{D}_{obs}|\boldsymbol{\beta}^{(\pi)})|\mathcal{D}_{obs}\right\}$ for suitably defined functions g and h. Thus, we maximize $Q_{(R)}(\boldsymbol{\beta}^{(R)}|\boldsymbol{\theta}')$ to update the parameters underlying the $\log R_{it}$ model in a conventional EM algorithm.

Consider the state-specific R_{it} model

$$\log R_{it} = \beta_i(t) \tag{S1.3}$$

where $\beta_i(t)$ is a nonparametric function for state i. As a saturated model, $\beta_i(t)$ can capture the effects of containment policies over time – in particular, model (S1.2) is nested in model (S1.3), as we can write $\beta_i(t) = \beta_{0i} + \beta_B(t) + \sum_k \boldsymbol{X}_{ik}(t)^\top \boldsymbol{\beta}_k$ as a special case. In practice, we estimate (S1.2) using regression splines, and therefore the estimator $\hat{\beta}_i(t)$ can only approximately capture the policy-effect model (S1.2). However, the error of spline approximation can be made arbitrarily small by increasing the number of spline knots (Huang et al., 2003).

This suggests the following two-stage estimation procedure for model (S1.2).

1. Use the EM algorithm from Section 3 to fit the saturated model (S1.3) in each state using regression splines for the $\log R_{it}$ mean model. Let $\hat{\boldsymbol{\theta}}^{(1)}$ denote the resulting parameter estimates across states, including the parameters for $\log R_{it}$ and ascertainment probabilities π_{it}.

2. Estimate the $\log R_{it}$ parameters $\beta_{0i}, \boldsymbol{\beta}_B, \boldsymbol{\beta}_k$ under model (S1.2) by maximizing $Q_{(R)}(\boldsymbol{\beta}^{(R)}|\hat{\boldsymbol{\theta}}^{(1)})$.

When model (S1.2) is correctly specified and nested in the saturated model (S1.3), and the ascertainment model is correctly specified, the first-stage estimates $\hat{\boldsymbol{\theta}}^{(1)}$ are consistent for the data-generating values $\boldsymbol{\theta}^*$. Therefore, since the maximizer of $Q_{(R)}(\boldsymbol{\beta}^{(R)}|\boldsymbol{\theta}^*)$ is a consistent estimator of the data-generating value $\boldsymbol{\beta}^{(R)*}$, the second-stage estimator is consistent under the above assumptions. We can calculate approximate SEs for the second-stage estimates using Louis' formula, as described in the previous section.

E Simulation Studies in More Realistic Settings

We further investigated the performance of our method under a misspecified model using a data generating model that more mimics the reality, where the reproductive rates follow smooth functional curves with respect to time, and the symptomatic, asymptomatic, and uninfected individuals have different chances of getting tested at different time periods throughout the epidemic. Additional simulation studies, including misspecification of the nuisance parameters (of the serial interval distribution and infection-reporting lag distribution) are shown in Supplementary Materials S2.

E.1 Data generation model:

Similar to the simulation setting in the main paper, we simulated epidemics lasting 1 year ($T = 365$ days) in a single region with a population size of $n_i = 8,000,000$. The data generation model is described as follows,

1. Set the reproductive numbers using a smooth function $R_{it} = 2e^{-10t} + 0.3cos^2(5\pi t/2) + t/3 + 0.65$. We chose this particular function to have three waves similar to the pattern experienced in the US. The exact form of the R_t curve is shown in Supplementary Figure 1.

2. Draw the number of newly infected cases $Y_{it} \sim \text{Poisson}(R_{it}\Lambda_{it}) + 5I_{\{t\leq 7\}}(t)$, where $\Lambda_{it} = \sum_{s=0}^{t-1} w_{t-s}Y_{is}$ is the infection potential, $Y_{i0} = 0$, and $I_{\{t\leq 7\}}(t)$ is the indicator function. Five extra infections were added to the first 7 days to represent infected people coming from outside into each region during the early stages of the epidemic. The serial interval distribution was assumed to have mean 4.7 and standard deviation 2.9 truncated to 30 days.

3. Draw $\boldsymbol{A}_{it} \sim \text{Multinomial}(Y_{it}, \phi)$, where $A_{i,t,k}$ is the number of cases potentially confirmed on day $t + k$ $(k = 1, \ldots, m_A)$ among the number of cases infected on day t, Y_{it}. The reporting lags were assumed to follow NegativeBinomial$(r = 5, \mu = 5)$ truncated to 21 days.

4. Calculate the number of cases that are potentially confirmed on day t as $M_{it} = \sum_{s=0}^{m_A} A_{i,t-s,s}$.

5. Assume that each case can be symptomatic with a probability p_S, and the ascertainment probabilities vary depending on whether an individual is symptomatic, asymptomatic, or uninfected, and those probabilities can also vary across time depending on the availability of tests. Detailed description of how to simulate ascertained and unascertained cases is given below.

E.2 Simulating ascertained and unascertained cases:

To simulate the ascertainment of cases in a more realistic scenario, we first generate the following,

$$M_{it}^{(S)} \sim \text{Binomial}(M_{it}, p_S); \quad M_{it}^{(A)} = M_{it} - M_{it}^{(S)},$$

where $p_S = 0.2$ is the probability that an infection is symptomatic, and $M_{it}^{(S)}$ and $M_{it}^{(A)}$ are the numbers of symptomatic and asymptomatic infections, respectively, among the M_{it} many potentially observed individuals on day t.

Next, we assume different probabilities of getting tested depending on the individual's infection and symptom status. Specifically, we assume the numbers of symptomatic, asymptomatic, and

uninfected individuals that are tested on day t in region i to be,

$$\begin{aligned}
\text{Symptomatic Tested: } T_{it}^{(S)} &\sim \text{Binomial}\left(M_{it}^{(S)}, p_{Si}(t)\right), \\
\text{Asymptomatic Tested: } T_{it}^{(A)} &\sim \text{Binomial}\left(M_{it}^{(A)}, p_{Ai}(t)\right), \text{ and} \\
\text{Uninfected Tested: } T_{it}^{(U)} &\sim \text{Binomial}\left(N_i - \sum_{s=0}^{t} Y_{is}, p_{Ui}(t)\right),
\end{aligned}$$

where $p_{Si}(t), p_{Ai}(t)$, and $p_{Ui}(t)$ are the probabilities of getting tested at time t given the individual is symptomatic, asymptomatic, or uninfected, respectively. We fix $p_{Si}(t) = 0.95$ throughout the epidemic, i.e., assume that the symptomatic individuals will get tested with very high probability (95%) regardless of time. With varying testing capabilities during the epidemic, we let the rest of the probabilities $p_{Ai}(t)$ and $p_{Ui}(t)$ to vary. Furthermore, to mimic the weekly cyclical pattern of the testing rates and the different waves of the epidemic, the functions $p_{Ai}(t)$ and $p_{Ui}(t)$ were designed to have a weekly periodic component, and a component corresponding to the different waves of the epidemic. The exact functional forms of $p_{Ai}(t)$ and $p_{Ui}(t)$ are shown in Supplementary Figure 1.

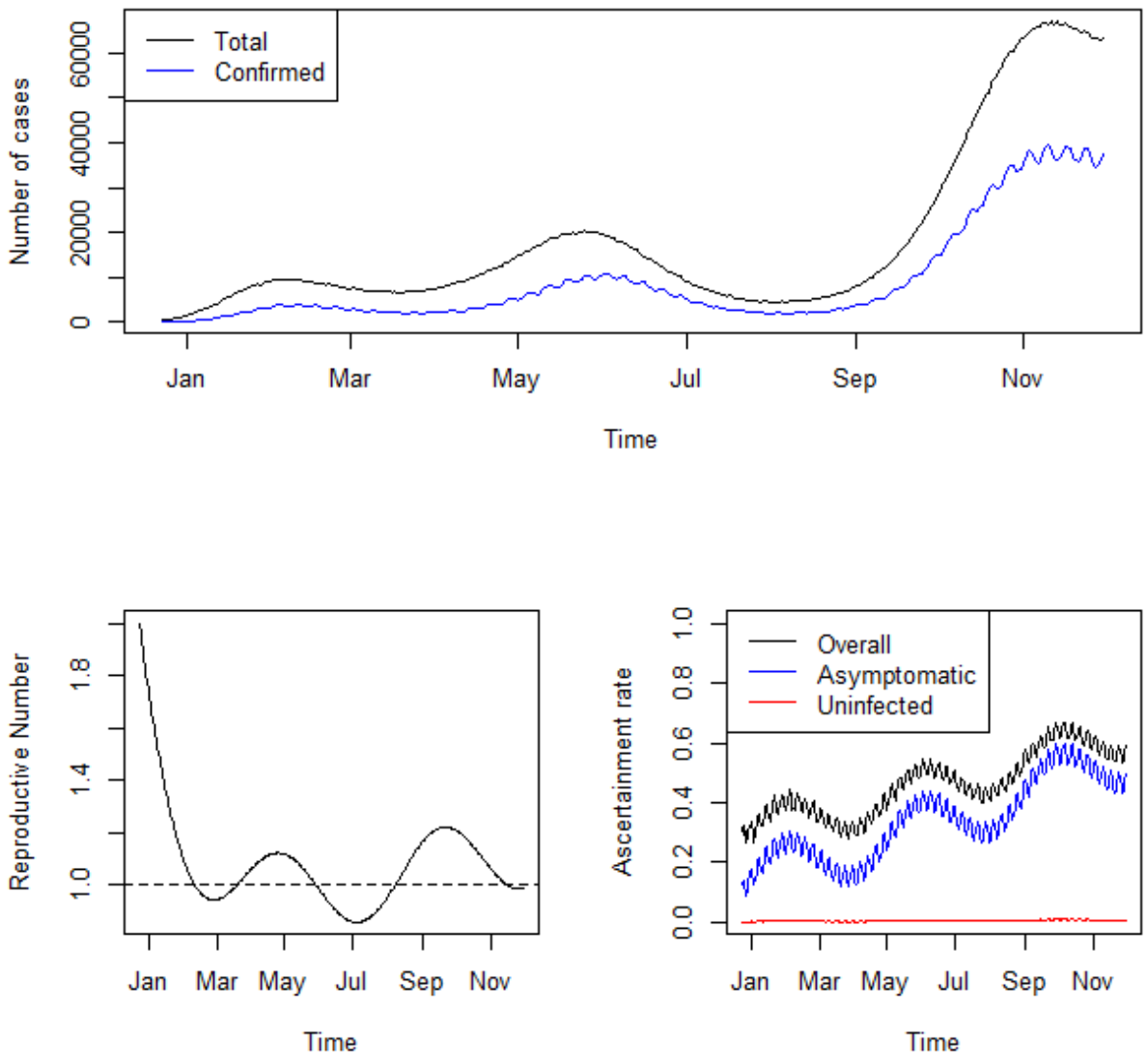
Once we have obtained the numbers of individuals tested under different symptom and infection status, the number of confirmed and unascertained cases, and the number of tests conducted on day t in region i can be calculated by,

$$\begin{aligned}
\text{Confirmed cases: } C_{it} &= T_{it}^{(S)} + T_{it}^{(A)}, \\
\text{Unascertained cases: } U_{it} &= M_{it} - C_{it}, \text{ and} \\
\text{Number of tests: } N_{it}^{(T)} &= C_{it} + T_{it}^{(U)}.
\end{aligned}$$

Here, we are assuming that the tests have full sensitivity, i.e., if an infected individual is tested, that individual will be tested positive with complete certainty. The overall ascertainment rate on day t for region i is then defined as $\pi_{it} = C_{it}/M_{it}$. We further generate $J_i = 6$ seroprevalence estimates for each region at times equally spread throughout the epidemic period, with each survey having sample sizes $N_i = 80,000$.

E.3 Simulation results

We assessed the performance of MERMAID (which will be a mis-specified model) by applying it on 500 epidemics simulated using the above-mentioned procedure. We correctly specified the lag distribution, and in each region i, we started the study periods on the first day $t = t_0$ such that $C_{it_0} \geq 50$. The serial interval distribution $\{w_s\}_{s=0}^{\infty}$, and the initial infection potential for the study period Λ_{it_0} are also correctly specified. To model the R_t and π_t curves, we used log-linear model



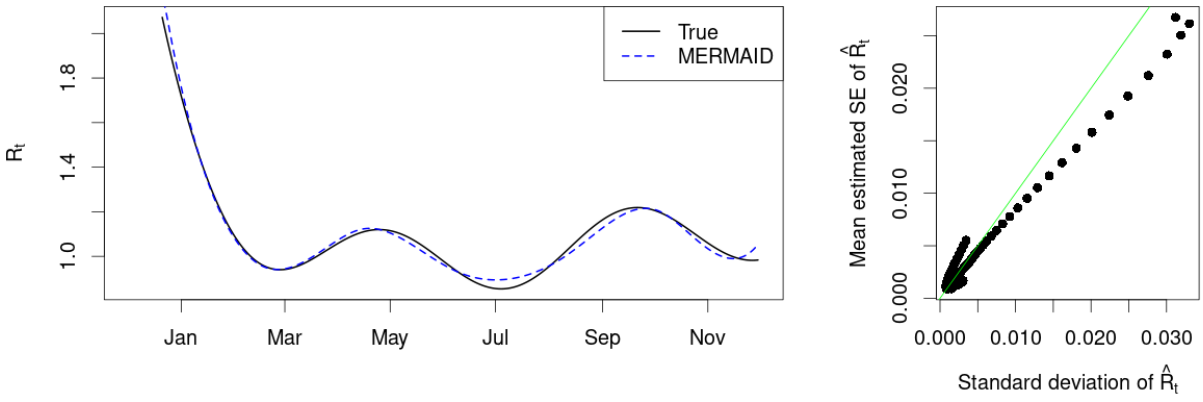
Supplementary S1 Figure 1: A more realistic simulation setting under model misspecification, where the ascertainment probabilities depend on whether the individuals are symptomatic, asymptomatic, or uninfected. The top panel shows the true total number of daily infections (black), and the daily number of confirmed cases (blue). The bottom left panel shows the form of the R_t curve. The bottom right panel shows the overall daily ascertainment rates (black), and the daily ascertainment probabilities for asymptomatic (blue) and uninfected (red) individuals. The ascertainment probabilities for symptomatic individuals were set at 95% for the entire duration of the epidemic.

with B-spline approximation and logit-linear model, respectively, as follows,

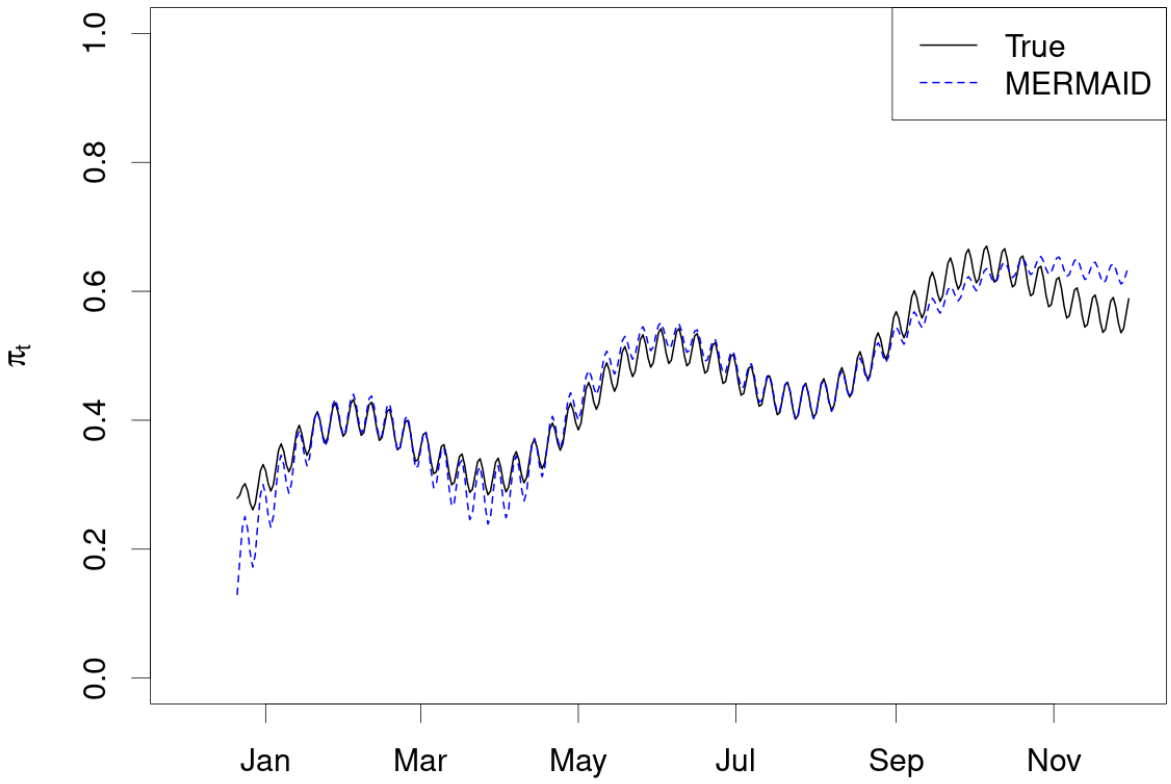
$$\log(R_{it}) = \alpha_{it}^{(R)} + \sum_{j=1}^{J} B_j(t)\beta_{ji}^{(R)},$$
$$\log\left(\frac{\pi_{it}}{1-\pi_{it}}\right) = \beta_{0i}^{(\pi)} + \log\left(N_{it}^{(T)}\right)\beta_{1i}^{(\pi)},$$

The offsets in the log-linear model for R_{it} are given by $\alpha_{it}^{(R)} = \log(1 - F_{it})$, where F_{it} is the fraction of the population that is immune to infection at time t. The seroprevalence log-likelihood was weighted by the factor $c_S = 25$, similar to the simulation setting presented in the main paper.

The results suggest that under this mis-specified model, the R_t estimates based on MERMAID remained robust. The standard errors for $\hat{R}_t$ were slightly under-estimated (Supplementary Figure 2). For the ascertainment probabilities (Supplementary Table 1 and Supplementary Figure 3), the estimates $\hat{\pi}_t$ had almost no bias throughout the duration of the epidemic except for a small amount of bias at the very beginning (January) and the end (November). The standard error estimates for $\hat{\pi}_t$ were fairly accurate with a little over-estimation during the first half of the year, and a little under-estimation in the later half. (Supplementary Table 1).



Supplementary S1 Figure 2: Estimates of R_{it} under misspecified models in a more realistic simulation scenario where R_t is a smooth function with respect to time, and the ascertainment probabilities depend on the symptomatic/asymptomatic/uninfected status of the individuals. The left panel shows the true R_t (black) and the estimated $\hat{R}_t$ based on MERMAID (blue dotted). The right panel compares the empirical standard deviations of $\hat{R}_t$s (x-axis) and the mean of the estimated standard errors of $\hat{R}_t$s (y-axis).



Supplementary S1 Figure 3: Estimates of ascertainment probabilities (π_t) under misspecified models in a more realistic simulation scenario where R_t is a smooth function with respect to time, and the ascertainment probabilities depend on the symptomatic/asymptomatic/uninfected status of the individuals. True π_t is shown by the solid black line and the estimated π_t based on MERMAID is shown by the dotted blue line.

Date	π_t	$\bar{\hat{\pi}}_t^{(i)}$	$\bar{\widehat{SE}}(\hat{\pi}_t^{(i)})$	$SD(\hat{\pi}_t^{(i)})$
Jan 1	0.322	0.284	4.69E-03	4.59E-03
Mar 1	0.358	0.355	3.19E-03	2.61E-03
May 1	0.385	0.400	2.33E-03	1.99E-03
Jul 1	0.484	0.492	2.61E-03	2.69E-03
Sep 1	0.568	0.545	3.85E-03	4.51E-03
Nov 1	0.598	0.640	6.19E-03	8.65E-03

Supplementary S1 Table 1: Estimates of ascertainment probabilities π_{it} with the true π_{it} is based on symptomatic/asymptomatic/uninfected status of the individuals. Values are shown on the first day of every second month. Shown are the mean estimated value, mean adjusted standard error, and standard deviation of estimates across simulation replicates.

F Preprocessing procedures for COVID-19 cases and PCR tests

F.1 COVID-19 reported case counts

We obtained daily reported COVID-19 cases across US states from three sources: (1) the CDC data repository (CDC 2021), (2) the COVIDTracking project (COVIDTracking 2021), and (3) the USAFacts.org webpage (USAFacts 2021). Substantial differences in reported case counts between sources were present in a number of regions (comparisons shown in Supplementary Materials S3-4). To obtain a consensus confirmed case count time series , we aggregated across the 3 data sources within each region using the following procedure. First, we applied an iterated moving average filter (KZ filter; Close et al. 2020) with length 3 days and 3 iterations to smooth daily counts in each region and from each data source. Second, we selected the two most concordant data sources in each region (smallest sum of absolute differences), and discarded the third more discordant source. Third, we merged the two selected case counts in each region over time by taking the maximum of the cumulative reported counts between the two sources on each day.

F.2 Merging confirmed cases with COVID-19 positive test counts

We obtained daily COVID-19 testing data (numbers of positive and negative PCR and other test specimens or individuals tested) from 2 sources: (1) the COVID Electronic Laboratory Reporting Program (CELR), conducted by the US federal government and available from the HealthData.gov data repository (HealthData.gov 2021) and (2) the COVIDTracking project, which primarily uses data from state-level sources (COVIDTracking 2021; Schechtman 2021). We similarly noted discordance between these two data sources (Supplementary Materials S3-4), with the first source generally appearing to provide more timely reports, as noted previously on the COVIDTracking webpage (Schechtman 2021). We therefore used the CELR test counts, applying the iterated moving average filter with length 3 days and 3 iterations to smooth irregularities as in the previous section.

We noted apparent time lags between confirmed case times series (described in Section F.1 and Schechtman 2021) and tests performed according to CELR. Because MERMAID relies on PCR testing data to model the ascertainment rate over time, we sought to minimize the discordance between the confirmed case count variable and PCR testing data prior to model fitting. To do so, we constructed a new confirmed case variable by combining data on positive PCR tests from CELR with the consensus confirmed case count data described in the previous section. CELR provides the total numbers of positive, negative, and inconclusive PCR tests on each day within

each region; positive tests may exceed confirmed cases when multiple tests are performed per case. We calculated a new confirmed-case time series by re-scaling the numbers of positive PCR tests reported by CELR to approximately match the numbers of confirmed cases obtained in Section F.1 within sliding windows. Specifically, we used least squares to estimate a time-varying scaling factor s_{it} parameterized by $\boldsymbol{b}_i$ in each US state i by minimizing $\sum_t (C_{it}^{(A)} - s_{it}(\boldsymbol{b}_i) C_{it}^{(B)})^2$ across days t. Here, $C_{it}^{(A)}$ is the number of confirmed cases (described in Section F.1), $C_{it}^{(B)}$ is the number of positive PCR tests reported by CELR, and $s(t; \boldsymbol{b}_i)$ is a B-spline function parameterized by $\boldsymbol{b}_i$. This procedure is equivalent to a varying-coefficient regression without an intercept. Here, the coefficient s_{it} is a scaling factor which captures the number of positive cases per positive PCR test at time t in state i. We used the re-scaled positive test counts $C_{it} \coloneqq s(t; \hat{\boldsymbol{b}}_i) C_{it}^{(B)}$ in subsequent analysis with MERMAID. The re-scaled confirmed case counts are consistent with PCR testing data from CELR with respect to the dates of reporting (i.e., not time-lagged), and consistent with the previous data sources (i.e., $C_{it}^{(A)}$) with respect to the total numbers of unique confirmed-positive individuals in larger time windows. Comparisons of all test and confirmed case time series are given in Supplementary Materials S3-4. In a majority of states, CELR data appears to show fewer irregularities (gaps in reporting) and prompter reporting than other sources.

References

CDC (2021). United States COVID-19 cases and deaths by state over time | data | centers for disease control and prevention. `https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36`. (Accessed on 04/05/2021).

Close, B., Žurbenko, I., and Sun, M. (2020). *kza: Kolmogorov-Žurbenko Adaptive Filters*. R package version 4.1.0.1.

COVIDTracking (2021). The data: The COVID Tracking Project. `https://covidtracking.com/data`. (Accessed on 04/05/2021).

HealthData.gov (2021). COVID-19 diagnostic laboratory testing (PCR testing) time series; healthdata.gov. `https://healthdata.gov/dataset/COVID-19-Diagnostic-Laboratory-Testing-PCR-Testing/j8mb-icvb`. (Accessed on 04/05/2021).

Huang, J. Z. et al. (2003). Local asymptotics for polynomial spline regression. *Annals of Statistics*, 31(5):1600–1635.

Schechtman, K. W. (2021). Analysis & updates Federal testing data's last mile. `https://covidtracking.com/analysis-updates/federal-testing-datas-last-mile`. (Accessed on 04/05/2021).

USAFacts (2021). Us COVID-19 cases and deaths by state. `https://usafacts.org/visualizations/coronavirus-covid-19-spread-map`. (Accessed on 04/05/2021).