Published in final edited form as:

Biometrics. 2012 December ; 68(4): 1238–1249. doi:10.1111/j.1541-0420.2012.01757.x.

A Hybrid EM and Monte Carlo EM Algorithm and Its Application to Analysis of Transmission of Infectious Diseases

Yang Yang¹, Ira M. Longini Jr.^{1,2}, M. Elizabeth Halloran^{1,2}, and Valerie Obenchain¹

¹Center for Statistics and Quantitative Infectious Diseases Fred Hutchinson Cancer Research Center, Seattle, WA 98109

²Department of Biostatistics, University of Washington, Seattle, WA 98195

Summary

In epidemics of infectious diseases such as influenza, an individual may have one of four possible final states: prior immune, escaped from infection, infected with symptoms, and infected asymptotically. The exact state is often not observed. In addition, the unobserved transmission times of asymptomatic infections further complicate analysis. Under the assumption of missing at random, data-augmentation techniques can be used to integrate out such uncertainties. We adapt an importance-sampling-based Monte Carlo EM (MCEM) algorithm to the setting of an infectious disease transmitted in close contact groups. Assuming the independence between close contact groups, we propose a hybrid EM-MCEM algorithm that applies the MCEM or the traditional EM algorithms to each close contact group depending on the dimension of missing data in that group, and discuss the variance estimation for this practice. In addition, we propose a bootstrap approach to assess the total Monte Carlo error and factor that error into the variance estimation. The proposed methods are evaluated using simulation studies. We use the hybrid EM-MCEM algorithm to analyze two influenza epidemics in the late 1970s to assess the effects of age and pre-season antibody levels on the transmissibility and pathogenicity of the viruses.

Keywords

Data augmentation; EM algorithm; Infectious disease; Missing data; Monte Carlo

1. Introduction

To assess transmissibility and pathogenicity of acute infectious diseases such as seasonal influenza in close contact groups, it is important to collect information about baseline susceptibility, infection outcomes and symptom outcomes at the individual level whenever possible. Infection outcomes here refer to infection status (infected or escaped) and infection time. Data on symptom outcomes include symptom status (present or absent) and symptom onset time. In more complex settings, the outcomes may incorporate additional information such as the viral type or genetic sequences of the pathogen associated with a given infection. Infection and symptom outcomes of each individual in turn provide information about the exposure history of his or her contacts. In most epidemiological studies, however, not all individual-level information is collected. Among all the crucial measurements, the most difficult one to obtain is the infection time of asymptomatic cases. Missing information also occurs in outbreaks of newly emerging pathogens, in which baseline immunological

Supplementary Materials

Web Appendices referenced in Sections 1 and 4, together with the data and computer code, are available with this paper at the Biometrics website on Wiley Online Library.

measures are generally not available. In such settings, a person with a high level of post-epidemic antibody but without symptoms or viral culture may have two potential states: pre-season immunity to infection, or infected but without symptoms.

Failure to account for unobserved key quantities will likely bias the assessment of transmissibility and the effects of risk factors (Sugimoto et al., 2011) and may underestimate the variances of parameter estimates. Under the assumption of missing at random (MAR), unobserved quantities can be integrated out for proper statistical inference (Little and Rubin, 2002). For the infection times of asymptomatic cases, MAR means that whether an infection is symptomatic or not does not depend on the time of infection, when conditioning on the observed exposure history determined by observed contacts with infectious sources and covariates that modify the transmissibility. Analytic integration is possible in certain models, for instance, the one addressing unmeasured heterogeneity in baseline susceptibility (Longini and Halloran, 1996). Numeric integration via data-augmentation techniques can be effectively used for unobserved infection times. Within the Bayesian framework, Markov chain Monte Carlo (MCMC) algorithms have been used in different study settings for infectious diseases (Gibson, 1997; Auranen et al., 2000; O'Neill et al., 2000; Cauchemez et al., 2004; Yang, Longini and Halloran, 2009). The use of these models by epidemiological investigators are limited by either the heavy computational burden or the need to specify prior distributions for parameters.

Another option for data augmentation is the EM algorithm which has been used to augment the data with infection times (Yang, Longini and Halloran, 2007) or the transmission trees (Kenah, 2010) when only symptomatic infections are modeled. Direct application of the EM algorithm to asymptomatic infections may be numerically impossible due to the high-dimensionality of the unobserved infection times. In this work, we explore the applicability of an MCEM algorithm based on importance sampling (Levine and Casella, 2001) to the setting of infectious diseases. Taking advantage of the assumption that the close contact groups are independent and that missing data of high dimension often exist in only a few of the observed contact groups, we propose a hybrid algorithm that applies the MCEM algorithm to contact groups with high dimensional missing data and the EM algorithm to the others. To further expedite the algorithm, we propose a bootstrap approach to assess the Monte Carlo error due to sampling and account for such error in the variance estimation.

Our investigation was motivated by the surveillance study of influenza infections in families with school-aged children that was implemented in Seattle, Washington, during 1975–1979 (Fox et al., 1982a,b). In this study, a total of 639 families were recruited during Octobers and Novembers of 1975 and 1976. Blood samples were collected at four-month intervals, from which influenza subtype-specific hemagglutinin antibody levels were measured via hemagglutinin inhibition (HI) titration both before and after the influenza season. The pre-season HI titers can be used to assess susceptibility to the circulating influenza strains, and a substantial increase between the pre- and post-season HI titers indicates the occurrence of an infection. Illness information and nose-throat swabs were collected during regularly scheduled visits or when a new illness was reported. Data are relatively complete for the type of analysis considered here for two of the four winters: 1977–1978 and 1978–1979. The circulating strains are H3N2 (A/Victoria and A/Texas) in the former and H1N1 (A/USSR) in the latter. There were small herald outbreaks of H3N2 in the spring of 1977 and H1N1 in the spring of 1978, which immunized fully or partially a proportion of the study families for the subsequent major outbreaks in the winters. Our interest is to assess the transmissibility of the two strains and the associated effects of age and prior-immunity level measured by the HI titers. The epidemic curves (timeline of illness onsets in symptomatic infections) of the two epidemics are shown in Figure 1. Inclusion/exclusion criteria of study households for analysis is given in Web Appendix A. Numbers of individuals by infection status, age group

and pre-season HI titers used in this analysis are shown in Table 1. The household sizes are distributed, in the form of size:number, as (2:6, 3:41, 4:76, 5:31, 6:11, 7:8, 8:2, 9:1) in the earlier epidemic and as (2:5, 3:25, 4:48, 5:20, 6:8, 7:3, 8:3, 9:1) in the later one. Both epidemics lasted for a few months, and each has 20–30% households with one or more asymptomatic infections. The large numbers of potential configurations of individual states in these households make the traditional EM algorithm computationally expensive.

2. Methods

2.1 Data Structure

Consider a prospective study for the transmission of influenza in H households each of size n_h , $h = 1, \dots, H$, and let $N = \sum_{h=1}^H n_h$ be the total number of people in the study. The transmission processes in different households are assumed to be independent. Let day T be the stopping day of the epidemic. Let z_i indicate the immune status of person i , $i = 1, \dots, N$, prior to the influenza season, where $z_i = 1$ for immune and 0 for susceptible. The prior immunity could be due to either vaccination or previous infection. Let y_i and s_i be the binary indicators for the infection and symptom outcomes of individual i during the epidemic season. We partition the study population into four possible final states at the end of the epidemic: (1) prior immunity (i.e., $z_i = 1$, $y_i = 0$, $s_i = 0$), (2) susceptible, but escaped infection (i.e., $z_i = 0$, $y_i = 0$, $s_i = 0$), (3) symptomatic infection (i.e., $z_i = 0$, $y_i = 1$, $s_i = 1$), and (4) asymptomatic infection (i.e., $z_i = 0$, $y_i = 1$, $s_i = 0$). Individuals with the latter three states are susceptible prior to the influenza season. Let $\mathbf{x}_{it}$ be the vector of covariates associated with individual i on day t . We assume $\mathbf{x}_{it}$ is fully observed for all individuals.

Let t_i^* and $\tilde{t}_i$ be the infection day and infectiousness onset day of person i . For people who have escaped infection during the epidemic, we set $t_i^* = \tilde{t}_i = \infty$. For people with prior immunity, t_i^* and $\tilde{t}_i$ are not defined. For symptomatic infections, we assume that the incubation period (the time from infection to symptom onset) and the latent period (the time from infection to infectiousness onset) are the same; hence, $\tilde{t}_i$ is also the symptom onset day. We further assume that all infections, whether symptomatic or asymptomatic, share the same distributions of the latent period and infectious period, and that these distributions are known from empirical studies. One more assumption is that, given a fixed duration of the infectious period, the infectiousness (i.e., infectivity) level is constant over the infectious period. Let $f(l)$ be the probability that the latent period lasts for l days, and $g(l)$ the probability that the infectious period of person i is l days or longer. Although not observed in general, the infection time, t_i^* , can be integrated out using $f(l)$ conditioning on $\tilde{t}_i$. The quadruplet $\mathbf{u}_i = (z_i, y_i, s_i, \tilde{t}_i)$ is the complete individual data associated with the epidemic that may be fully or partially observed.

2.2 Likelihood based on complete data

We assume $z_i \sim \text{Bernolli}(\alpha)$, where α is the proportion of individuals with prior immunity, and $s_i \sim \text{Bernolli}(\phi)$, where ϕ is the probability of being symptomatic given infection. We assume the infection outcomes are generated by a chain binomial model (Yang, Longini and Halloran, 2007). Each susceptible individual is exposed to the risk of infection from casual contacts outside the household in addition to close contacts with infective household members. The casual contacts are generally not observed, and are, therefore, modeled as exposure to an unknown infection source that is assumed to be identical across the study population. For a susceptible individual, the daily unadjusted probability of infection by the unknown source is b , and the daily unadjusted probability of infection by an symptomatic infective household member is p . Let $\mathbf{x}_{ijt}$ be the subset of $(\mathbf{x}_{it}, \mathbf{x}_{jt})$ that are relevant to the close contact between individuals i and j on day t . We assume the prior immune status is

only associated with time-independent covariates, and denote the time-independent subset of $\mathbf{x}_{it}$ by $\mathbf{x}_i$. Let c_i be the household of person i and $\mathbf{1}(\text{condition})$ be the indicator function (1 if *condition* is true, 0 otherwise). Covariates can be adjusted for by the following logistic models:

$$\begin{aligned}\text{logit}(\alpha_i) &= \text{logit}(\alpha) + \mathbf{x}_i^\top \boldsymbol{\beta}_\alpha, & \text{logit}(\varphi_{it}) &= \text{logit}(\varphi) + \mathbf{x}_{it}^\top \boldsymbol{\beta}_\varphi, \\ \text{logit}(b_{it}) &= \text{logit}(b) + \mathbf{x}_{it}^\top \boldsymbol{\beta}_b, & \text{logit}(p_{ijt}) &= (-\infty)^{\mathbf{1}(c_j \neq c_i)} (\text{logit}(p) + \mathbf{x}_{it}^\top \boldsymbol{\beta}_p)^{\mathbf{1}(c_j = c_i)},\end{aligned}$$

where α_i , φ_{it} , b_{it} and p_{ijt} are effective counterparts of α , φ , b and p that are specific to each person or person-day, and $\boldsymbol{\beta}_\alpha$, $\boldsymbol{\beta}_\varphi$, $\boldsymbol{\beta}_b$ and $\boldsymbol{\beta}_p$ are covariate effects.

The probability of person i escaping infection during and up to day t are, respectively, $e_{it} =$

$(1 - b_{it}) \prod_{j: c_j = c_i} (1 - \theta^{1-s_j} p_{ijt} g(t - \tilde{t}_j + 1))$ and $q_{it} = \prod_{l=1}^t e_{il}$, where θ is the infectiousness level of asymptomatic cases relative to symptomatic cases. We assume that θ is known to avoid non-identifiability for small numbers of outcomes. Let $\boldsymbol{\psi} = \{b, p, \alpha, \varphi, \boldsymbol{\beta}_\alpha, \boldsymbol{\beta}_\varphi, \boldsymbol{\beta}_b, \boldsymbol{\beta}_p\}$ be the set of all the parameters. We suppress the covariates in all subsequent expressions to simplify notation. If elements in $\mathbf{u}_i$ are fully observed, the likelihood contributed by individual i is given by

$$L_{(i)}(\boldsymbol{\psi} | \mathbf{u}_i) = \alpha_i^{z_i} (1 - \alpha_i)^{1-z_i} \left\{ q_{iT}^{1-y_i} \left(\sum_t f(\tilde{t}_i - t) q_{i(t-1)} (1 - e_{it}) (\varphi_{it})^{s_i} (1 - \varphi_{it})^{1-s_i} \right)^{y_i} \right\}^{1-z_i}.$$

For household h , let $\mathbf{O}_h = \{\mathbf{u}_i: c_i = h \text{ and all elements of } \mathbf{u}_i \text{ are observed}\}$ and $\mathbf{U}_h = \{\mathbf{u}_i: c_i = h \text{ and some or all elements of } \mathbf{u}_i \text{ are not observed}\}$. Let $\mathbf{O} = \{\mathbf{O}_h: h = 1, \dots, H\}$ and $\mathbf{U} = \{\mathbf{U}_h: h = 1, \dots, H\}$. The household-level and population-level likelihoods based on the complete data are

$$L_h(\boldsymbol{\psi} | \mathbf{O}_h, \mathbf{U}_h) = \prod_{i: c_i = h} L_{(i)}(\boldsymbol{\psi} | \mathbf{u}_i)$$

and

$$L(\boldsymbol{\psi} | \mathbf{O}, \mathbf{U}) = \prod_{h=1}^H L_h(\boldsymbol{\psi} | \mathbf{O}_h, \mathbf{U}_h).$$

2.3 Hybrid EM-MCEM algorithm

We assume independence between households, and therefore imputation of missing quantities is performed at the household level. Let δ_h be the dimension of domain, or equivalently the number of possible realizations, of $\mathbf{U}_h$. Let $\{\mathbf{U}_{hk}^*: k = 1, \dots, \delta_h\}$ be the collection of all possible realizations of $\mathbf{U}_h$ for household h , $h = 1, \dots, H$. We set $\delta_h = 1$ if $\mathbf{U}_h$ is empty. For non-empty $\mathbf{U}_h$ we expect $\delta_h > 1$. The exact scale of δ_h depends on the data. For example, if $\tilde{t}_i$ of individual i of household h is the only unobserved quantity in that household, then the range of $\mathbf{U}_h$ is determined by $1 + I - \tilde{t}_i - T$ and $\delta_h = T - I$, where I is the minimum duration of the latent period. When T is large and the infectiousness onset

days are missing for multiple individuals in the household, δ_h could be too large for the EM algorithm to enumerate all possibilities. To avoid impractical enumeration, the MCEM algorithm proposed by Levine and Casella (2001) draws importance samples of $\mathbf{U}$ via MCMC conditioning on reasonable estimates of the parameters before the EM iterations, and weights the complete-data-based likelihood over these samples in the E-step. The number of importance samples is increased at each iteration if the conditional MC error swamps the changes in parameter estimates over the last iteration. The computational burden of this MCEM algorithm is still too heavy for the type of data we consider. We further reduce the burden by: (1) fixing the number of samples and factoring the overall MC error into the variance calculation, and (2) implementing MCEM on households with large, and EM on households with small, values of δ_h . Let J be the minimum value of δ_h of households on which MCEM is to be implemented.

2.3.1 The general algorithm—The hybrid EM-MCEM algorithm adapted from the importance-sampling-based MCEM algorithm of Levine and Casella (2001) is the following:

1. Choose the value of J to partition the households into three groups: $\Delta_{OBS} = \{h: \delta_h = 1\}$, $\Delta_{EM} = \{h: 1 < \delta_h < J\}$, and $\Delta_{MCEM} = \{h: \delta_h \geq J\}$. EM is implemented on households in Δ_{EM} , and MCEM is implemented on households in Δ_{MCEM} .
2. Choose an integer K to be the number of importance samples for the MCEM algorithm. The value of K is chosen to be reasonably large, depending on the computational feasibility and how well the MCMC samples are mixing.
3. Choose an initial value $\boldsymbol{\psi}^{(0)}$ for $\boldsymbol{\psi}$. For household $h \in \Delta_{MCEM}$, draw K samples of $\mathbf{U}_h$ from $\Pr(\mathbf{U}_h | \mathbf{O}_h, \boldsymbol{\psi}^{(0)})$ using a MCMC algorithm, and let these samples be $\hat{\mathbf{U}}_{hk}$, $k = 1, \dots, K$.
4. Set $\hat{\boldsymbol{\psi}}^{(0)} = \boldsymbol{\psi}^{(0)}$
5. At iteration $r = 0$,
 - a. Update the conditional probabilities for all $h \in \Delta_{EM}$:

$$\lambda_{hk}^{(r)} = \frac{L_h(\hat{\boldsymbol{\psi}}^{(r)} | \mathbf{O}_h, \mathbf{U}_{hk}^*)}{\sum_{l=1}^{\delta_h} L_h(\hat{\boldsymbol{\psi}}^{(r)} | \mathbf{O}_h, \mathbf{U}_{hl}^*)}, k=1, \dots, \delta_h,$$

and the importance weights for all $h \in \Delta_{MCEM}$:

$$\omega_{hk}^{(r)} = \frac{L_h(\hat{\boldsymbol{\psi}}^{(r)} | \mathbf{O}_h, \hat{\mathbf{U}}_{hk})}{L_h(\boldsymbol{\psi}^{(0)} | \mathbf{O}_h, \hat{\mathbf{U}}_{hk})}, k=1, \dots, K.$$

- b. Maximize

$$Q(\boldsymbol{\psi}, \hat{\boldsymbol{\psi}}^{(r)}) = \sum_{h \in \Delta_{OBS}} \ln L_h(\boldsymbol{\psi} | \mathbf{O}_h) + \sum_{h \in \Delta_{EM}} \sum_{k=1}^{\delta_h} \lambda_{hk}^{(r)} \ln L_h(\boldsymbol{\psi} | \mathbf{O}_h, \mathbf{U}_{hk}^*) + \sum_{h \in \Delta_{MCEM}} \frac{1}{\omega_h^{(r)}} \sum_{k=1}^K \omega_{hk}^{(r)} \ln L_h(\boldsymbol{\psi} | \mathbf{O}_h, \hat{\mathbf{U}}_{hk})$$

with regard to $\boldsymbol{\psi}$ to find $\hat{\boldsymbol{\psi}}^{(r+1)}$, where $\omega_h^{(r)} = \sum_{k=1}^K \omega_{hk}^{(r)}$ for $h \in \Delta_{MCEM}$.

Repeat this step until convergence in the estimates of $\boldsymbol{\psi}$, and denote the final estimate by $\hat{\boldsymbol{\psi}}$.

The traditional MCEM algorithm of Levine and Casella (2001) is a special case of the hybrid algorithm with $J = 2$. Like the traditional algorithm, the performance of this hybrid algorithm may be sensitive to the initial value $\psi^{(0)}$. We follow the suggestion of Levine and Casella (2001) that a few initial iterations be carried out with all weights set to 1 and that new samples be drawn based on updated estimates at each iteration, and $\psi^{(0)}$ be set to the estimates of the final iteration.

2.3.2 The importance sampling step—For the types of data we consider, it is convenient to use the Gibbs sampler to draw the importance samples in step (3). Let Λ_j be the collection of all possible realizations of complete data for each $u_j \in U_h$. The size of Λ_j is often small enough so that u_j can be directly sampled when conditioning on $\{u_j: j \neq i, c_j = c_i\}$. Let $U_{h(-i)} = \{u_j: j \neq i \text{ and } u_j \in U_h\}$. The conditional probability is

$$\Pr(u_i = u_i^* | O_h, U_{h(-i)}, \psi) = \frac{L_h(\psi | u_i^*, U_{h(-i)}, O_h)}{\sum_{u_i \in \Lambda_i} L_h(\psi | u_i, U_{h(-i)}, O_h)}.$$

For each household h , we start with some initial value of U_h , and alternate the sampling of u_j over all individuals with incomplete u_j until $K_0 + K$ samples of U_h are obtained. The K_0 burn-in samples are discarded and the last K samples are used in step (3). The numbers of importance samples need not be equal across households. We use the same value K for all households in Δ_{MCEM} for simplicity and illustration. How well the samples are mixing could be assessed by visually examining the distribution of samples over iterations for selected u_j 's. Quantification of the degree of mixing is possible if the infectiousness onset time $\tilde{t} = \{t_i: i = 1, \dots, N\}$ is the only variable that is subject to incomplete observation, but would be difficult when there is ambiguity about the final states, e.g., a person could either escape infection or be asymptomatic if infected, because the sampling space is then discrete and non-ordinal.

2.3.3 Variance estimation—For point estimation, one maximizes

$E_{U|O, \hat{\psi}^{(r)}} \ln L(\psi | O, U) = \sum_{h=1}^H E_{U_h | O_h, \hat{\psi}^{(r)}} \ln L_h(\psi | O_h, U_h)$ at each EM iteration. Taking advantage of the independence between households, each $E_{U_h | O_h, \hat{\psi}^{(r)}} \ln L_h(\psi | O_h, U_h)$ can be numerically evaluated using MCMC importance sampling separately, allowing the hybrid of EM and MCEM as well as different numbers and values of importance weights across households in Δ_{MCEM} . However, the estimation of the variance requires evaluation of

$$E_{U|O, \hat{\psi}} \left(\frac{d \ln L(\psi | O, U)}{d \psi} \right) \left(\frac{d \ln L(\psi | O, U)}{d \psi} \right)^{\tau},$$

which is not equal to

$$\left(E_{U|O, \hat{\psi}} \frac{d \ln L(\psi | O, U)}{d \psi} \right) \times \left(E_{U|O, \hat{\psi}} \frac{d \ln L(\psi | O, U)}{d \psi} \right)^{\tau}.$$

To facilitate variance evaluation, we need an equal number of importance samples for all households with $\delta_h > 1$. This can be attained by generating K samples based on $\psi^{(0)}$ for all households in Δ_{EM} , or by generating K new importance samples based on any parameter

value $\tilde{\psi} \approx \hat{\psi}$ for all households in Δ_{EM} or Δ_{MCEM} . The rationale for the latter approach is that, if K is sufficiently large, the algorithm should converge to the same final estimates $\hat{\psi}$ no matter which initial estimates it starts with, as long as the initial estimates are fairly close to $\hat{\psi}$. The former approach is a special case of the latter with $\tilde{\psi} = \psi^{(0)}$. To be general, we elaborate the latter approach. Denote the new importance samples by $\tilde{U}_{hk}$, $k = 1, \dots, K$, $h \in \Delta_{EM} \cup \Delta_{MCEM}$. To distinguish this set of importance sample from the one used for point estimation, we refer to the former as the variance-estimation, and the latter as the point-estimation, importance samples. Let $\tilde{U}_k = \{\tilde{U}_{hk}: h = 1, \dots, H\}$. Let $\tilde{\omega}_k = L(\hat{\psi}|\mathbf{O}, \tilde{U}_k)/L(\tilde{\psi}|\mathbf{O}, \tilde{U}_k)$, $k = 1, \dots, K$, and $\tilde{\omega} = \{\tilde{\omega}_k: k = 1, \dots, K\}$. Assuming that the MC error due to importance sampling can be ignored, the covariance matrix of the parameter estimates $\hat{\psi}$ is estimated by the inverse of the observed information matrix (Louis, 1982):

$$\widehat{V}^{-1}(\hat{\psi}, \tilde{\psi}, \tilde{U}) = \left(\frac{1}{\sum_{k=1}^K \tilde{\omega}_k} \sum_{k=1}^K \tilde{\omega}_k \frac{d \ln L(\psi|\mathbf{O}, \tilde{U}_k)}{d\psi} \right) \left(\frac{1}{\sum_{k=1}^K \tilde{\omega}_k} \sum_{k=1}^K \tilde{\omega}_k \frac{d \ln L(\psi|\mathbf{O}, \tilde{U}_k)}{d\psi} \right)^{\tau} \\ - \frac{1}{\sum_{k=1}^K \tilde{\omega}_k} \sum_{k=1}^K \tilde{\omega}_k \left\{ \frac{d^2 \ln L(\psi|\mathbf{O}, \tilde{U}_k)}{d\psi^2} + \left(\frac{d \ln L(\psi|\mathbf{O}, \tilde{U}_k)}{d\psi} \right) \left(\frac{d \ln L(\psi|\mathbf{O}, \tilde{U}_k)}{d\psi} \right)^{\tau} \right\} \Big|_{\psi=\hat{\psi}}.$$

Another special case worth noting is $\tilde{\psi} = \hat{\psi}$ which yields $\tilde{\omega}_k = 1$ for all k . For the traditional MCEM algorithm, there is no need to draw new importance samples, and one can use $\tilde{U} = \hat{U}$ and $\tilde{\psi} = \psi^{(0)}$, as long as the numbers of importance samples are equal across households.

2.3.4 Assessing the MC error—In traditional MCEM algorithms, there are steps to assess whether the conditional MC error (MC variation of $\psi^{(r+1)}|\psi^{(r)}$) swamps out the E-M step, and if so, how to increase the number of importance samples. The assessment of the conditional MC errors in previous work was mostly based on the central limit theorem, so that the sampling of $\mathbf{u}_i \in U_h$, which is computationally expensive, can be skipped. Booth and Hobert (1999) gave a sandwich formula for the conditional MC variance when the samples are independent. Using subsamples drawn at the renewal times of the MCMC samples, Levine and Casella (2001) estimated the conditional MC variance for the weighted average of the score functions, but not for the parameter estimates themselves.

We found in simulation studies that the assessemnt of conditional MC error leads to extremely slow convergence rate, even with a moderate sample size (500 people) and a moderate tolerance level of relative error in parameter estimates, such as 10^{-6} . Alternatively, it is more practical to choose a reasonably large K and factor the total MC error directly into the variance estimates of the parameters. By total MC error we mean the variation in the final parameter estimates due to the sampling. Unlike the conditional MC error, the total MC error needs to be assessed only once after convergence; therefore, sampling- and resampling-based approaches are computationally affordable.

An intuitive approach is to generate via MCMC multiple sets of new point-estimation importance samples based on $\psi^{(0)}$, to which we refer as the exact approach. We propose a more efficient resampling-based approach as follows: For each household $h \in \Delta_{MCEM}$, assume that the K point-estimation importance samples of U_h are well mixed such that their empirical distribution approximates $\Pr(U_h|\psi^{(0)}, \mathbf{O}_h)$. We draw via bootstrap M sets of point-estimation importance samples, denoted by $U_{hk(m)}^{\#}$, $k = 1, \dots, K$, $m = 1, \dots, M$, independently from the K original samples. For each m , we obtain parameter estimates $\hat{\psi}_m^{\#}$ by implementing step (5) of the MCEM algorithm based on $\{U_{hk(m)}^{\#}: h \in \Delta_{MCEM}, k = 1, \dots, K\}$. The sample variance of $\{\hat{\psi}_m^{\#}: m = 1, \dots, M\}$ measures the MC error of the MCEM estimator $\hat{\psi}$. Let $\widehat{V}_m^{\#} = \widehat{V}^{-1}(\hat{\psi}_m^{\#}, \hat{\psi}_m^{\#}, \tilde{U}_m)$ be the estimated covariance matrix associated with

$\hat{\psi}_m^\#$. To further reduce the computational burden, the calculation of $\hat{V}_m^\#$ can be based on the same set of variance-estimation importance samples for all m , e.g., $\psi_m = \hat{\psi}$ and $\tilde{O}_m = \tilde{O}$, where $\tilde{O}$ is drawn from $\Pr(U_h | O_h, \hat{\psi})$ using MCMC for all $h \in \Delta_{EM} \in \Delta_{MCEM}$. The covariance matrix of $\hat{\psi}$ with the MC error factored in is estimated by

$$\hat{V}^\# = \frac{1}{M} \sum_{m=1}^M \left(\hat{\psi}_m^\# - \frac{1}{M} \sum_{m=1}^M \hat{\psi}_m^\# \right)^2 + \frac{1}{M} \sum_{m=1}^M \hat{V}_m^\#. \quad (1)$$

3. Simulation Study

We simulate 30-day epidemics in a population composed of 100 households under the following setting. The household sizes are distributed, in the form of size:probability, as (3:0.1, 4:0.2, 5:0.3, 6:0.2, 7:0.1, 8:0.05, 9:0.05), with a mean size of 5.4. The parameter values are set to $b = 0.01$, $p = 0.05$, $\alpha = 0.2$, and $\phi = 0.7$. A single time-independent binary covariate was generated for each individual to modify the transmission probabilities via $\text{logit}(b_{ij}) = \text{logit}(b) + \beta x_i$ and $\text{logit}(p_{ij}) = \text{logit}(p) + \beta x_j$, where the corresponding odds ratio is $\text{OR} = \exp(\beta) = 0.6$. The covariate has no effect on prior immunity or pathogenicity. The following configuration for the natural history, similar to seasonal influenza (Elveback, Fox and Ackerman, 1976), are used: a latent period with a mean of 2 days and $(f(1), f(2), f(3)) = (0.21, 0.58, 0.21)$, and an infectious period with $(g(1), \dots, g(8)) = (1, 1, 1, 0.8, 0.6, 0.3, 0.1, 0)$. This setting produces on average an attack rate (proportion of infections among the total population) of about 38% with around 50 asymptomatic infections for whom data-augmentation is necessary. For estimation we assume θ is known. For the maximization step, a tolerance rate of 10^{-5} for the relative error in every parameter is used to determine convergence.

We first investigate, via both the traditional MCEM and the hybrid EM-MCEM algorithms, the performance of the resampling approach for assessing the total MC error by comparing it with the exact approach. For illustration, only the infectiousness onset times $\tilde{t}$ of asymptomatic cases are subject to missingness, and the range of possible values for each missing $\tilde{t}_i$ is 2–30 days, assuming a minimum latent period of one day. We set $K_0 = 200$, $J = 100$ and $M = 500$, and vary K from 500 to 5000. The MC variances of the estimates for b , p and OR obtained from the two approaches under the two algorithms are given in Table 2. The MC variance for the estimates of α and ϕ are constantly 0 as the proportions of pre-immune individuals in the whole population and asymptomatic individuals among all infections do not vary over the new sampling runs. The bootstrap variance estimates approximate the exact estimates reasonably well for all three parameters at all values of K . The absolute differences in the estimated total MC variances between the two approaches decreases with increasing K . Decrease in relative distance, for example, the absolute difference scaled by either estimate, is not expected, because the total MC variance itself converges to 0 with increasing K .

With the current simulation setting, the time ratio of the bootstrap approach to the exact approach is at the scale of 10^{-5} for the sampling phase and 0.6–0.7 for the whole process of estimating the total MC variance, regardless of the algorithm used or the value of K . Relative to the traditional MCEM algorithm, the hybrid EM-MCEM algorithm reduced the total MC variance by more than 50% for all three parameters at all values of K . Table 2 also suggests that the MC variance is roughly proportional to $1/K$, as expected. With a tolerance rate of 10^{-5} for the relative error in $\hat{b}$, if the true value of b is at the 10^{-2} scale, a scale of 10^{-7} is needed for the standard deviation of $\hat{b}$, or equivalently, 10^{-14} for the MC variance, which translates into a sample size of about 5 million for the traditional MCEM algorithm

and 1.5 million for the hybrid algorithm. It is therefore much more practical to factor the MC error into the variance estimation using (1) rather than to control it by increasing the number of importance samples. In addition, the component $\frac{1}{M} \widehat{\mathbf{V}}_m^\#$ in (1), which measures the uncertainty inherent in epidemic-generation, is at the scales of 10^{-6} for b , 10^{-5} for p , and 10^{-2} for the OR (not shown in the table), implying that the MC error is ignorable in magnitude under this simulation setting.

We then investigate the performance in estimation of the hybrid EM-MCEM algorithm versus the traditional MCEM algorithm. In addition to the infectiousness onset times of asymptomatic infections, one more layer of missingness is introduced in this investigation. Among people whose final states are either prior immunity ($z_i = 1, y_i = 0, s_i = 0$) or escaped infection ($z_i = 0, y_i = 0, s_i = 0$), there is a proportion (denoted by ξ) with their exact states unknown. We vary θ over 0.25, 0.5 and 0.75 to examine the impact of the amount of missing data on the inference. As θ is assumed known for inference, sensitivity analysis is performed by deviating the assumed value to the two extreme values, 0 and 1, for the hybrid EM-MCEM algorithm. The MC variation is ignorable in magnitude and thus is not considered in the calculation of variances and confidence interval (CI) coverages.

Results of comparing the two algorithms are based on 500 simulated epidemics and presented in Table 3. When the assumed value of θ is correct, most parameter estimates appear to be unbiased, regardless of the value of ξ , except for a slight upward bias for OR which will diminish for larger population sizes or longer epidemics. Interestingly, higher values of ξ , which correspond to higher uncertainty levels of the data, only slightly increase the standard deviation (SD) estimates for all parameters except for ϕ . However, setting $\xi = 1.0$ leads to non-convergence in most simulated epidemics. This observation suggests that it is good enough to determine the prior immunity status in a small proportion of uninfected people, but no attempt of such determination at all will cause a nonidentifiability issue if one intends to estimate the prior immunity level. That estimation for ϕ is not affected by the value of ξ is expected as the information about ϕ only relies on the infections. The traditional MCEM algorithm yielded almost identical results as the hybrid algorithm, with only a slight difference in the CI coverage for α .

Misspecification of the value of θ in estimation does affect the inference about b and p but not other parameters. Assuming $\theta = 0$ leads to minor overestimation and slightly lower CI coverage than the desired level for b and p . Conversely, the direction of bias changes to downwards when $\theta = 1$ is assumed; nevertheless, the CI coverage for b seems relatively insensitive to the bias. The estimate for the odds ratio would not be affected by the value of θ as long as the covariate value of a susceptible is independent of his or her exposure level, e.g., independent of whether or not being exposed to asymptomatic infections. For the proportion of pre-season immunity, α , and the probability of symptom onset given infection, ϕ , their information is determined by whether infection occurs or not rather than by the infectiousness level of asymptomatic infections,

With $\xi = 0.5$, $\theta = 0.5$, and other simulation parameters the same as in Table 3, the computational efficiency is compared between the traditional MCEM and the hybrid EM-MCEM algorithms in Figure 2 for a few settings of J and K . Panel (a) shows a highly skewed distribution of δ_h , the household-specific numbers of possible realizations of missing final states, over all households and 100 simulated epidemics, with the majority of the values below 50 and a few larger than one million. For each setting of J and K , the average time in seconds over the 100 epidemic is measured for each of three stages: importance sampling, EM iterations, and variance calculation. The measured times are shown in panels (b), (c) and (d) for $K=500, 1000$ and 2000 respectively. In each panel, $J=2$

corresponds to the traditional MCEM algorithm, and the hybrid algorithm approaches the traditional EM algorithm as $J \rightarrow \infty$.

As J increases, the times for the sampling and EM iterations stages first decrease until J reaches thousands and begin to rise thereafter, the exact turning point depending on K . The time for variance calculation is minimal when $J = 2$, as no new sampling is necessary, and is constant when $J > 2$ but constitutes only a small fraction of the overall time. In most scenarios, the overall computational time seems to be primarily driven by the sampling stage, except when J is extremely large, e.g., 10^6 , where the EM iterations stage become dominant and the overall time becomes unfavorable as compared to $J = 2$. The slight increases in the time of the sampling stage at large values of J are attributed to the initial EM iterations to obtain $\phi^{(0)}$ in this stage. In terms of computational efficiency under our simulation setting, the hybrid algorithm with a wide range of J outperforms the traditional MCEM algorithm, and the relative efficiency reaches 50% at appropriate values of J . Not surprisingly, the traditional EM algorithm is the most time-inefficient with such a high dimension of individual-level missing data.

4. Data Analysis

The daily probability of infection due to the unknown source accounts for the risk due to casual contacts outside households; therefore it is reasonable to associate its variation over time with the changes in the observed total number of cases during each epidemic course, assuming that the study households are representative of all infected households. For simplicity, we assign to each day the average daily number of cases in the 7-day window centered around that day, and refer to the moving averages as the expected numbers. These expected numbers are then divided by their average over all days, so that the rescaled expected numbers, denoted by W_t for day t , have a mean 1. The time-dependent daily probability of infection due to the unknown source is given by

$\text{logit}(b_{it}) = \text{logit}(b) + x_{it}^T \beta_b + \ln(W_t)$. The rescaled expected numbers are shown by the dashed lines in Figure 1. In this analysis, age groups (0: < 18 years, 1: 18+ years) and pre-season HI titers of susceptible individuals are adjusted for in the regression models:

$$\begin{aligned}\text{logit}(\varphi_{it}) &= \text{logit}(\varphi) + \beta_{\varphi 1} \text{AGE}_i + \beta_{\varphi 2} \text{TITER}_i, \\ \text{logit}(b_{it}) &= \text{logit}(b) + \beta_{b1} \text{AGE}_i + \beta_{b2} \text{TITER}_i + \ln(W_t), \\ \text{logit}(p_{ijt}) &= (-\infty)^{\mathbf{1}_{(c_j \neq c_t)}} (\text{logit}(p) + \beta_{p1} \text{AGE}_i + \beta_{p2} \text{TITER}_i)^{\mathbf{1}_{(c_j = c_t)}},\end{aligned}$$

with the constraints $\beta_{b1} = \beta_{p1}$ and $\beta_{b2} = \beta_{p2}$. The constraints are based on the assumption that the covariate effects on transmissibility are identical regardless of infection sources, which is plausible as we only consider the covariates of the susceptible individuals. Pre-season HI titer levels are grouped into four and three categories for the 1977–1978 and 1978–1979 epidemics, respectively, as defined in Table 3. Accordingly, TITER_j is a vector of three and two binary indicators for the two epidemics, respectively. In this analysis, $K_0 = 200$ and $K = 1000$. The sampled infectiousness onset days of asymptomatic cases in two households of each epidemic were plotted in Web Figure 1 (Web Appendix B) and show satisfactory mixing.

Results are shown in Table 4, with the relative infectiousness level θ , which we do not estimate directly, varying over 0, 0.5 and 1. Estimates of parameters at $\theta = 0.5$ are reported as the primary results. Interestingly, the two subtypes of influenza had similar transmissibility in terms of both b and p . The secondary attack rate (SAR) is the baseline probability that a susceptible is infected by a case during his or her infectious period.

Mathematically, $SAR = 1 - \prod_{l=0}^{\delta-1} (1 - pg(l))$. The SAR presented in Table 4 indicates that the probability a susceptible child with the lowest HI titer level gets infected when exposed to a symptomatic infective household member is about 0.27 for both subtypes in these epidemics. The H3N2 strain was somewhat more pathogenic than the H1N1 strain, with the probability of developing symptoms being 0.81 for the former and 0.62 for the latter, but their confidence intervals are overlapping. For the effects of age group and pre-season HI titer level, odds ratios instead of regression coefficients are presented, and are denoted by $OR_{bp,covariate}$ for transmissibility and $OR_{\phi,covariate}$ for pathogenicity, covariate=AGE or TITER.

Children were more susceptible to both infection and the disease (symptoms) than adults, although the CIs for $OR_{\phi,AGE}$ do include 1. The age effects on pathogenicity are similar between the two subtypes, adults having about 40–55% reduction in the odds of developing symptoms. The effects of pre-season HI titer levels on transmissibility share the same trend between the two subtypes, that is, the level adjacent to the reference level was not protective, but the next and higher levels were. For H3N2, the titers of 40 and 80+ reduced the odds of infection by 50% and 78%, respectively. For H1N1, the titer of 40 or above was highly protective, reducing the odds of infection by 91%, but the titers of 10–20 did not show protection. A further examination of the H1N1 data found that the majority of cases were children (144/149), and the majority of children in the second HI titer level (63/80) had titer values of 10. In addition, the majority of adults (158/220) had titer values of 20 or more, and three out of the five adult cases had a pre-season titer value of 10. An analysis with a finer grouping of pre-season HI titers (< 10, 10, 20, 40) but without modeling pathogenicity for the H1N1 epidemic yielded OR estimates of 0.92 (95% CI:0.62, 1.36), 0.21 (95% CI:0.065, 0.69), and 0.08 (95% CI:0.01, 0.61), confirming the protective effect of titers ≥ 20. For H3N2, the titers of 40+ reduced the odds of disease by 60–70%. For H1N1, only a small number of cases had titers ≤ 10, so these titers are combined into one level for the estimation regarding the OR for pathogenicity. It seems surprising that, compared to the base titer level, the higher level doubled the odds of disease, but this effect is not statistically significant.

The variation in θ changes the estimates for transmission probabilities to a certain extent, with a general association of lower estimates with higher infectiousness levels of asymptomatic infections. If the extra exposure to asymptomatic infections does not change the ratio of existing exposure levels to symptomatic infections between infected and escaped people, which is satisfied under MAR (being symptomatic or not is independent of the time of infection given observed data), then such association is expected as the same proportion of infections among initial susceptibles is explained by more infective sources when asymptomatic infections are assumed infectious. The estimation of p for the H1N1 epidemic is quite sensitive to the value of θ , leading to a 50% jump in the SAR estimate when θ moves from 0 to 1. The estimates for the ORs are highly resistant to the variation in θ for both subtypes. Further sensitivity analyses were performed with shorter and longer means for the latent (1.6 and 2.5 days) and infectious (3.8 and 5.8 days) periods within reasonable ranges. Very mild changes occurred in the estimates for b and p in response to the variation of the latent-period distribution. With a longer infectious period, the estimates for p decreased moderately, but the estimates for the SARs remained unchanged. This is expected because a reasonable variation of the distribution of the infectious period should not drastically change the total amount of risks delivered by all cases. All OR estimates are highly robust to the changes in the assumptions about natural history.

5. Discussion

The hybrid MCEM algorithm is an extension of the traditional importance-sampling-based MCEM algorithm in Levine and Casella (2001) to account for missing data in settings where independent units have different dimensions of domain of missing data and integration over extremely large domains may be numerically impractical. The use of this algorithm is demonstrated in one such setting, infectious disease transmission data among close contact groups in which individuals may have unknown final disease states. Via simulating influenza transmission within households we show that the hybrid algorithm is computationally more efficient than the traditional MCEM or EM algorithms. The optimal partition point J may be difficult to find in advance, mainly because the number of EM iterations is unknown. A practical guideline is to apply MCEM only to units with very high dimension of missing data. If the clusters of transmission under study is much larger than households, such as school transmission studies, there could be much higher numbers of asymptomatic infections, and the computational efficiency of the MCEM component of the hybrid algorithm would be even more appreciable.

To avoid the slow convergence problem that often occurs when conditional MC error is assessed in the presence of high-dimensional missing data, we fix the number of importance samples and assess the total MC error. Such assessment can be further expedited via the bootstrap of importance samples. However, this approach does not guarantee monotonic increase in the likelihood over the EM iterations. One way to avoid far-from-optimal estimates is to start the algorithm from multiple initial estimates. If the multiple estimation processes do not agree on the final estimates, then K should be increased. The hybrid EM-MCEM algorithm can also be coupled with the approach of assessing conditional MC error and increasing the number of importance samples in Levine and Casella (2001), whenever computationally affordable, and should converge faster than the traditional MCEM algorithm, as the hybrid algorithm has smaller MC errors as shown in Table 2.

The EM algorithm is known to converge slowly in some applications, which we have not experienced in our investigation. In most cases, there were no more than ten EM iterations before convergence. When slow convergence is of concern, various EM accelerators may be considered, for example, the quasi-Newton accelerator QN1 proposed by Jamshidian and Jennrich (1997). Some accelerators such as the conjugate gradient accelerator (Jamshidian and Jennrich, 1993) require knowledge about the likelihood based on the observed data and are therefore not applicable to the data considered here as $L(\boldsymbol{\psi}|\mathbf{O})$ is generally intractable. The parameter expansion (PX) approach (Liu, Rubin and Wu, 1998) is another option for acceleration. A possible use of this approach is to use the regression model

$\text{logit}(b_{it}) = \text{logit}(b) + \mathbf{x}_{it}^T \boldsymbol{\beta}_b + \gamma \ln(W_i)$ to adjust for the relative infectivity of the unknown source, where the extra parameter $\gamma \in (0, \infty)$. At iteration k of the PX EM algorithm, perform the E-step and the M-step with $(\boldsymbol{\psi}^{(k-1)}, \gamma^{(k-1)} = 1)$. Denote the obtained parameter estimates by $(\boldsymbol{\psi}_*^{(k)}, \gamma_*^{(k)})$, the formal estimates used for the next iteration is given by $\gamma^{(k)} = 1$ and $\boldsymbol{\psi}^{(k)} = \boldsymbol{\psi}_*^{(k)}$ except that $\text{logit}(b^{(k)}) = \text{logit}(b_*^{(k)})/\gamma_*^{(k)}$ and $\boldsymbol{\beta}_b^{(k)} = \boldsymbol{\beta}_{b*}^{(k)}/\gamma_*^{(k)}$. The PX EM and the usual EM algorithms eventually yield identical parameter estimates. The rationale is to use γ to allow more flexibility in the variance of the samples of $\mathbf{U}_h$ whose sampling was only guided by the risk from the unknown source. Yet whether such an expansion scheme is appropriate or not needs further research.

We discussed the application of the hybrid EM-MCEM algorithm for infectious disease data under the prospective design in which households are followed from the beginning of an epidemic. The algorithm can be generalized to the case-ascertained design in which households are enrolled only when index cases are ascertained by symptom onsets. However,

caution should be taken for determining the sampling range of the infectiousness onset times for asymptomatic cases, as the starting time of the epidemic is generally unknown. It may be reasonable to include in the sampling range a few days before the symptom onset time of the index case, so that all possible realizations of exposure of susceptibles to the asymptomatic infections are taken into account. If an asymptomatic person could be infected much earlier than the index case, then prior immunity can be considered as a possible state in addition to asymptomatic infection, so as to avoid a too wide sampling range.

Using only final infection states but no information on symptom onset times and assuming that asymptomatic infections are as infectious as symptomatic infections, a final size model (Longini et al., 1982) estimated the infection SARs as 0.21 (95% CI:0.17, 0.25) for the 1977–1978 H3N2 epidemic and 0.31 (95% CI:0.22, 0.39) for the 1978–1979 H1N1 epidemic. No covariate was explicitly adjusted for in those analyses, but only subjects with HI titers ≥ 640 (nearly the whole population as only three individuals have higher titers) were considered susceptible in the H3N2 epidemic, and only subjects < 20 years old with HI titers ≥ 10 were considered susceptible in the H1N1 epidemic. In addition, the risk of infection from the unknown source was assumed constant. Applying similar cutoffs for the HI titers and assumption about relative infectivity of asymptomatic infections, our infection SAR estimates are 0.16 (95% CI:0.12, 0.21) for the H3N2 epidemic and 0.27 (95% CI:0.19, 0.37) for the H1N1 epidemic. The discrepancy in the SAR estimates for H3N2 is most likely a consequence of different inclusion/exclusion criteria of households for analysis. The numbers of households used in Longini et al. (1982) are 159 and 93 for the two epidemics, both smaller than the numbers we used. The distribution of the final sizes was given in Longini et al. (1982) for the H1N1 epidemic but not for the H3N2 epidemic, and our H1N1 data give a fairly similar distribution.

Conditioning on the pre-season HI titer level, strong age effects on reducing susceptibility to infection were detected in both epidemics, which may be indicative of the role of either a T-cell response or the neuraminidase antibody level (not available) in preventing transmission. The larger age effect on transmissibility for H1N1 relative to H3N2 is likely due to the fact that it was the first emergence of the H1N1 virus after vanishing in late 1950s (Xu et al., 2010), and therefore only adults of 20 years or above had previous exposure to this virus.

As θ increases over the range of $[0, 1]$, the full likelihood for the complete data increases for the H3N2 epidemic and decreases for the H1N1 epidemic (results not shown). However, we did not pursue a profile likelihood approach for the estimation of θ for two reasons. First, an estimate of $\theta > 1$ would be difficult to interpret in biology as influenza is a respiratory disease whose transmission is primarily driven by symptoms such as coughing and sneezing. Second, how reliable θ can be estimated and the sample size required for such estimation are under future investigation. Our assumptions concerning the knowledge about θ as well as the distribution of the infectious period are made to avoid non-identifiability of key quantities such as p and covariate effects in the presence of uncertainty about the final status. Previous studies have shown that, even with final outcomes completely observed, a relative large population is necessary for more flexibility in modeling the infectious period (Yang, Longini and Halloran, 2009; Kenah, 2010).

Our simulation study indicates that the inference is not greatly compromised even when the proportion of people with uncertain final states is relatively high but is totally impossible when such proportion reaches 100%. The usefulness of a validation set (individuals with lab-confirmed final states) with appropriate statistical methods has been previously found and emphasized in the setting of infectious diseases (Halloran et al., 2003; Yang et al., 2010). However, such usefulness should by no means diminish the importance of the efforts to ascertain as complete information as possible via carefully designed studies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Institute of General Medical Sciences MIDAS grant U01-GM070749 and the National Institute of Allergy and Infectious Diseases grant R37-AI032042.

References

- Auranen K, Arjas E, Leino T, Takala AK. Transmission of pneumococcal carriage in families: a latent Markov process model for binary longitudinal data. *Journal of the American Statistical Association*. 2000; 95:1044–1053.
- Booth JG, Hobert JP. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society B*. 1999; 61:265–285.
- Cauchemez S, Carrat F, Viboud C, Valleron AJ, Böelle PY. A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. *Statistics in Medicine*. 2004; 23:3469–3487. [PubMed: 15505892]
- Elveback LR, Fox JP, Ackerman E. An influenza simulation model for immunization studies. *American Journal of Epidemiology*. 1976; 103:152–165. [PubMed: 814808]
- Fox JP, Hall CE, Cooney MK, Foy HM. Influenza virus infections in Seattle families, 1975–1979. I. Study design, methods and the occurrence of infections by time and age. *American Journal of Epidemiology*. 1982a; 116:212–227. [PubMed: 7114033]
- Fox JP, Hall CE, Cooney MK, Foy HM. Influenza virus infections in Seattle families, 1975–1979. II. Pattern of infection in invaded households and relation of age and prior antibody to occurrence of infection and related illness. *American Journal of Epidemiology*. 1982b; 116:228–242. [PubMed: 7114034]
- Gibson GV. Markov chain Monte Carlo methods for fitting spatio-temporal stochastic models in plant epidemiology. *Applied Statistics*. 1997; 46:215–233.
- Halloran ME, Longini IM, Gaglani MJ, Piedra PA, Chu H, Herschler GB, Glezen WP. Estimating efficacy of trivalent, cold-adapted, influenza virus vaccine (CAIV-T) against influenza A (H1N1) and B using surveillance culture. *American Journal of Epidemiology*. 2003; 158:305–311. [PubMed: 12915495]
- Jamshidian M, Jennrich RI. Conjugate gradient acceleration of the EM algorithm. *Journal of the American Statistical Association*. 1993; 88:221–228.
- Jamshidian M, Jennrich RI. Acceleration of the EM algorithm by using Quasi-Newton methods. *Journal of the Royal Statistical Society B*. 1997; 59:569–587.
- Kenah E. Contact intervals, survival analysis of epidemic data, and estimation of R_0 . *Biostatistics*. 201010.1093/biostatistics/kxq068
- Levine RA, Casella G. Implementations of the Monte Carlo EM algorithm. *Journal of Computational and Graphical Statistics*. 2001; 10:422–439.
- Little, RJA.; Rubin, DB. *Statistical Analysis with Missing Data*. New York: John Wiley; 2002.
- Liu C, Rubin DB, Wu YN. Parameter expansion to accelerate EM: the PX-EM algorithm. *Biometrika*. 1998; 85:755–770.
- Longini IM, Koopman JS, Monto AS, Fox JP. Estimating household and community transmission parameters for influenza. *American Journal of Epidemiology*. 1982; 115:736–751. [PubMed: 7081204]
- Longini IM, Halloran ME. A frailty mixture model for estimating vaccine efficacy. *Journal of the Royal Statistical Society C*. 1996; 45:165–173.
- Louis TA. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society B*. 1982; 44:226–233.

- O'Neill P, Balding DJ, Becker NG, Eerola M, Mollison D. Analysis of infectious disease data from household outbreaks by Markov chain Monte Carlo methods. *Applied Statistics*. 2000; 49:517–542.
- Scharfstein DO, Halloran ME, Chu H, Daniels MJ. On estimation of vaccine efficacy using validation samples with selection bias. *Biostatistics*. 2006; 7:615–629. [PubMed: 16556610]
- Sugimoto, JD.; Yang, Y.; Halloran, ME.; Oulstad, B.; Bagwell, D.; Dean, B.; Mascola, L.; Bancroft, E.; Longini, IM. Accounting for unobserved immunity and asymptomatic infection in the early household transmission of the pandemic influenza A (H1N1) 2009. 2011. Under review
- Xu R, Ekiert DC, Krause JC, Hai R, Crowe JE Jr, Wilson IA. Structural basis of preexisting immunity to the 2009 H1N1 pandemic influenza virus. *Science*. 2010; 328:357–360. [PubMed: 20339031]
- Yang Y, Longini IM, Halloran ME. A data-augmentation method for infectious disease incidence data from close contact groups. *Computational Statistics and Data Analysis*. 2007; 51:6582–6595. [PubMed: 18704156]
- Yang Y, Longini IM, Halloran ME. A Bayesian model for evaluating influenza antiviral efficacy in household studies with asymptomatic infections. *Biostatistics*. 2009; 10:390–403. [PubMed: 19202152]
- Yang Y, Halloran ME, Daniels M, Longini IM, Burke DS, Cummings DAT. Modeling competing infectious pathogens from a Bayesian perspective: Application to influenza studies with incomplete laboratory results. *Journal of the American Statistical Association*. 2010; 105:1310–1322. [PubMed: 21472041]

Figure 1.

Epidemic curves of illness onsets of symptomatic cases in the Seattle influenza seasons. The dashed lines denote the relative infectivity levels of the unknown source over time, calculated using rescaled 7-day moving averages of daily case numbers. (a) 1977–1978 season, H3N2; (b) 1978–1979 season, H1N1.

Figure 2.

Comparison of the computational efficiency between the traditional MCEM algorithm and the hybrid EM-MCEM algorithm for disease transmission in independent households. Panel (a) shows the distribution of δ_h , the numbers of possible realizations of U_h , across all households over 100 simulated epidemics. Black diamonds along the horizontal axis indicate the values of J used in other panels, Panels (b), (c) and (d) plot the average computational times of importance sampling (black), EM iterations (white) and variance calculation over simulated epidemics versus J , the minimum value of δ_h to apply the MCEM algorithm.

\$watermark-text

\$watermark-text

\$watermark-text

Summary of infection outcomes in terms of the numbers of escapes, symptomatic (sym.) and asymptomatic (asym.) infections, by age group and pre-season antibody level, for the influenza H3N2 epidemic during 1977–1978 and the H1N1 epidemic during 1978–1979 in Seattle.

Table 1

Status Category	1977–1978, H3N2				1978–1978, H1N1			
	Total	Escape	Infection		Total	Escape	Infection	
			Sym.	Asym.			Sym.	Asym.
Population	741	567	121	53	479	330	98	51
Age Group								
Child	401	270	94	37	259	115	95	49
Adult	340	297	27	16	220	215	3	2
Baseline antibody titer level [†]								
1	107	69	30	8	188	81	66	41
2	210	134	57	19	177	136	31	10
3	190	155	19	16	113	112	1	0
4	227	202	15	10				
Unknown	7	7	0	0	1	1	0	0

[†] HI titer levels for H3N2: (1) 10, (2) 20, (3) 40, (4) 80.

HI titer levels for H1N1: (1) < 10, (2) 10 and 20, (3) 40.

Table 2

Comparing the total MC variances based on $M = 500$ new importance samplings between the bootstrap approach and the exact approach. Results for α and φ are not presented as the estimates are constant. A single 30-day epidemic in 100 households was simulated, with $b = 0.01$, $p = 0.05$, $\alpha = 0.2$, $\varphi = 0.7$, $\theta = 0.5$, and $OR = 0.6$. θ is assumed known in estimation. The burn-in number of importance samples is $K_0 = 200$.

K	Traditional MCEM ($J = 2$)						Hybrid EM-MCEM ($J = 100$)					
	$b(\times 10^{-11})$		$p(\times 10^{-9})$		$OR(\times 10^{-8})$		$b(\times 10^{-11})$		$p(\times 10^{-9})$		$OR(\times 10^{-8})$	
	Boot.	Exact	Boot.	Exact	Boot.	Exact	Boot.	Exact	Boot.	Exact	Boot.	Exact
500	9.77	9.64	17.88	18.20	13.52	13.80	3.13	3.37	6.19	6.20	5.34	5.43
1000	4.91	5.14	8.64	9.29	6.32	6.53	1.44	1.59	2.78	2.94	2.54	2.67
2000	2.40	2.29	4.44	4.70	3.46	3.46	0.74	0.87	1.48	1.65	1.41	1.50
5000	0.91	0.86	1.55	1.51	1.21	1.15	0.31	0.31	0.60	0.56	0.56	0.49

Table 3

Performance of the MCEM algorithms: mean estimates, standard deviations (SD) and 95% CI coverages are based on 500 simulated 30-day epidemics among 100 households each of size 5, with $b = 0.01$, $p = 0.05$, $\alpha = 0.2$, $\varphi = 0.3$ and $\theta = 0.5$. θ is assumed known. Burn-in and final numbers of importance samples are $K_0 = 200$ and $K = 1000$.

Assumed value of θ	Parameter	True Value	$\xi = 0.25$			$\xi = 0.5$			$\xi = 0.75$					
			Mean	x SD	95% CI Coverage	Mean	x SD	95% CI Coverage	Mean	x SD	95% CI Coverage			
Traditional MCEM ($J = 2$)														
0.5	b	0.01	0.010	0.0015	94.8%	0.010	0.0015	94.4%	0.010	0.0016	95.6%			
	p	0.05	0.050	0.0098	94.6%	0.050	0.010	95.2%	0.050	0.011	95.6%			
	φ	0.70	0.70	0.041	93.4%	0.70	0.041	93.4%	0.70	0.041	93.4%			
	α	0.20	0.20	0.020	94.0%	0.20	0.024	94.8%	0.20	0.034	95.8%			
	OR	0.60	0.61	0.11	94.8%	0.61	0.12	95.4%	0.61	0.12	95.2%			
Hybrid EM-MCEM ($J = 1000$)														
0.5	b	0.01	0.010	0.0015	94.8%	0.010	0.0015	94.4%	0.010	0.0016	95.6%			
	p	0.05	0.050	0.0098	94.6%	0.050	0.010	95.2%	0.050	0.011	96.0%			
	φ	0.70	0.70	0.041	93.4%	0.70	0.041	93.4%	0.70	0.041	93.4%			
	α	0.20	0.20	0.020	93.8%	0.20	0.024	95.2%	0.20	0.033	95.4%			
	OR	0.60	0.61	0.11	94.8%	0.61	0.12	95.4%	0.61	0.12	95.2%			
0	b	0.01	0.011	0.0015	91.4%	0.011	0.0016	92.2%	0.011	0.0016	92.6%			
	p	0.05	0.055	0.011	92.6%	0.055	0.011	93.2%	0.055	0.012	92.8%			
	φ	0.70	0.70	0.041	93.4%	0.70	0.041	93.4%	0.70	0.041	93.4%			
	α	0.20	0.20	0.020	94.0%	0.20	0.024	95.8%	0.20	0.033	95.6%			
	OR	0.60	0.61	0.11	95.0%	0.61	0.12	95.4%	0.61	0.12	95.0%			
1	b	0.01	0.0099	0.0015	95.2%	0.0099	0.0015	95.6%	0.010	0.0016	95.4%			
	p	0.05	0.044	0.0085	91.0%	0.044	0.0087	90.2%	0.044	0.0091	92.4%			
	φ	0.70	0.70	0.041	93.4%	0.30	0.041	93.4%	0.70	0.041	93.4%			
	α	0.20	0.20	0.020	94.0%	0.20	0.024	95.2%	0.20	0.033	95.4%			
	OR	0.60	0.61	0.11	95.0%	0.61	0.12	95.4%	0.61	0.12	95.2%			

Table 4

Analysis of the influenza H3N2 epidemic during 1977–1978 and the H1N1 epidemic during 1978–1979 in Seattle. Estimates and 95% CIs (in parentheses) are presented by the infectiousness level of asymptomatic cases relative to symptomatic cases (θ). The hybrid EM-MCEM algorithm with $K_0 = 200$, $K = 1000$, and $J = 500$ is used for inference. A tolerance rate of 10^{-5} for the relative error in every parameter is used to determine convergence during the EM iterations.

Parameter	1977–1978, H3N2			1978–1979, H1N1 [†]		
	$\theta = 0$	$\theta = 0.5$	$\theta = 1$	$\theta = 0$	$\theta = 0.5$	$\theta = 1$
$k \times 10^{-4}$	10.66 (7.58,15.0)	10.12 (7.16,14.29)	9.86 (6.97,13.94)	9.11 (7.35,11.30)	8.77 (7.02,10.96)	8.79 (7.02,10.99)
$p \times 10^{-2}$	6.49 (3.99,10.39)	6.47 (4.08,10.12)	5.78 (3.70,8.92)	7.48 (4.40,12.43)	6.28 (3.74,10.35)	4.74 (2.76,8.01)
ϕ		0.81 (0.65,0.91)			0.62 (0.52,0.70)	
SAR	0.27 (0.18,0.40)	0.27 (0.18,0.39)	0.25 (0.16,0.36)	0.31 (0.19,0.46)	0.27 (0.17,0.40)	0.21 (0.12,0.32)
OR _{ϕ} AGE	0.26 (0.19,0.38)	0.26 (0.18,0.37)	0.26 (0.18,0.37)	0.045 (0.017,0.11)	0.046 (0.018,0.12)	0.048 (0.019,0.12)
OR _{ϕ} TITER						
2 vs 1	1.20 (0.80,1.79)	1.21 (0.81,1.81)	1.23 (0.82,1.83)	0.77 (0.53,1.13)	0.76 (0.52,1.11)	0.76 (0.52,1.11)
3 vs 1	0.51 (0.32,0.82)	0.50 (0.32,0.81)	0.50 (0.31,0.80)	0.093 (0.012,0.70)	0.091 (0.012,0.68)	0.090 (0.012,0.68)
4 vs 1	0.22 (0.13,0.37)	0.22 (0.13,0.36)	0.22 (0.13,0.37)			
OR _{ϕ} AGE		0.55 (0.25,1.19)			0.41 (0.059,2.92)	
OR _{ϕ} TITER						
2 vs 1		0.87 (0.34,2.19)			2.25 (0.94,5.40)	
3 vs 1		0.31 (0.11,0.90)				
4 vs 1		0.39 (0.13,1.22)				

[†]Pre-season antibody titer levels 2 and 3 are combined for OR _{ϕ} TITER.