Fitting mechanistic epidemic models to data: a comparison of simple Markov chain Monte Carlo approaches

Michael Li, Jonathan Dushoff, Ben Bolker

January 31, 2017

Abstract

Background

Simple mechanistic epidemic models are widely used for forecasting and parameter estimation of infectious diseases based on noisy case reporting data. Despite the widespread application of models to emerging infectious diseases, we know little about the comparative performance of standard computational-statistical frameworks in these contexts. Here we build a simple stochastic, discrete-time, discrete-state epidemic model with both process and observation error and use it to characterize the effectiveness of different flavours of Bayesian Markov chain Monte Carlo (MCMC) techniques. We explore the limitations of different platforms and quantify parameter estimation accuracy, forecasting accuracy, and computational efficiency across combinations of modeling decisions (e.g. discrete vs. continuous latent states, levels of stochasticity) and computational platforms (JAGS, NIMBLE, Stan).

Results

Simulations showed that models incorporating at least one source of population-level variation (i.e., dispersion in either the transmission process or the observation pro-

cess) provide reasonably good forecasts and parameter estimates, while models that incorporate only individual-level variation can lead to inaccurate (or overconfident) results. Models using continuous-approximations to the transmission process showed improved computational efficiency without loss of accuracy.

Conclusion

Simple models of disease transmission and observation can be fitted reliably to simple simulations, as long as population-level variation is taken into account. Continuous approximations can improve computational efficiency using more advanced MCMC techniques.

Keywords: MCMC HMC TSIR Dispersion Moment-matching

1 Introduction

Simple homogeneous population models have been widely used to study emerging infectious disease outbreaks. Although such models can provide important insights — including estimated epidemic sizes and effects of intervention strategies, as well as short-term forecasts — they neglect spatial, individual-level and other heterogeneities which are often important. Decades of work have created frameworks that enable researchers to construct analytical models to capture many aspects of infectious disease epidemics. But many challenges remain. In particular, estimating parameters (and associated uncertainties) is always challenging, especially models incorporating multiple forms of heterogeneity, and especially during the early stages of an epidemic. Using complex models that are insufficiently supported by data can lead to unstable parameter estimates (Ludwig and Walters, 1985) — in many cases, researchers are forced to revert to simple models.

In the past few decades, researchers have begun to adopt Bayesian approaches

to disease modeling problems. Bayesian Markov Chain Monte Carlo (MCMC) is a powerful, widely used sampling-based estimation approach. Despite the widespread use of MCMC in epidemic modeling (Morton and Finkenstädt, 2005; O'Neill, 2002), however, there have been relatively few systematic studies of the comparative performance of statistical frameworks for disease modeling.

In this paper, we apply relatively simple MCMC approaches to data from simulated epidemics that incorporate stochasticity in both transmission and observation, and account for multiple generation infectious periods. We compare model approaches of varying complexity, including a fitting model that matches the simulation model, and we also explore three different MCMC platforms: JAGS (Plummer et al., 2003), NIMBLE (de Valpine et al., 2016) and Stan (Carpenter et al., 2016). We quantify and compare parameter estimation accuracy, forecasting accuracy, and computational efficiency across combinations of these modeling decisions.

2 Methods

We generated test data using a simple framework that combines a *transmission process* based on a simple discrete-time model with an *observation process* to account for incomplete reporting. Both processes are assumed to be stochastic. We then fit the observed cases from these simulations using Bayesian methods that model the underlying true number of infections as a latent (i.e., unobserved) variable. Our Bayesian fitting models explore an approach that matches the assumptions of the simulation model, and also various simplifications: in particular, we explore simpler methods of accounting for variation in both the transmission process and the observation process, and the use of continuous rather than discrete latent variables.

2.1 Simulation Model

The transmission process of our dual-process framework is based on the Reed-Frost (R-F) chain binomial model, which can also be described as a discrete-time, stochastic compartmental SIR model (Ludwig, 1973). To account for the possibility that some fraction of the population may be beyond the scope of the epidemic — geographically or socially isolated, genetically resistant, vaccinated or immune — we assume that only a proportion P_{eff} of the total population is effectively susceptible to infection. Then, for every time step, we assume that only a proportion P_{rep} of the number of new infections are actually observed. We model both transmission and observation using a beta-binomial (rather than binomial) distribution to account for additional sources of variation (i.e., overdispersion) in both processes. The equations are:

$$N_{\text{eff}} = P_{\text{eff}} N \tag{1}$$

$$S_1 = N_{\text{eff}} - I_1 \tag{2}$$

$$\Phi_t = \sum_{i=1}^{\ell} k(i) I_i \tag{3}$$

$$I_{t+1} \sim \text{BetaBin}(1 - e^{-\Phi_t}, S_t, \delta_P) \tag{4}$$

$$S_{t+1} = S_t - I_{t+1} \tag{5}$$

$$\text{Obs}_t \sim \text{BetaBin}(P_{\text{rep}}, I_t, \delta_{\text{obs}}). \tag{6}$$

where Φ_t is the force of infection at time t; N_{eff} is the effective population size; and ℓ is the number of lags.

We use the standard parameterization of the beta binomial, meaning that larger values of the dispersion parameters (δ_P and δ_{obs}) correspond to less variability (the beta-binomial converges to the binomial distribution as δ_{obs} becomes large).

We extend the R-F model by allowing the infectious period to last longer than one step, using a transmission kernel $k(i)$ based on a truncated negative binomial

distribution:

$$\tilde{k}(i) = i^{(G_S-1)} \times \exp\left(\frac{-i}{G_P \times \ell}\right), i = 1, ..., \ell \tag{7}$$

$$k(i) = \frac{\mathcal{R}_0}{N_{\text{eff}}} \times \frac{\tilde{k}(i)}{\sum_{i=1}^{\ell} \tilde{k}(i)}, i = 1, ..., \ell \tag{8}$$

Here, $\mathcal{R}_0$ represents the basic reproductive number and G_S and G_P are shape and position parameters, respectively.

2.2 Fitting Model

2.2.1 Transmission and Observational Process Errors

The transmission (eq. 4) and observation (eq. 6) processes in the simulation model are both defined as beta-binomial (BB) processes. In fitting, we used the BB to match the simulation model, but also tried several simpler alternatives: binomial (B), Poisson (P), and negative-binomial (NB) processes. Process B does not allow for overdispersion, while NB does not incorporate the size of the pool from which a value is chosen; that is, it is theoretically possible for a NB sample of the number of infections to be larger than the current susceptible population (although this is extremely unlikely when the *per capita* infection probability is small). Process P neglects both of these phenomena. Figure 1 illustrates the relationship of the four discrete distributions.

2.2.2 Latent Continuous Transmission process

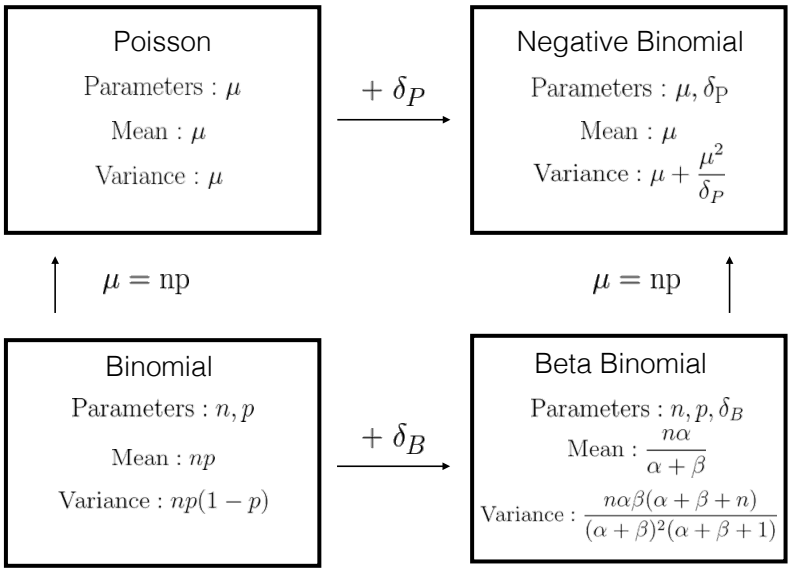

Figure 1: Discrete distribution relationships. For beta-binomial distribution (bottom right panel), we used an alternative parameterization α and β, where $\alpha = \frac{\delta_P}{1-p}$ and $\beta = \frac{\delta_P}{p}$. Moving from the top to bottom row adds a size parameter (replacing μ with np). Moving from left to right adds a dispersion parameter δ.

Continuous Approximation (Hybridization) $\mathrm{Gamma}(\mathrm{shape} = \alpha, \mathrm{rate} = \beta)$	
Poisson $\alpha = \mu$ $\beta = 1$	Negative Binomial $\alpha = \mu\beta$ $\beta = \frac{\delta_P}{\delta_P + \mu}$
Binomial $\alpha = np\beta$ $\beta = \frac{1}{(1-p)}$	Beta Binomial $\alpha = np\beta$ $\beta = \frac{\delta_B + p(1-p)}{(1-p)(\delta_B + np(1-p))}$

Figure 2: Continuous approximation of discrete distributions via moment matching. Distributions in Figure 1 were matched to a Gamma distribution with equivalent first and second moments.

Another simplification we considered was treating the unobserved number of underlying cases as a continuous variable. To do this, we matched the first two moments of the discrete distribution to a Gamma distribution (Figure 2). One advantage of the continuous approximation approach is that it allows us to scale our latent variable to help with model convergence (see below); it also allows the use of MCMC sampling procedures such as Hamiltonian Monte Carlo (HMC).

2.2.3 Multiple Scale Decorrelation

The proportion of the population assumed to be effectively susceptible (P_{eff}) and the reporting proportion (P_{rep}) have very similar effects on observed incidence. We therefore expect them to be hard to identify separately, so we reparameterized the model so that it uses a single parameter P_{effrep} for their product.

$$\widehat{P}_{\text{eff}} = P_{\text{effrep}}^{1-\rho} \tag{9}$$

$$\widehat{P}_{\text{rep}} = P_{\text{effrep}}^{\rho} \tag{10}$$

We also expect that this parameterization will improve statistical convergence, since it makes it possible to change the poorly constrained value of ρ without changing P_{effrep}. For similar reasons, we experimented with measuring infected individuals on a "reporting" scale in our continuous-variable models (see below).

2.3 Bayesian Markov Chain Monte Carlo

In Bayesian MCMC, model parameters are sampled from the posterior distribution by a reversible Markov chain whose stationary distribution is the target posterior distribution. Classical MCMC techniques include the Metropolis-Hasting algorithm (Hastings, 1970), Gibbs sampling (Geman and Geman, 1984), and slice sampling (Neal, 2003). Recently, convenient implementations of a powerful MCMC technique

called Hamiltonian Monte Carlo (HMC: also called hybrid MC) (Duane et al., 1987) have become available. HMC uses the concept of Hamiltonian dynamics to create a proposal distribution for the M-H algorithm, together with the leap-frog algorithm and the No U-Turn sampler (Hoffman and Gelman, 2014). HMC requires more computational effort per sample step compared to other MCMC techniques, but because subsequent steps are less correlated it also produces more effective samples per sample step (Hoffman and Gelman, 2014).

2.3.1 Platforms

Many software platforms implement the automatic construction of MCMC samplers for user-defined models. One of the most widely used platforms is JAGS (Just a Gibbs Sampler); despite its name, it combines a variety of MCMC techniques to fit models. NIMBLE (Numerical Inference for Statistical Models for Bayesian and Likelihood Estimation) is a more recent platform that allows users to flexibly model and customize different algorithms and sample techniques for MCMC. Neither JAGS nor NIMBLE has yet implemented HMC. One of the relatively few platforms that currently implements HMC is Stan, which provides full Bayesian inference for continuous-variable models based on the No-U-Turn sampler, an adaptive form of HMC.

2.3.2 Simulation and Evaluations

The typical (frequentist) statistical simulation scheme fits multiple realizations to data generated from a fixed set of parameters that is determined *a priori* and evaluates the match of the parameter estimates to the true values. Our simulation test scheme, based on a Bayesian perspective, sampled multiple sets of the parameters from the same prior distribution that was used in the fitting process and simulated one realization for each parameter set. All model variants were used to fit each realization (Table 1 and 2 in the appendix give more detail about parameters and priors).

Forecasts were made by simulating forward using parameters sampled from the fitted posterior distributions.

We used four summary statistics to evaluate total cases predicted over the forecast window (disaggregated forecasts are analyzed in the supplementary material), mean generation interval, and parameter estimates. The mean generation interval is defined by:

$$\text{Mean Generation Interval} = \frac{\sum_{i=1}^{\ell} i\hat{k}(i)}{\sum_{i=1}^{\ell} \hat{k}(i)} \tag{11}$$

We used bias, root mean square error (RMSE), and coverage to assess model fit. We also assessed model efficiency using time per effective sample. All errors used were proportional errors, calculated as:

$$\epsilon_i = \log\left(\frac{\text{med}(\hat{\theta}_i)}{\theta_i}\right) \tag{12}$$

We then calculated bias and RMSE as:

$$\text{Bias} = \text{median}(\epsilon) \tag{13}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{100}(\epsilon_i^2)}{100}} \tag{14}$$

3 Results

The full model (which matches the simulation model) provides generally good forecasts and parameter estimates when looking at either bias (Figure 3, or RMSE (Figure 4), except for estimates of P_{eff} using JAGS.

In general, models with any kind of dispersion in the transmission process, *or*

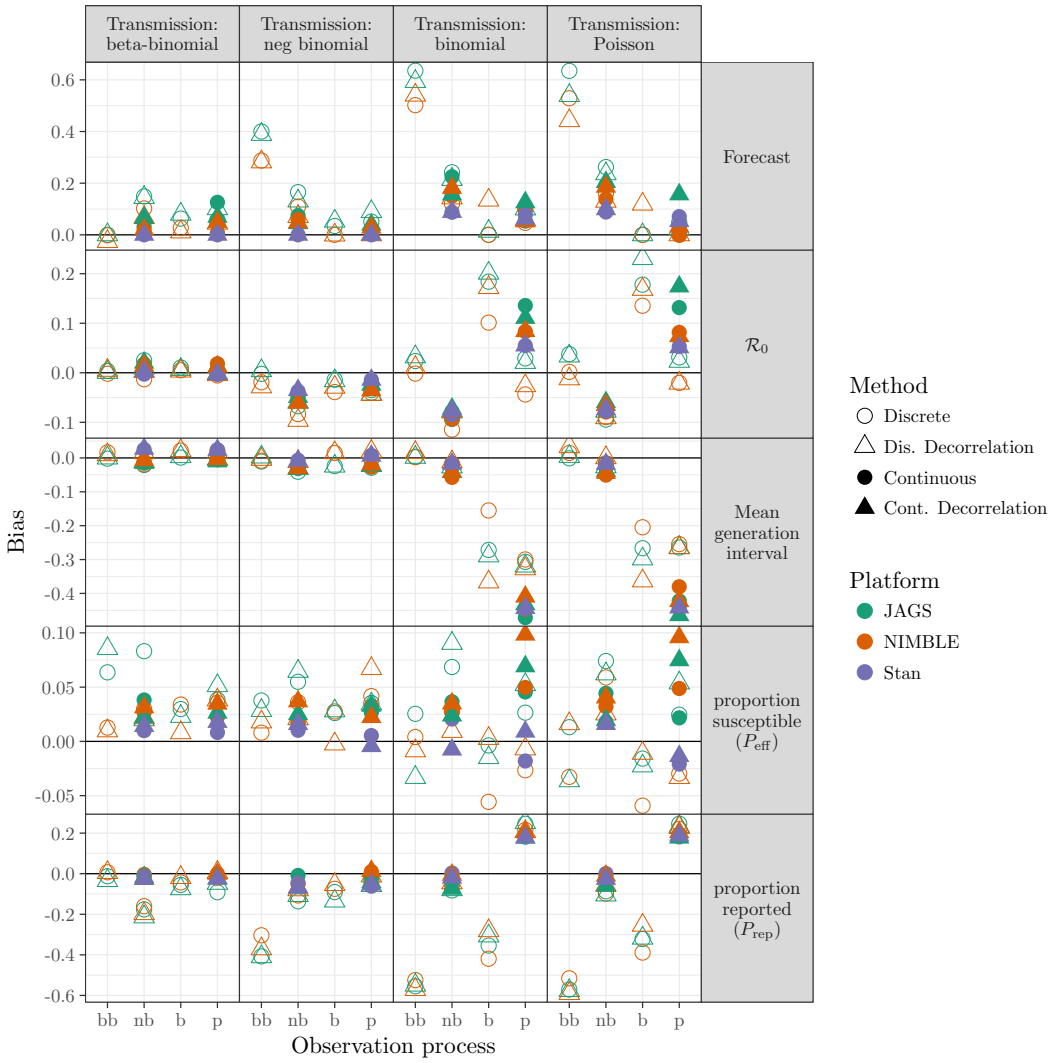

Figure 3: Comparison of bias (based on proportional errors) for forecast, MGI and parameters using models described in Sect. 2.2 across different platforms described in Sect. 2.3.1. Models with transmission process dispersion (first and second left column panels) and models with observation process dispersion (first and second left column within each panel) have low bias and models without dispersion have large bias (moving from left to right columns between and within panels. Continuous latent state models (solid points) and Stan (purple points) are available for negative binomial (second column within each panel) and poisson (fourth column within each panel) observational process. The black line shows bias of zero.

with negative binomial dispersion in the observation process, did well. The exception is that models that combined negative binomial transmission dispersal with beta binomial observation dispersal produced biased forecasts and estimates of P_{rep}.

There are no clear differences in the quality of model fit due to multi-scale decorrelation, latent continuous transmission process or platform.

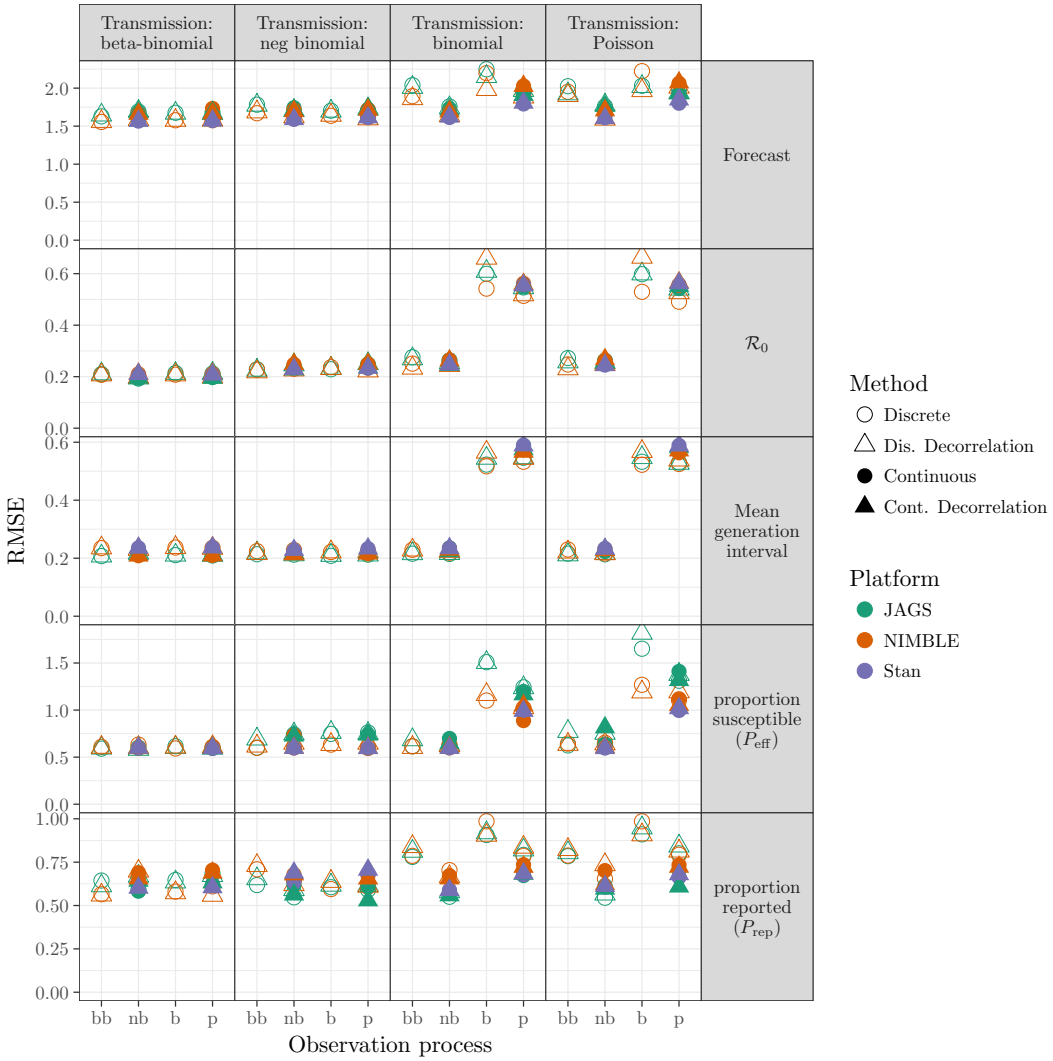

Figure 4: Comparison of RMSE (based on proportional errors) for all fitting model variants. The layout matches that of Figure 3. Patterns across models and platforms are similar to those seen in Figure 3. Short-term forecasts have generally high error, even when bias is low, reflecting inherent uncertainty in the system. The cross-correlated parameters P_{eff} and P_{effrep} also show high error but not high bias.

Figure 5 shows the statistical coverage of our estimates. Similar to the results shown for bias and RMSE (Figure Figure 3 and Figure 4), we find generally good

coverage (i.e., close to the nominal value of 0.9) for models with dispersion in the transmission process, except that the negative-binomial transmission process model undercovers across the board (coverage ≈ 0.8 for all observation process models and platforms) for forecasts and P_{rep}. For models without dispersion in transmission, models with dispersion in the observation process have low coverage (≈ 0.8) for most parameters, while the beta-binomial process model has low coverage (≈ 0.4) for P_{rep} and models without any dispersion have uniformly low coverage.

There are noticeable efficiency differences between platforms and transmission-process approaches (continuous vs. discrete), as measured by time per effective sample size, shown in Figure 6. For a given platform, models using continuous latent variables are generally more efficient than discrete latent processes. Comparing models with continuous latent variables between platforms (Figure 5, second and fourth column of every panel), Stan (using HMC) is the most efficient platform, followed by NIMBLE and JAGS. For discrete latent-state models, NIMBLE is more efficient than JAGS.

4 Discussion

This paper fits models with a variety of simplifications to simulated epidemic data with multiple sources of heterogeneity, using several different platforms. Using models that include some form of overdispersion is necessary for robust fits, but models that include overdispersion only in the transmission process can work as well as or better than the full model. Including overdispersion only in the observation process (if implemented as a negative binomial distribution) also provides relatively robust fits to these data. Simplifying the models by using continuous rather than discrete latent variables increased efficiency with little effect on fits.

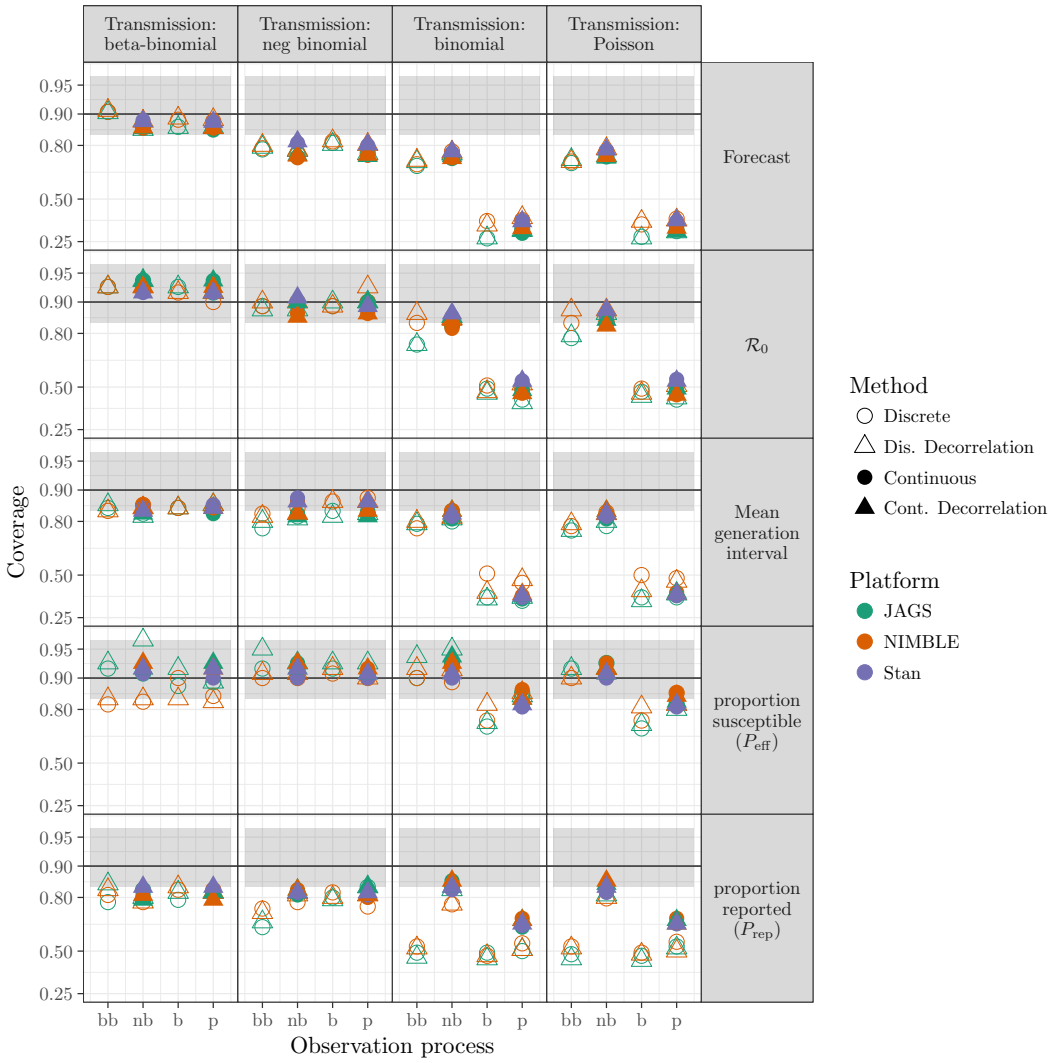

Figure 5: Comparison of coverage probability for forecast and parameters. Models with transmission process dispersion (left column panels) and models with observation process dispersion (first and second left column within each panel) have coverage near the nominal value of 0.9 for all parameters and model variants. The black line shows the nominal coverage, and the grey ribbon the 95% binomial confidence interval based on the number of simulated fits. Vertical axis is plotted on a logit scale.

4.1 Ceilings

The effects of using distributions with ceilings (i.e. binomial and beta binomial distributions) instead of their less realistic counterparts without ceilings (Poisson and negative binomial) was relatively small. In our framework, ceilings only apply in models with discrete latent variables; the primary effect of such ceilings is to reduce

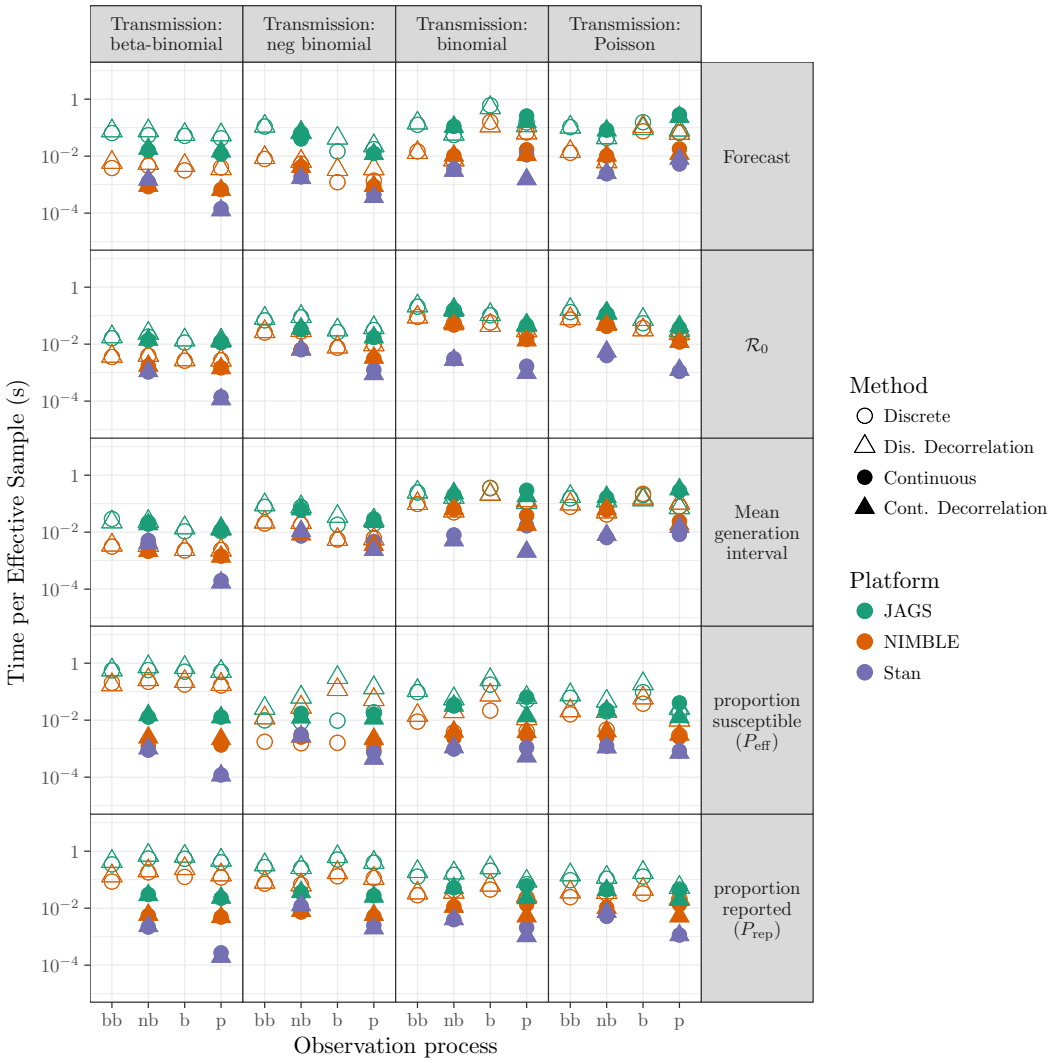

Figure 6: Comparison of efficiency for all fitting model variants (the layout of the figure is also the same as Figure 3)

variance as probabilities (of infection or of sampling) become large.

4.2 Overdispersion

Accounting for overdispersion had more impact on our fits than the presence or absence of ceilings. In particular, models with no overdispersion in either process lacked flexibility and tended to be over-confident (that is, they showed low coverage). However, models that account for overdispersion in only one process (either transmission

or observation) tended to be reliable for estimating parameters such as $\mathcal{R}_0$, mean generation interval, and short-term forecasts, particularly when overdispersion was implemented through negative binomial (a less constrained distribution than the beta binomial). However, parameters such that are more closely tied to the details of a particular model structure used (such as the overdispersion parameters for the observation and transmission processes) must change when the overdispersion model changes, in order to compensate for missing sources of variability.

Several authors (e.g., (King et al., 2015; Taylor et al., 2016)) recommend accounting for process as well as observation error in estimates of $\mathcal{R}_0$ and in forecasts, to avoid over-confident estimates. Our exploration does not include any cases where process error is completely absent — even our "dispersion-free" processes incorporate sampling error in the process. However, we find that neglecting overdispersion can still lead to over-confident and unreliable estimates.

4.3 Latent vs Observable

We are interested in two aspects of the epidemic that are not directly observable: reporting rate and total effective population size. Classic infectious disease models ignore both of these aspects, relying on the constancy of reporting rate and the non-sensitivity of e.g. $\mathcal{R}_0$ estimates to a constant degree of underreporting (Clarkson and Fine, 1985). While we want to use as much observable information as possible and make as few assumptions as possible about unobservable aspects of the epidemic, underreporting is of huge practical importance. Thus, modeling observation error explicitly is required if we want reliable estimates of uncertainty (King et al., 2015). If reporting error is modeled with a ceiling, then underreporting is a necessary component of reporting error (i.e., reporting is always biased downward as well as noisy). Allowing overdispersion, especially without a ceiling (i.e, a negative-binomial model of the reporting process), decouples variance and bias in the reporting process.

We have shown how simple techniques can improve accuracy and efficiency in modeling epidemics, but much remains to be done. Our fitting model neglects many different forms of heterogeneity and epidemic phenomena — among them spatial, age and social structure — that may be important in modeling epidemics. We have yet to explore more advanced Bayesian MCMC techniques that can potentially improve accuracy, such as redundant parameterizations, block sampling, or sequential Monte Carlo frameworks (Del Moral et al., 2012; Gelman et al., 2014; He et al., 2009; Yang et al., 2014).

5 Conclusion

We have presented a comparison of simple MCMC approaches to fit epidemic data. We learned two things about fitting epidemic data. First, modeling different processes with dispersion (BB and NB) is a naive but effective way to add uncertainty in the model; models without such uncertainty are likely to be over-confident and less accurate at forecasting. Second, approximating discrete latent state process with continuous processes can aid efficiency without losing robustness of fit. This allows more efficient fitting in the classic framework (e.g., JAGS and NIMBLE), and also allows us to use the more advanced HMC technique (which we implemented via Stan).

6 Acknowledgments

We would like to thank Ebola challenge organizers for organizing the Ebola model challenge. We also would like to thank Fred Adler for his thoughtful comments. Lastly, we thank McMaster University, NSERC Discovery grant and CIHR Ebola grant for funding this project.

References

Carpenter, B., A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, and A. Riddell (2016). Stan: A probabilistic programming language. *J Stat Softw*.

Clarkson, J. A. and P. E. M. Fine (1985). The efficiency of measles and pertussis notification in England and Wales. *International Journal of Epidemiology 14*(1), 153–168.

de Valpine, P., D. Turek, C. J. Paciorek, C. Anderson-Bergman, D. T. Lang, and R. Bodik (2016). Programming with models: writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics* (just-accepted), 1–28.

Del Moral, P., A. Doucet, and A. Jasra (2012). An adaptive sequential monte carlo method for approximate bayesian computation. *Statistics and Computing 22*(5), 1009–1020.

Duane, S., A. D. Kennedy, B. J. Pendleton, and D. Roweth (1987). Hybrid monte carlo. *Physics letters B 195*(2), 216–222.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2014). *Bayesian data analysis*, Volume 2. Chapman & Hall/CRC Boca Raton, FL, USA.

Geman, S. and D. Geman (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence* (6), 721–741.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika 57*(1), 97–109.

He, D., E. L. Ionides, and A. A. King (2009). Plug-and-play inference for disease dynamics: measles in large and small populations as a case study. *Journal of the Royal Society Interface*.

Hoffman, M. D. and A. Gelman (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research 15*(1), 1593–1623.

King, A. A., M. D. de Cellès, F. M. G. Magpantay, and P. Rohani (2015, May). Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola. *Proc. R. Soc. B 282*(1806), 20150347.

Ludwig, D. (1973). Mathematical models for the spread of epidemics. *Computers in biology and medicine 3*(2), 137–139.

Ludwig, D. and C. J. Walters (1985, June). Are Age-Structured Models Appropriate for Catch-Effort Data? *Canadian Journal of Fisheries and Aquatic Sciences 42*(6), 1066–1072.

Morton, A. and B. F. Finkenstädt (2005). Discrete time modelling of disease incidence time series by using markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 54*(3), 575–594.

Neal, R. M. (2003). Slice sampling. *Annals of statistics*, 705–741.

O'Neill, P. D. (2002). A tutorial introduction to bayesian inference for stochastic epidemic models using markov chain monte carlo methods. *Mathematical biosciences 180*(1), 103–114.

Plummer, M. et al. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, Volume 124, pp. 125. Vienna.

Taylor, B. P., J. Dushoff, and J. S. Weitz (2016, November). Stochasticity and the limits to confidence when estimating of Ebola and other emerging infectious diseases. *Journal of Theoretical Biology 408*, 145–154.

Yang, W., A. Karspeck, and J. Shaman (2014). Comparison of filtering methods for the modeling and retrospective forecasting of influenza epidemics. *PLoS Comput Biol 10*(4), e1003583.