Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/MBE

MBE, 16(5): 4299–4313.
DOI: 10.3934/mbe.2019214
Received: 25 January 2019
Accepted: 21 April 2019
Published: 16 May 2019

Research article

Comparative assessment of parameter estimation methods in the presence of overdispersion: a simulation study

Kimberlyn Roosa[1,*], **Ruiyan Luo**[1] **and Gerardo Chowell**[1,2]

[1] Department of Population Health Sciences, School of Public Health, Georgia State University, Atlanta, GA, USA

[2] Division of International Epidemiology and Population Studies, Fogarty International Center, National Institute of Health, Bethesda, MD, USA

* **Correspondence:** Email: kroosa1@student.gsu.edu.

Abstract: The Poisson distribution is commonly assumed as the error structure for count data; however, empirical data may exhibit greater variability than expected based on a given statistical model. Greater variability could point to model misspecification, such as missing crucial information about the epidemiology of the disease or changes in population behavior. When the mechanism producing the apparent overdispersion is unknown, it is typically assumed that the variance in the data exceeds the mean (by some scaling factor). Thus, a probability distribution that allows for overdispersion (negative binomial, for example) may better represent the data. Here, we utilize simulation studies to assess how misspecifying the error structure affects parameter estimation results, specifically bias and uncertainty, as a function of the level of random noise in the data. We compare results for two parameter estimation methods: nonlinear least squares and maximum likelihood estimation with Poisson error structure. We analyze two phenomenological models the generalized growth model and generalized logistic growth model to assess how results of parameter estimation are affected by the level of overdispersion underlying in the data. We use simulation to obtain confidence intervals and mean squared error of parameter estimates. We also analyze the impact of the amount of data, or ascending phase length, on the results of the generalized growth model for increasing levels of overdispersion. The results show a clear pattern of increasing uncertainty, or confidence interval width, as the overdispersion in the data increases. While maximum likelihood estimation consistently yields narrower confidence intervals and smaller mean squared error, differences between the two methods were minimal and not practically significant. At moderate levels of overdispersion, both estimation methods yielded similar performance. Importantly, it is shown that issues of parameter uncertainty and bias in the presence of overdispersion can be mitigated with the inclusion of more data.

Keywords: parameter uncertainty; overdispersion; phenomenological models; generalized growth model; epidemiological models; parameter estimation

1. Introduction

Mathematical modeling offers a quantitative framework for investigating dynamics of infectious disease epidemics and guiding decisions regarding the type and intensity of public health control interventions. Phenomenological models are often useful for forecasting epidemic trajectories, whereas, mechanistic models allow researchers to evaluate the effects of interventions or the roles of different factors on transmission dynamics (e.g., mixing patterns or environmental factors). To be useful in specific outbreak contexts, dynamic models are often calibrated using infectious disease outbreak data that typically correspond to time series of new cases, where a case corresponds to an observable (reportable) event. It should be noted that the time-series data curve corresponds to only one realization of a stochastic process, and unfortunately, generating more data realizations in a carefully controlled environment is not feasible in the context of real outbreaks occurring in natural environments.

When calibrating models to data via some fitting process (also known as data assimilation), the solution of the dynamic model for a given set of parameter values and initial conditions is typically considered to be the mean solution, which is embedded into a counting process characterized by a statistical model (e.g. Poisson, Negative Binomial). For instance, a researcher fits an SEIR-type model (mean signal) to weekly series of newly reported cases of Ebola in West Africa assuming a Poisson error structure around the case series data. In this inference framework, the *equidispersion* property of the Poisson distribution (where the mean is equal to the variance) simplifies the inference process, limits the number of degrees of freedom, and indirectly reduces potential issues of parameter non-identifiability [1]. Moreover, in real-time analyses of evolving outbreaks, one has to link the observable with the unobservable by adjusting, for instance, for delays associated with incubation periods, time to diagnosis, or reporting delays [2].

Importantly, visual inspection of time series data could suggest an apparently larger variability than the mean signal linked to the model. One potential source of this *apparent overdispersion* effect could arise from systematic deviations of the model (mean) to the data due to model misspecification (e.g., the model incorrectly specifies the length of the incubation period or neglects another important mechanism involved in the dynamic process) [3]. Hence, researchers could fix this lack of model fit by identifying and incorporating other key process components in the model, thus resolving the apparent overdispersion issue. Alternatively, there is *actual overdispersion*, where the variability in the data is larger than expected. For example, count data in the early phase of an infectious disease outbreak would follow a Poisson distribution with completely homogenous mixing and a constant infectious period across individuals; however, these are unrealistic assumptions for real outbreak data, as the counts often deviate from a Poisson distribution in the presence of individual heterogeneity and randomly distributed infectious periods [2]. When actual overdispersion is detected, the researcher may reconsider the statistical model by considering those that allow for the variance to be larger than the mean (e.g., Negative Binomial) [1,2]. Hence, identifying the relevant sources of overdispersion is critical in the modeling process, as it could lead to poor description of the data and predictive power and underestimated standard errors and confidence intervals [2,4].

Fortunately, simulation studies can be utilized to evaluate the impact of various forms of misspecification when calibrating a model to data. In a previous paper, we outline a simple computational bootstrap-based method for assessing parameter identifiability [5]. This method involves repeated sampling from the deterministic model solution to simulate multiple data sets from which the parameters are re-estimated [5–7]. This allows us to detect parameter non-identifiability that could arise from model structure or the amount of information that can be extracted from the available data. In this paper, we evaluate the effects of misspecification of the error structure on bias and uncertainty associated with parameter estimates using simple dynamic transmission models. Specifically, we focus on modeling varying levels of data overdispersion stemming from randomness in the counting process that shapes the time series data, rather than systematic misspecifications linked to the dynamic model. We utilize Monte Carlo simulation, an approach similar to the parametric bootstrapping approach, to assess parameter estimates and their uncertainty as a function of the level of random noise in the data, and we compare results using two common parameter estimation methods: nonlinear least squares (LSQ) and maximum likelihood estimation with a Poisson error structure (Poisson-MLE).

2. Methods

2.1. Phenomenological models

For each of the following model examples, daily time series incidence (total number of new cases) curves were simulated directly from the model equation. All simulations and analyses were performed in Matlab 2017 (Mathworks, Inc).

2.1.1. Example 1: Generalized growth model (GGM)

Models used to study the growth patterns of infectious disease outbreaks often assume exponential growth in the absence of control interventions (compartmental models, for example) [8, 9]; however, growth patterns are likely slower than exponential for some diseases depending on the mode of transmission and the population structure. For example, Ebola spreads only via close contact, so in a constrained population contact structure, sub-exponential growth patterns would be expected [10]. The generalized growth model (GGM) includes a deceleration of growth parameter, also referred to as a scaling of growth parameter, p (range: [0, 1]) that relaxes the assumption of exponential growth [11]. A value of $p = 0$ represents constant (linear) growth, while a value of $p = 1$ indicates exponential growth. If $0 < p < 1$, the growth pattern is characterized as sub-exponential or polynomial.

The GGM is as follows:

$$\frac{dC(t)}{dt} = C'(t) = rC(t)^p$$

where $C(t)$ describes the cumulative number of cases at time t, $C(t)$, the derivative of $C(t)$, is the incidence curve, r is the growth rate parameter ($r > 0$), and p is the deceleration of growth parameter [11].

Allowing for a range of growth scaling in the model allows for applications to outbreak data of various different diseases. For example, the GGM has been applied to forecast outbreaks of a range of diseases, including foot and mouth disease [12], Zika [13], pandemic influenza [14], HIV/AIDS [15], and Ebola [16, 17]. For the example presented here, we assume a growth rate $r = 0.4$ and a deceleration of growth rate $p = 0.9$ (Table 1) for the simulated data. Similar values have been used to characterize pandemic influenza outbreaks.

Table 1. Descriptions and values of parameters used in simulation for the generalized growth model (Example 1).

Parameter	Description	True value (in simulations)	Estimation limits
r	Rate of change (growth rate)	0.4	[0, 10]
p	Deceleration constant ($0 \leq p \leq 1$)	0.9	[0, 1]

2.1.2. Example 2: Generalized logistic growth model (GLM)

While the GGM can model early epidemic growth, the function is strictly increasing, and thus cannot be used to fit entire epidemic curves (as it is assumed epidemic growth will slow at some point in time). The generalized logistic growth model (GLM) is an extension of the GGM that includes a parameter K that classifies the carrying capacity or final size of the epidemic. The GLM is as follows:

$$C'(t) = rC(t)^p(1 - \frac{C(t)}{K})$$

where $C(t)$ describes the cumulative number of incident cases at time t, and $C(t)$ is the incidence curve [11]. Again, r is the intrinsic growth rate, p is the deceleration of growth parameter, and K is the final epidemic size. For the GLM, we use the same r and p values from Example 1 and set the final epidemic size $K = 10,000$ (Table 2).

Table 2. Descriptions and values of parameters used in simulation for the generalized logistic growth model (Example 2).

Parameter	Description	True value	Estimation limits
r	Rate of change (growth rate)	0.4	[0, 10]
p	Deceleration constant ($0 \leq p \leq 1$)	0.9	[0, 1]
K	Final epidemic size/carrying capacity	10000	[0, 1000000]

2.2. Data error structure

The Poisson distribution is a commonly assumed error structure for count data which has equal mean and variance [1,4]. However, empirical data may exhibit greater variability than expected based on a Poisson distribution, which implies a systematic model misspecification or missing crucial information about the disease [4]. If the mechanism producing the overdispersion is known, it could be remedied by revising the model; however, when it is unknown, it is typically assumed that the variance in the data exceeds the mean (by some scaling factor) [1,7]. Relative to the Poisson distribution, the negative binomial distribution requires an additional parameter to model count data with varying levels of overdispersion. In the case of equidispersion (variance equal to the mean), the Poisson distribution is a special case of the negative binomial distribution.

For the simple phenomenological models employed here, the data are realizations of random counts $y_t = y_1, y_2, \ldots, y_n$, following the defined distribution, where t represents time point ($t = 1, 2, , n$) and n is the total number of observations [7]. We model the error using the negative binomial distribution, with mean $E(Y) = \mu$ and variance $Var(Y) = \sigma^2$. Let $d = \sigma^2/\mu$ represent the variance-to-mean ratio [1,2]. Thus $d = 1$ yields the Poisson distribution, and $d > 1$ indicates presence of overdispersion, for which the negative binomial distribution is a common choice. It should be noted that $d < 1$ represents underdispersion, though this is less common in empirical data [3]. Here, we consider data simulated with the following values: $d = [1, 2, 20, 40, 60, 80, 100]$, which represent different levels of overdispersion.

2.3. Simulated data

We utilize a Monte Carlo simulation to quantify the uncertainty of parameter estimates. This approach is an iterative process that involves simulating random error around the data in each iteration to generate a sample of data sets from which to estimate parameters [18]. The estimated parameters from all iterations simulate the sampling distributions from which we construct the confidence intervals of parameters. This method is similar to the parametric bootstrapping approach (derived from the general bootstrap method [6]), which first fits the model of interest to the available time-series data to obtain the best-fit estimate of the parameters; however, here we are simulating the data directly from the model, so the 'best-fit estimate' to the simulated data is essentially the model with the given parameter values (Θ_{true}). This way we know the true parameter values of the data and can assess the performance of different parameter estimation approaches considering different error structures.

We generate time series data directly from the model equation, setting parameters to values of interest (Tables 1 and 2); this yields the data y_t with solution $f(t, \Theta_{true})$, which is used to generate $M = 500$ simulated data sets for each variance-to-mean ratio. Assuming Poisson error structure, each new observation is sampled from a Poisson distribution with mean = $f(t, \Theta_{true})$, which is the incidence curve, at each time point t. This results in 500 estimated parameter sets $\hat{\Theta}_i$, where $i = 1, 2, , M$.

To analyze scenarios of data overdispersion, or of variance greater than Poisson mean, we utilize the negative binomial distribution with variance-to-mean ratios (d) of 2, 20, 40, 60, 80, and 100. To simulate negative binomial noise, the variance at each time point t is given by multiplying $f(t, \Theta_{true})$, the mean, by the specified variance-to-mean ratio. The above steps (1-4) are repeated for each value of d, assuming a negative binomial error structure, in place of Poisson. Thus, we obtain empirical distributions for each estimated parameter at each of the seven variance-to-mean ratios, where $d = 1$ indicates Poisson noise.

2.4. Parameter estimation

For each example, we estimate parameters for each time-series curve using the two general estimation methods. For both methods, one can use numerical optimization methods available in Matlab or R (R Core Team). The methods are as follows:

2.4.1. Nonlinear least squares (LSQ)

Least squares estimation yields the best fit solution to the time series data by searching for the parameter set $\hat{\Theta} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m)$ that minimizes the sum of squared deviations between the data y_t and $f(t, \Theta)$ [7]. That is,

$$\hat{\Theta} = argmin \sum_{t=1}^{n} (f(t, \Theta) - y_t)^2$$

For the presented examples, $f(t, \Theta) = C'(t|\Theta)$ is the function of the incidence rate at time t that depends on the set of parameters Θ. For the GGM (Example 1), $\Theta = (r, p)$. For the GLM (Example 2), $\Theta = (r, p, K)$. Then, $\hat{\Theta}$ is the parameter set that yields the smallest differences between the data and model. To run nonlinear least squares estimation (LSQ), we utilize the *fmincon* function in Matlab 2017, which finds the minimum of a constrained nonlinear multivariable function, with the interior-point algorithm (default). Further, we restrict the bounds for the parameters (Tables 1 and 2).

This parameter estimation method gives the same weight to all of the data points. LSQ also does not require a specific distributional assumption for y_t, except for the first moment $E[y_t] = f(t, \Theta)$; meaning that the mean at time t is equal to the count (e.g., number of cases) at time t [19]. It is of interest to study the impact of data overdispersion on LSQ parameter estimates.

2.4.2. Poisson-MLE

The goal of maximum likelihood estimation (MLE) is to derive parameter estimates given a model that dictates the dynamical process and data with variability assumed to follow a specific probability distribution. Consider the probability density function (PDF) that specifies the probability of observing data y_t given the parameter set Θ, or $p(y_t|\Theta)$ [19]. MLE aims to determine the values of the parameter set that maximizes the likelihood function $L(\Theta|y_t) = \prod_{t=1}^{n} p(y_t|\Theta)$ [2, 19, 20]. The resulting parameter set is called the MLE estimate, the most likely to have generated the observed data. Specifically, the MLE estimate is obtained by maximizing the corresponding log-likelihood function. For count data (y_t) with variability characterized by the Poisson distribution, the log-likelihood function is given by:

$$L(\Theta|y_t) = \sum_{t=1}^{n} [y_t log(f(t, \Theta)) - f(t, \Theta)]$$

and the Poisson-MLE estimate is expressed as

$$\hat{\Theta} = argmax \sum_{t=1}^{n} [y_t log(f(t, \Theta)) - f(t, \Theta)]$$

We again utilize the *fmincon* function with the same parameter bounds as defined for LSQ (Tables 1 and 2). It is of interest to compare the performance of Poisson-MLE and LSQ with simulation in the context of increasing levels of variability in the data.

We utilize Poisson-MLE and LSQ to estimate the parameters for each simulated data set, resulting in 500 estimated parameter sets $\hat{\Theta}_i$ where $i = 1, 2, , M$ (per level of overdispersion). We repeat the steps for each specified level of overdispersion, with variance-to-mean ratios given by $d = 1$ (Poisson), 2, 20, 40, 60, 80, and 100.

2.5. Performance

To compare the parameter estimation methods within each model example, we assess the empirical distributions of the estimates obtained from the Monte Carlo simulation. For each method, we calculate the 95% confidence intervals (CIs) for each parameter using the 2.5 and 97.5 percentiles. The width of the confidence intervals is used to compare the uncertainty, or precision, of parameter estimates at each level of overdispersion. As LSQ had wider confidence intervals in the majority of the simulations (with only one exception), the relative width difference between two confidence intervals is calculated as: $\%diff = \dfrac{width_{LSQ} - width_{MLE}}{width_{LSQ}} \cdot 100\%$.

Further, we use the mean squared error (MSE) to quantify accuracy, or how close the estimated values are from the true parameter value across the entire distribution of parameter estimates. MSE is calculated as: $MSE = \dfrac{1}{M}\sum_{i=1}^{M}(\theta_{true} - \hat{\theta}_i)^2$, where θ_{true} represents the true parameter value (in the simulated data) and $\hat{\theta}_i$ represents the estimated parameter value for the i^{th} bootstrap sample.

3. Results

3.1. Example 1: GGM

For the GGM, we simultaneously estimate both model parameters, r and p, and compare results for the two estimation methods: nonlinear least squares (LSQ) and maximum likelihood estimation (MLE). We use an ascending phase length (amount of data fit to) of 45 days, and we will later assess how the amount of data used impacts the results. Based on both LSQ and Poisson-MLE, the level of overdispersion in the data had little effect on the mean estimated parameter values. This can be seen in Figure 1, as the mean estimates at each level are distributed randomly around and very closely to the true value line. For both methods (LSQ and Poisson-MLE), the amount of uncertainty surrounding the parameter estimates increases as the level of overdispersion increases 95% confidence intervals become increasingly wider for higher variance-to-mean ratios (Figure 1). The percent differences of CI widths (for each d) are nearly equivalent for both r and p, so to avoid repetition, we will discuss percent difference in terms of r. Results indicate that even small levels of overdispersion can impact the uncertainty; for MLE, there is a 43% increase in CI width from $d = 1$ (r = 0.4 (0.392, 0.408)) to $d = 2$ (r = 0.4 (0.388, 0.411)), and for LSQ a 38% increase from $d = 1$ (r = 0.4 (0.389, 0.412)) to $d = 2$ (r = 0.4 (0.383, 0.415)) is observed. The largest increase in CI width is seen between $d = 2$ (see CIs in previous sentence) to $d = 20$ (MLE: r = 0.399 (0.368, 0.436); LSQ: r = 0.402 (0.348, 0.458)), with a 201.33% increase for MLE and a 246.54% increase for LSQ. After $d = 20$, the variance-to-mean ratio increases by 20. For each of these increases in variance-to-mean ratio, the % increase in CI width ranges from 13% − −47% for MLE and 11% − −37% for LSQ.

Comparing the methods to each other, we can see that MLE consistently yields narrower confidence intervals, or less uncertainty, compared to LSQ. Across the levels of overdispersion, the relative difference between LSQ and MLE confidence interval widths ranges from 28-38.5% for r and 29.5-38.5% for p, showing that while both methods' CIs are increasing, the relative difference between them is remaining stable. While these relative differences may seem high, actual CI width differences between the methods are small. With Poisson data, a difference in CI width of 0.0073 is observed for r and 0.0031 for p, comparing LSQ to MLE. For a large amount of overdispersion ($d = 100$), the CI difference is 0.0708 (29.27%) for r and 0.032 (31.19%) for p. In regards to application of the GGM specifically, these differences in widths do not yield practically meaningful differences in parameters estimated.

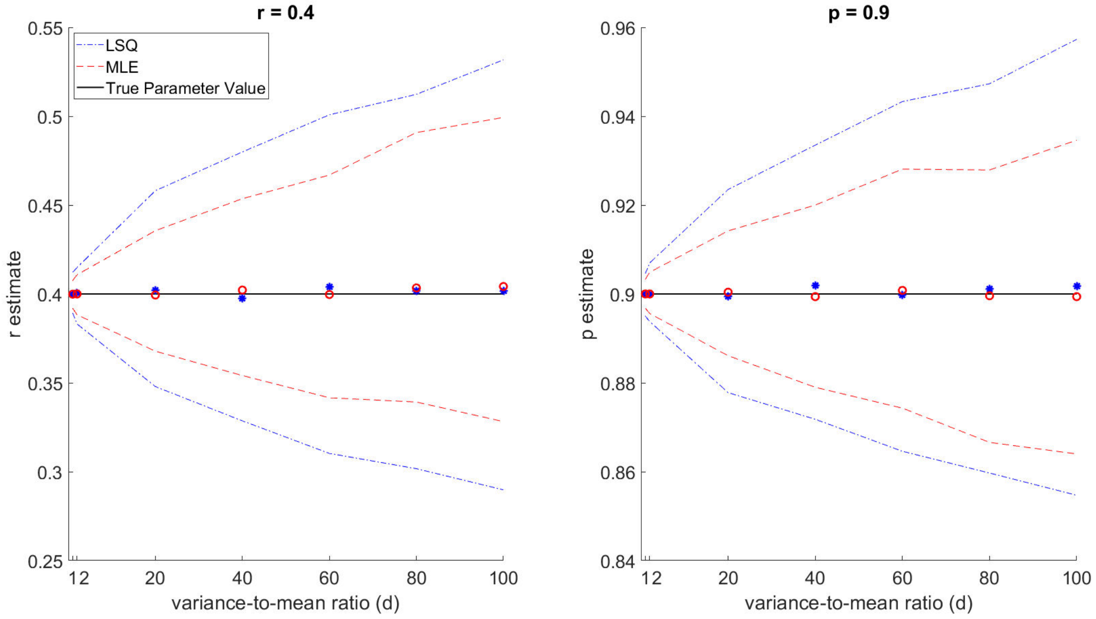
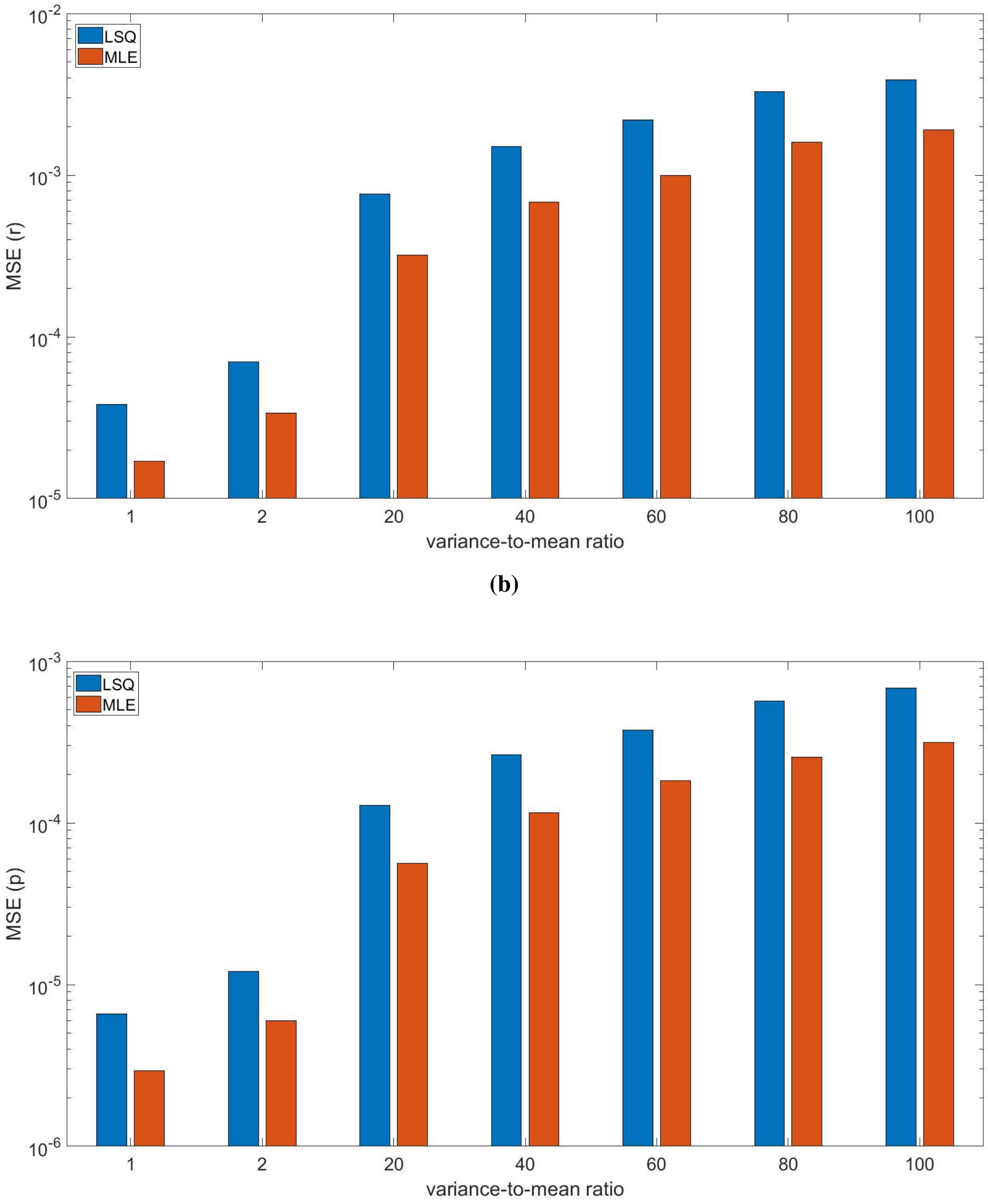


Figure 1. GGM parameter estimation results for increasing levels of variance assumed in the data. Mean estimates (circles) and 95% confidence intervals (dashed lines) are shown for the two estimation methods: nonlinear least squares (LSQ) and maximum likelihood estimation (MLE).

In Figure 2, we see that the MSE increases for higher levels of noise in the data. At each level of overdispersion, MLE and LSQ have very small diferences in MSE, but MLE consistently has lower MSE. The MSE for LSQ ranges from 2.05–2.38 times the MSE for MLE for r and 2.01-2.30 times for p. Again, while MLE is relatively more accurate than LSQ, the practical differences are small. For example, the largest difference in MSE seen between the two methods was less than 0.002.

It is known that parameter estimation results depend on the amount of information in the data available to fit the model to. For this purpose, we also performed the analyses for four different variance-to-mean ratios ($d = 1, 10, 50, 100$) at increasing lengths of the ascending phase, ranging from 15 to 40 days (increments of 5 days). Across each level of overdispersion, the overall pattern was consistent: fitting to more data (longer ascending phase length) resulted in smaller confidence intervals, and thus, lower uncertainty of parameter estimates. While the widths of the confidence intervals vary significantly across the levels of overdispersion, this general pattern is clearly seen for both estimation methods (Figure 3). For the Poisson error structure ($d = 1$), Poisson-MLE should yield the true uncertainty (confidence intervals) of estimates, as the method is based on the correct error structure; whereas, LSQ assumes constant variance. It is shown that, for the Poisson case, LSQ yields wider confidence intervals compared to MLE (CI width differences - LSQ - MLE - for r: 0.040, 0.060, 0.038, 0.018, 0.014, 0.007 for ascending phase length, t = 15, 20, 25, 30, 35, and 40 days respectively), and the difference between the two methods decreases as the ascending phase increases, excluding from 15 to 20 days (Figure 3). The % difference in width between the two methods ranges from 12.43–31.07% (for all t), but as previously stated, the practical significance of these differences is small.

(a)



(b)

Figure 2. Mean squared error of the distribution of GGM parameter estimates (M=500) for increasing levels of error assumed (variance-to-mean ratio) is shown for both estimation methods: nonlinear least squares (LSQ) and maximum likelihood estimation (MLE). Results for r are shown in (a), and results for p are in (b). Note that MSE is in log-scale.

Figure 3 also shows that higher levels of overdispersion require longer ascending phase lengths to reduce parameter uncertainty. For example, under Poisson assumptions, MLE yields a confidence interval length of about 0.15 for estimates of r using an ascending length of 20 days. When overdispersion is present, longer ascending phases are needed for MLE to yield comparable confidence intervals. Comparing to the CI width of 0.15 for r for MLE with 20 days of data, a variance-to-mean ratio of 10 yields similar uncertainty (MLE CI width: 0.17) with 30 days of data; and further, a variance-to-mean ratio of 50 yields similar uncertainty (MLE CI width: 0.17) with 40 days of data. This indicates that uncertainty of parameter estimates in the presence of overdispersion can be mitigated with the inclusion of more data points. This idea is seen in Figure 3, in that the confidence bounds quickly converge to the true value line as more data points are used, even for extreme cases of overdispersion (e.g., variance-to-mean ratio = 100).

Similarly, at each level of overdispersion, the MSE decreases for longer ascending phase lengths (Figure 4), indicating that more data yields higher accuracy of parameter estimates. For example, for data with Poisson error structure, each 5 day increase in ascending phase yielded between 55% − 72% decrease, in descending order, for r estimated with MLE. Poisson-MLE results for p and LSQ results (r and p) are nearly equivalent and follow the same patterns. It can be seen that, for increasing variance-to-mean ratios, more data points are required before the mean estimate falls on the true value line (Figure 3). For example, looking at the plots of r estimates, the mean value falls on the line around 20 days, 25 days, 30 days, and 35 days for the increasing levels of overdispersion (σ^2 = 1, 10, 50, 100), indicating that these would be the minimum amount of data from which the signal can be accurately detected.

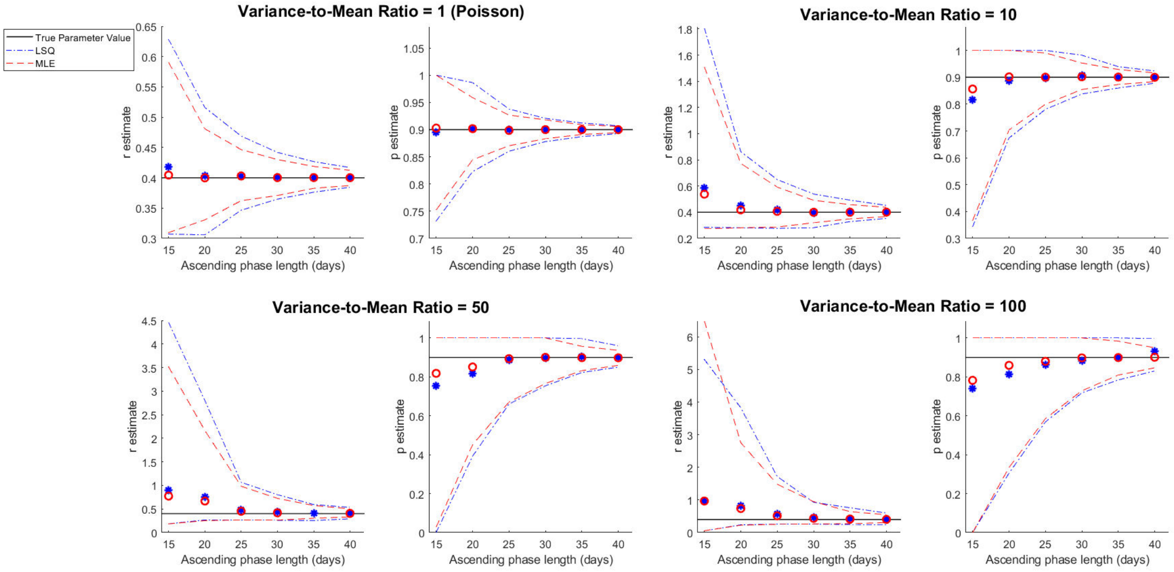


Figure 3. GGM parameter estimates as amount of data available (length of ascending phase) increases. Variance-to-mean ratios of: 1, 10, 50, 100. Mean estimates are represented by the circles and 95% confidence intervals are represented with dashed lines.

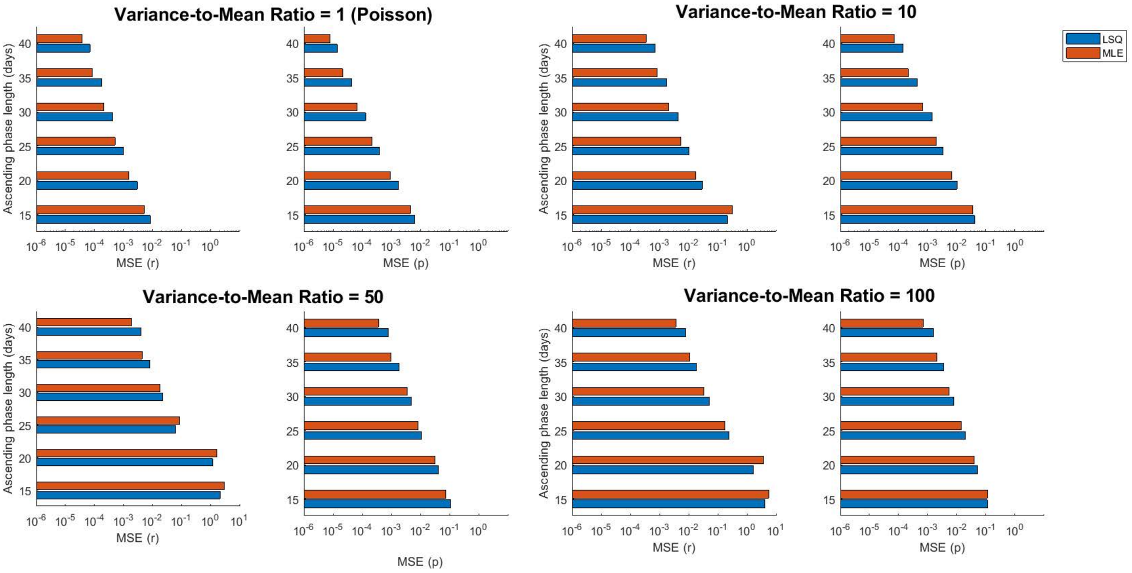


Figure 4. MSE of GGM parameter estimates for each estimation method (LSQ, MLE) as amount of data available (length of ascending phase) increases. Variance-to-mean ratios of: 1, 10, 50, 100.

3.2. *Example 2: GLM*

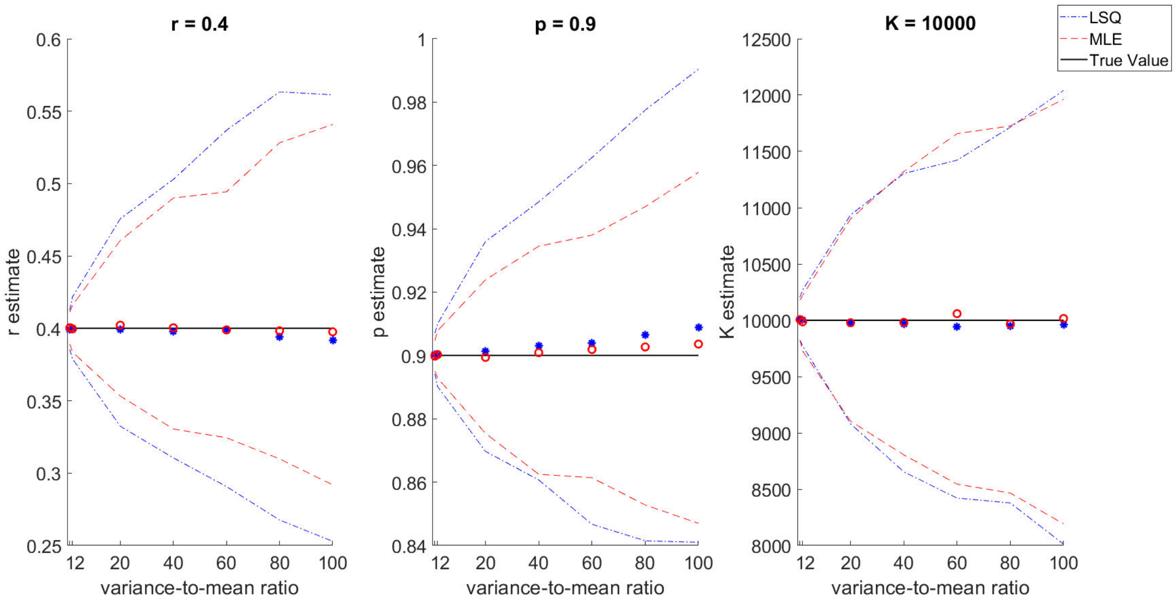


Figure 5. GLM parameter estimation results for increasing levels of variance assumed in the data. Mean estimates (circles) and 95% confidence intervals (dashed lines) are shown for the two estimation methods: nonlinear least squares (LSQ) and maximum likelihood estimation (MLE).

For the generalized logistic growth model (GLM), we jointly estimate all three parameters: r, p, and K. Again, we see that the uncertainty surrounding the parameters (width of the confidence intervals) increases with the increasing variance-to-mean ratio (Figure 5). Again, the largest % difference observed was from $d = 2$ to $d = 20$, with increases in CI width of 232%, 226%, and 264% (r, p, and K, respectively) for MLE, and 235%, 235%, and 272% for LSQ. Percent increases were comparable for the two methods, with LSQ consistently yielding slightly wider CIs for r and p, and nearly equivalent CIs for K (Figure 5).

Comparing the methods, we find similar results for the uncertainty of r and p as with the GGM. LSQ consistently yields CIs that are 17.1-33.8% wider for r and 18-33.9% wider for p. Again, these CI width differences are small for the parameters (¡0.08 for all), but yield high relative differences due to the small values. For parameters of larger magnitude, like K, the relative difference between the methods is much smaller. The relative CI width difference comparing LSQ to MLE ranges from -3.7 - 6.8% across the levels of overdispersion. At $d = 60$, LSQ has a smaller CI width, hence the negative 3.7%.

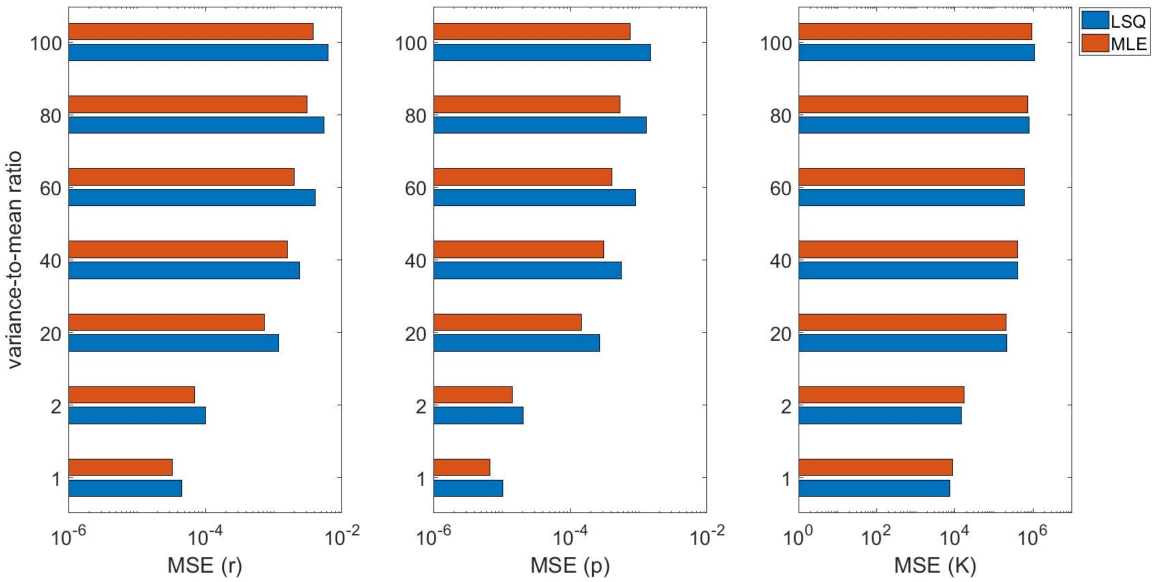


Figure 6. Mean squared error of the distribution of GLM parameter estimates (M=500) for increasing levels of error assumed (variance-to-mean ratio) is shown for both estimation methods: nonlinear least squares (LSQ) and maximum likelihood estimation (MLE). Results for r, p, and K are presented from left to right.

For both LSQ and MLE, the mean parameter estimates remain around the true value line for r and K. The estimates for p begin to deviate from the true line in an upward trajectory, with the mean estimates from LSQ rising faster (so further from the true value) than MLE. The MSE for LSQ ranges from 1.4–2.05 times and 1.46-2.25 times the MSE for MLE (r and p, respectively). These differences are again very small, with the largest difference in MSE for r and p being less than 0.0025. Further, for K estimates, MLE and LSQ yielded minimal differences in MSE, or accuracy (Figure 6). For K, the MSE for LSQ ranged from 0.84–1.08 times the MSE for MLE. This indicates that the MSE is at times larger for LSQ and at times larger for MLE; for $d = 1, 2, 40$, and 60, the MSE is smaller for LSQ, and for $d = 20, 80, 100$, the MSE is lower for MLE. This may indicate that LSQ can provide more accurate estimates than MLE for data with lower levels of overdispersion, but MLE provides more accurate estimates for highly overdispersed data.

4. Discussion and conclusions

The results of the simulations show a clear pattern of increasing uncertainty (assessed by CI width) as the variance-to-mean ratio (d), or overdispersion, increases. The examples (Figures 1 and 5) show near linear growth for the upper and lower bounds of the estimates, increasing as overdispersion increases. Results for r and p for both GGM and GLM reveal that MLE consistently yields more accurate (lower MSE) and precise (smaller CI width) estimates compared to LSQ, though it should be noted that practical differences in estimates are small. For the large parameter, K, differences between methods are even smaller. Across all levels of overdispersion, the relative difference in CI width is less than 6.9%. Further, neither of the methods consistently yields lower MSE for K, indicating comparable MSE, or accuracy, of the methods for large-scale parameters.

For the generalized logistic model (Example 2), we used the entire incidence curve to fit the model, but as explained above, the GGM (Example 1) is strictly increasing and is not flexible enough to detect a peak or decline phase. To illustrate how the amount of data affects the results of parameter estimation for the GGM, we used an increasing number of days (ranging from 15-40 days) for the ascending phase and conducted the analysis across the range for four different levels of overdispersion. It is clearly seen (Figure 3) that using more data for the ascending phase decreases the uncertainty of parameter estimates, for all levels of overdispersion in the data.

Not only do the confidence interval widths significantly decrease as the ascending phase increases, but the mean estimates trend toward the true value, resulting in smaller MSE values as well (Figure 3, 4). In terms of bias, our simulation study indicates that using too few data points in the presence of overdispersion may result in biased parameter estimates. Each 5 day increase in ascending phase length yielded an improvement in MSE. Further, both estimation methods are based on unbiased estimating equations, and thus there is essentially no difference in terms of bias between the two methods. Figure 4 shows minimal differences in MSE between the two methods for each ascending phase length within each variance-to-mean ratio.

The general descending pattern of the percent improvement in MSE (for 5 day increases in ascending phase) is mostly consistent for each level of overdispersion, though at variance-to-mean ratios of 50 and 100, this pattern is not seen until t = 25. This suggests that the signal still cannot be distinguished from the noise for ascending phases lower than 25 days, and thus an increase of 5 days does not provide much improvement in model fit. While significantly more data is required in the presence of overdispersion, the results suggest that even when overdispersion is suspected, uncertainty and bias of estimation results can be mitigated as more data become available.

These analyses were conducted using data simulated directly from the model corresponding to each example. This is a limitation in that we cannot generalize these results to scenarios with real-world data issues. Further, the results are specific to the parameter values noted, and thus cannot be generalized to all configurations of the models. Because of the difference in results between small parameters (r, p) and large parameters (K) in this study, future studies should look into this pattern

(MLE slightly outperforms LSQ for small parameters, but performs equivalently for large parameters). It would be of interest to investigate whether this holds true when including other parameters or models. Similar to the time-varying analysis illustrated here for the GGM, future studies could also conduct analyses on only the ascending phase of other models, for example the GLM, as we looked at the entire incidence curve. Further, infectious disease outbreak data, as time series, are naturally correlated, but the fitting performed in our analyses were based on marginal distributions in each time period, assuming independence. Future studies could include a covariance structure while fitting to outbreak data, which would require specification of the correlation structure specific to the disease application.

Overall, the results demonstrate two simple estimation methods that work well, and nearly equivalently, in the presence of little to no overdispersion, but may have significant uncertainty as the level of overdispersion increases depending on the amount of available data. It is also shown that more data is needed to provide precise confidence intervals in the presence of increasing levels of overdispersion, which implies that the utilization of more data can resolve potential identifiability issues when high levels of overdispersion are suspected. For both models shown, LSQ and MLE-Poisson provide little difference in results with regards to both parameter accuracy and precision.

Acknowledgments

We thank Dr. Ping Yan (Public Health Agency of Canada) for interesting discussions relating to overdispersion and parameter estimation.

KR is supported by a Georgia State University Second Century Initiative (2CI) doctoral fellowship.

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

1. P. McCullagh and J. A. Nelder, *Generalized linear models.* Monographs on statistics and applied probability. London ; New York : Chapman and Hall, 1983., 1983.
2. P. Yan and G. Chowell, Quantitative methods for investigating infectious disease outbreaks, Submitted for publication, 2019.
3. R. Williams, Heteroskedasticity, 2015.
4. C. Dean and E. Lundy, Overdispersion, 2014. In *Wiley StatsRef: Statistics Reference Online*.
5. K. Roosa and G. Chowell, Assessing parameter identifiability in compartmental dynamic models using a computational approach: application to infectious disease transmission models, *Theor. Biol. Med. Mod.*, **16** (2019), 1.
6. B. Efron and R. Tibshirani, *An introduction to the bootstrap.* Monographs on statistics and applied probability: 57. New York: Chapman and Hall, c1993., 1993.
7. G. Chowell, Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts, *Infect. Dis. Model.*, **2** (2017), 379–398.

8. R. Anderson and R. May, *Infectious Diseases of Humans: Dynamics and Control*, New York ; Oxford University Press, 1991., 1991.
9. O. Diekmann, J. A. Heesterbeek and J. A. Metz, On the definition and the computation of the basic reproduction ratio r0 in models for infectious diseases in heterogeneous populations, *J. Math. Biol.*, **28** (1990), 365–382.
10. G. Chowell, C. Viboud, J. M. Hyman, et al., The western africa ebola virus disease epidemic exhibits both global exponential and local polynomial growth rates, 2014.
11. C. Viboud, L. Simonsen and G. Chowell, A generalized-growth model to characterize the early ascending phase of infectious disease outbreaks, *Epidemics*, **15** (2016), 27–37.
12. D. W. Shanafelt, G. Jones, M. Lima, et al., Forecasting the 2001 foot-and-mouth disease epidemic in the uk, *ECOHEALTH*, **15** (2018), 338–347.
13. G. Chowell, D. Hincapie-Palacio, J. Ospina, et al., Using phenomenological models to characterize transmissibility and forecast patterns and final burden of zika epidemics, *PLOS Currents Outbreaks*, 2016.
14. G. Chowell, H. Nishiura and L. M. A. Bettencourt, Comparative estimation of the reproduction number for pandemic influenza from daily case notification data, *J. R. Soc. Interface*, **4** (2007), 155–166.
15. L. Dinh, G. Chowell and R. Rothenberg, Growth scaling for the early dynamics of hiv/aids epidemics in brazil and the influence of socio-demographic factors, *J. Theor. Biol.*, **442** (2018), 79–86.
16. B. Pell, Y. Kuang, C. Viboud, et al., Using phenomenological models for forecasting the 2015 ebola challenge, *Epidemics*, 22(The RAPIDD Ebola Forecasting Challenge), (2018), 62–70.
17. T. Ganyani, K. Roosa, C. Faes, et al., Assessing the relationship between epidemic growth scaling and epidemic size: The 201416 ebola epidemic in west africa, *Epidemiol. Infect.*, **147** (2018), e27.
18. C. Z. Mooney, *Monte Carlo Simulation*. Sage University Paper series on Quantitiative Applications in the Social Sciences. Thousand Oaks, CA: Sage, 1997.
19. I. J. Myung, Tutorial on maximum likelihood estimation, *J. Math. Psychol.*, **47** (2003), 90.
20. K. Kashin, Statistical inference: Maximum likelihood estimation, 2014.