Applying the stochastic Galerkin method to epidemic models with individualised parameter distributions

D. B. Harman[1] P. R. Johnston[2]

(Received 24 December 2015; revised 27 July 2016)

Abstract

There are many different models to help predict the likely course an epidemic will take. However, the parameters within these models are often not known with certainty. It is important for this uncertainty to be incorporated into these models to ensure accurate predictions. This article considers the stochastic Galerkin method to solve an SIR model with uncertainty in its parameters. A data set from an influenza outbreak in a boarding school is then investigated. Rather than just finding the 'best' values for the parameters, several possible probability distributions for the parameters in the SIR model are determined. The stochastic Galerkin method is then used to determine the mean solution of the model as well as its variance.

DOI:10.21914/anziamj.v57i0.10394 gives this article, © Austral. Mathematical Soc. 2016. Published August 23, 2016, as part of the Proceedings of the 12th Biennial Engineering Mathematics and Applications Conference. ISSN 1445-8810. (Print two pages per sheet of paper.) Copies of this article must not be made otherwise available on the internet; instead link directly to the DOI for this article.

Contents

1 Introduction

The ability to accurately predict the course of an epidemic is very important. Because of this, there are many different mathematical models for epidemics. The most common of these are compartment models which were first derived by Kermack and McKendrick [7]. From the compartment models, a system of differential equations which models the epidemic can easily be determined.

While compartment epidemic models are relatively easy to derive, the parameters within these models are often not known with certainty [1]. It is important to include this uncertainty in the model in order to account for a range of possible outcomes. One way of representing the uncertainty in the parameters is to make them functions of random variables so that the mean and variance of the model can be calculated. This can be achieved using

Monte Carlo sampling but can be computationally expensive due to its slow convergence [11]. The stochastic Galerkin method is much more efficient.

While there is extensive literature on the stochastic Galerkin method, there is little concerned with its application to epidemic modelling [9, 8, 5, 10, 1, 3]. Roberts [9] derives a probability distribution for the uncertain parameter $\mathcal{R}_0$ from a data set and then applies the stochastic Galerkin method. Santonja and Chen-Charpentier [10] use a small data set, and so assume uniform distributions for the uncertain parameters. Other researchers simply assume the probability distributions of the uncertain parameters [8, 5, 1, 3].

In Section 2, the stochastic Galerkin method is briefly explained and applied to an SIR epidemic model. A data set from an epidemic that spread through a boarding school in England is then investigated in Section 3. Several possible probability distributions for each of the parameters in the SIR model are found and the stochastic Galerkin method is then applied to find the mean solution and its variance. These results are then compared with the original data set.

2 The SIR model

The SIR model is one of the simplest and most well known of the epidemic compartment models [7]. In the SIR model, each person in the population is placed into one of three compartments. The individual is either susceptible S, infected I, or recovered R.

The system of differential equations for the SIR model (without births or deaths) is

$$\frac{dS}{dt} = -\beta SI\,, \quad \frac{dI}{dt} = \beta SI - \gamma I\,, \quad \frac{dR}{dt} = \gamma I\,, \tag{1}$$

where β is the 'contact rate' and $1/\gamma$ is the average recovery time from the disease [4]. For simplicity, S, I and R are normalised so that $S+I+R=1$. As dS/dt and dI/dt do not depend on R, the system is solved numerically using only the first two equations with R simply given by $R=1-S-I$.

While β and γ are usually assumed to be constants, they are rarely known with certainty [1]. This uncertainty should be incorporated into the model to ensure the model returns accurate results. To represent the uncertainty in β and γ, they are defined as functions of independent random variables:

$$\beta = f(\xi_1)\,, \quad \gamma = g(\xi_2)\,, \tag{2}$$

where f and g are known functions. The independent random variables ξ_1 and ξ_2 have known probability density functions $w_1(\xi_1)$ and $w_2(\xi_2)$, respectively, and probability spaces $(\Omega_1, \mathcal{F}_1, \mathcal{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathcal{P}_2)$, respectively.

Now that β and γ are functions of random variables, equation (1) can no longer be trivially solved using a single call to an ODE solver (such as MATLAB's `ode45`). The mean and variance of the model could be determined using Monte Carlo sampling, but this can be computationally expensive depending upon the distributions of ξ_1 and ξ_2. A more efficient alternative is to use the stochastic Galerkin method.

2.1 Applying the stochastic Galerkin method to the SIR model

To apply the stochastic Galkerin method, we expand

$$\begin{aligned} S(t,\xi_1,\xi_2) &= \sum_{i,j=0}^{\infty} S_{ij}(t)\Psi_i(\xi_1)\Phi_j(\xi_2)\,, \\ I(t,\xi_1,\xi_2) &= \sum_{i,j=0}^{\infty} I_{ij}(t)\Psi_i(\xi_1)\Phi_j(\xi_2)\,, \end{aligned} \tag{3}$$

where $\Psi_i(\xi_1)$ and $\Phi_j(\xi_2)$ are orthogonal polynomials whose weight functions are $w_1(\xi_1)$ and $w_2(\xi_2)$, respectively [2]. The deterministic functions $S_{ij}(t)$ and $I_{ij}(t)$, which only depend upon time, need to be determined.

Substituting equations (2) and (3) into equation (1) gives

$$\begin{aligned}\sum_{i,j=0}^{\infty} \frac{dS_{ij}(t)}{dt}\Psi_i(\xi_1)\Phi_j(\xi_2) &= -f(\xi_1) \sum_{i,j,m,n=0}^{\infty} [S_{ij}(t)I_{mn}(t)\Psi_i(\xi_1)\Phi_j(\xi_2) \\ &\qquad \times \Psi_m(\xi_1)\Phi_n(\xi_2)]\,, \\ \sum_{i,j=0}^{\infty} \frac{dI_{ij}(t)}{dt}\Psi_i(\xi_1)\Phi_j(\xi_2) &= f(\xi_1) \sum_{i,j,m,n=0}^{\infty} [S_{ij}(t)I_{mn}(t)\Psi_i(\xi_1)\Phi_j(\xi_2) \\ &\qquad \times \Psi_m(\xi_1)\Phi_n(\xi_2)] - g(\xi_2)\sum_{i,j=0}^{\infty} I_{ij}(t)\Psi_i(\xi_1)\Phi_j(\xi_2)\,. \end{aligned} \tag{4}$$

Multiplying through by $\Psi_u(\xi_1)\Phi_v(\xi_2)$ $(u, v = 0, 1, 2, \ldots)$, integrating over the probability space, and truncating the expansions at the Pth order gives

$$\begin{aligned}\frac{dS_{uv}}{dt} &= \frac{-1}{\langle(\Psi_u)^2, (\Phi_v)^2\rangle} \sum_{i,m=0}^{P}\sum_{j=0}^{P-i}\sum_{n=0}^{P-m} S_{ij}I_{mn}\langle f\Psi_i\Phi_j\Psi_m\Phi_n, \Psi_u\Phi_v\rangle\,, \\ \frac{dI_{uv}}{dt} &= \frac{1}{\langle(\Psi_u)^2, (\Phi_v)^2\rangle} \sum_{i,m=0}^{P}\sum_{j=0}^{P-i}\sum_{n=0}^{P-m} S_{ij}I_{mn}\langle f\Psi_i\Phi_j\Psi_m\Phi_n, \Psi_u\Phi_v\rangle \\ &\qquad - \frac{1}{\langle(\Psi_u)^2, (\Phi_v)^2\rangle} \sum_{i=0}^{P}\sum_{j=0}^{P-i} I_{ij}\langle g\Psi_i\Phi_j, \Psi_u\Phi_v\rangle\,, \end{aligned} \tag{5}$$

where the inner product is defined as

$$\langle F, G\rangle = \int_{\Omega_2}\int_{\Omega_1} F(\xi_1, \xi_2)G(\xi_1, \xi_2)w_1(\xi_1)w_2(\xi_2)\, d\xi_1 d\xi_2\,.$$

By appropriately choosing the orthogonal polynomials (the weight function of $\Psi_i(\xi_1)$ is equal to $w_1(\xi_1)$ and the weight function of $\Phi_i(\xi_2)$ is equal to $w_2(\xi_2)$), many of the inner products trivially evaluate to zero. This gives a system of $2\binom{P+2}{2}$ *deterministic* differential equations that can be numerically solved, for example, using MATLAB's `ode45`.

While uncertainty was introduced into the SIR model using random variables, the final system of equations is deterministic and therefore only needs to be solved once. This represents a significant speed increase over methods such as Monte Carlo sampling, which requires the model to be solved numerous times to determine the mean and variance.

2.2 Determining mean and variance from the stochastic Galerkin solution

Once $S_{ij}(t)$ and $I_{ij}(t)$ are determined, the mean and variance of the susceptible and infected populations is determined directly from the stochastic Galerkin expansions [12]. The mean solution E for the fraction of infected individuals in the population I, is

$$\mathsf{E}[I(t, \xi_1, \xi_2)] = I_{00}(t)\,,$$

and the variance is

$$\operatorname{Var}[I(t, \xi_1, \xi_2)] = \sum_{i=0}^{P} \sum_{j=0}^{P-i} [I_{ij}(t)]^2 \left\langle (\Psi_i)^2, (\Phi_j)^2 \right\rangle - [I_{00}(t)]^2\,.$$

Therefore, once the stochastic Galerkin expansion is found, the mean solution and variance are straightforwardly calculated from the expansions. The mean solution is simply the zero order term while the variance is the sum of the squares of the remaining terms (along with a constant factor).

3 Influenza outbreak in a boarding school

In the previous section, the stochastic Galerkin method was applied to an SIR model with uncertainty in the parameters. This method is now applied to a data set associated with an influenza outbreak that occurred within a boarding school in the North of England [6].

To begin the new school term, students returned to the boarding school. The school had 763 male students. One of the students returned infected with influenza and as a result, many of the students in the boarding school became infected. Figure 1 shows the fraction of students at the boarding school infected with influenza on a given day.

When a student began showing symptoms, they were confined to bed. Because of this, accurate records of the number of infected students at any time were kept. Also, as it is a boarding school, it is assumed that the students enrolled were effectively isolated from the surrounding population. This makes it a unique and almost ideal data set to investigate.

3.1 Fitting the SIR model to the influenza data

Once a data set is obtained, the next step is to find an epidemic model that fits the data. As the British Medical Journal [6] does not mention any students becoming reinfected after recovering, an SIR model seems the most reasonable.

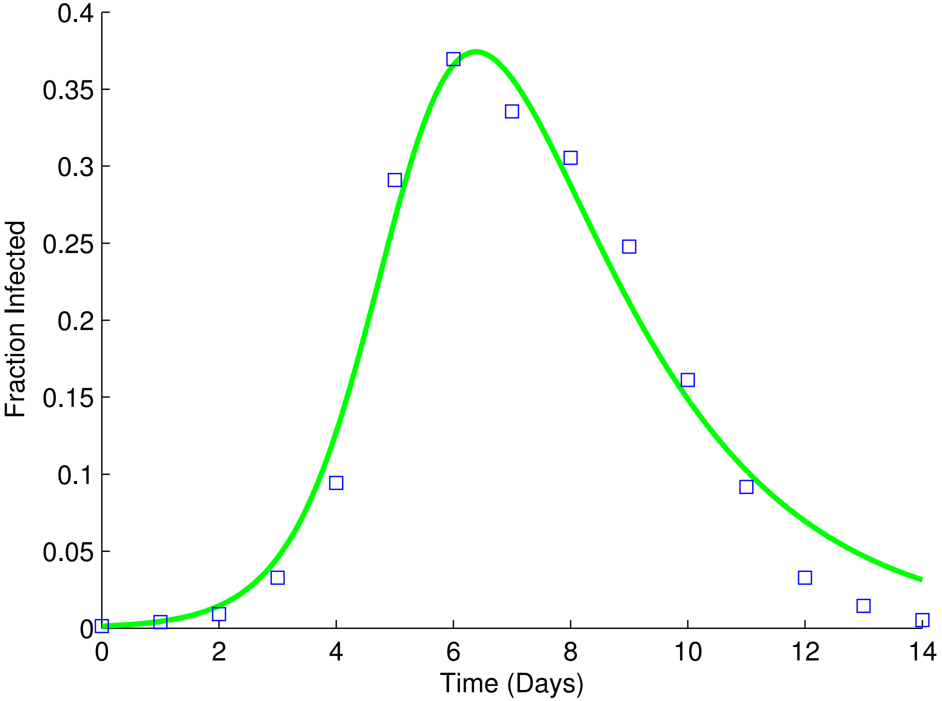
The 'best' values for β and γ are obtained using a simple least squares error formula. For chosen values of β and γ, the error associated with those values compared to the known data points is

$$E_{\beta,\gamma} = \sqrt{\sum_{k=0}^{14} [I_{\beta,\gamma}(k) - I_D(k)]^2}, \tag{6}$$

where $I_D(k)$ is the fraction of infected students on day k according to the data and $I_{\beta,\gamma}(k)$ is the fraction of infected students on day k according to the SIR model with the chosen values of β and γ.

To minimise the error, the best values for β and γ are approximately 1.665 and 0.453, respectively. This gives an error of $E_{1.665,0.453} \approx 0.086$. These values of β and γ were obtained using MATLAB's `fminsearch`. Figure 1

Figure 1: Influenza spreading within a small boarding school. Blue squares are known data points. The green line is the 'best fit' using an SIR model with $\beta = 1.665$ and $\gamma = 0.453$.



shows the normalised data as well as the solution from the SIR model using these values of β and γ. From the graph it is seen that the SIR model fits the data reasonably well, with the exception of the last three data points.

3.2 Determining 'plausible' values for β and γ

Assume for the moment that the influenza outbreak spread outside the relatively isolated confines of the boarding school and into the surrounding population. For example, it could easily be spread outside the school by a teacher who worked at the school but did not live at the school. What information can be determined from the boarding school outbreak that could be applied to modelling the spread of the infection in the surrounding population?

In the previous section, the 'best' values for β and γ were determined from the error formula. Even though these values for β and γ model the outbreak reasonably well, the 'tail' of the outbreak (the last three data points in particular) are modelled rather poorly. If these values of β and γ are used to predict the course of the epidemic in the surrounding population, then the same problem is likely to occur. Additionally, while using these values for β and γ give a 'best guess' of what would happen in the surrounding population, the method provides no information on possible confidence intervals or what else could possibly happen.

In order to determine what could happen if the infection were to spread to the surrounding population, instead of simply using the values of β and γ that minimise the error, a range of values for β and γ are considered. Such a range is obtained by keeping the error below some predefined threshold. For this study, three different error thresholds are considered: $E_{\beta,\gamma} < 0.15$, $E_{\beta,\gamma} < 0.25$ and $E_{\beta,\gamma} < 0.35$.

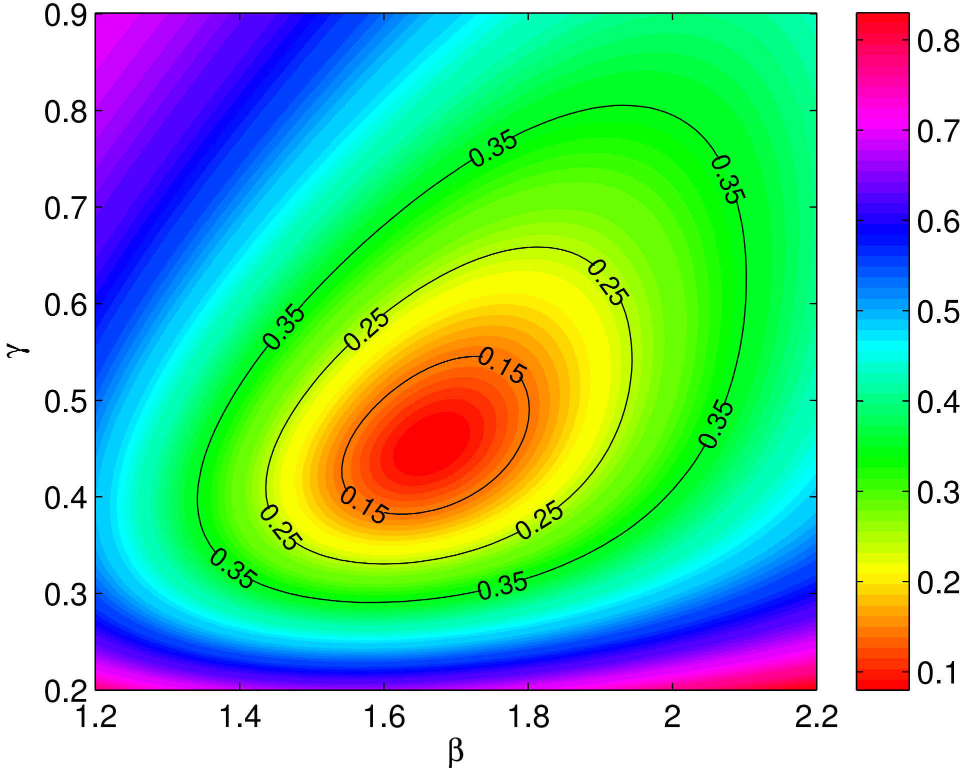
Figure 2 shows a heat map of error $E_{\beta,\gamma}$ for $1.2 \leqslant \beta \leqslant 2.2$ and $0.2 \leqslant \gamma \leqslant 0.9$. The lowest error occurs at the 'best fit' values which are approximately 1.665 and 0.453 for β and γ, respectively. As β or γ move away from the 'best' values, the error increases.

3.3 Determining probability distributions for β and γ

Before the stochastic Galerkin method is applied, expressions for the probability density functions of β and γ need to be determined.

From Figure 2, it is seen that the region of plausible values for a given error threshold forms a simple closed shape. Therefore, many of the points

Figure 2: Error in the SIR model for different values of β and γ compared to known data points. The error is given by equation (6) and the colours in the plot describe this error.



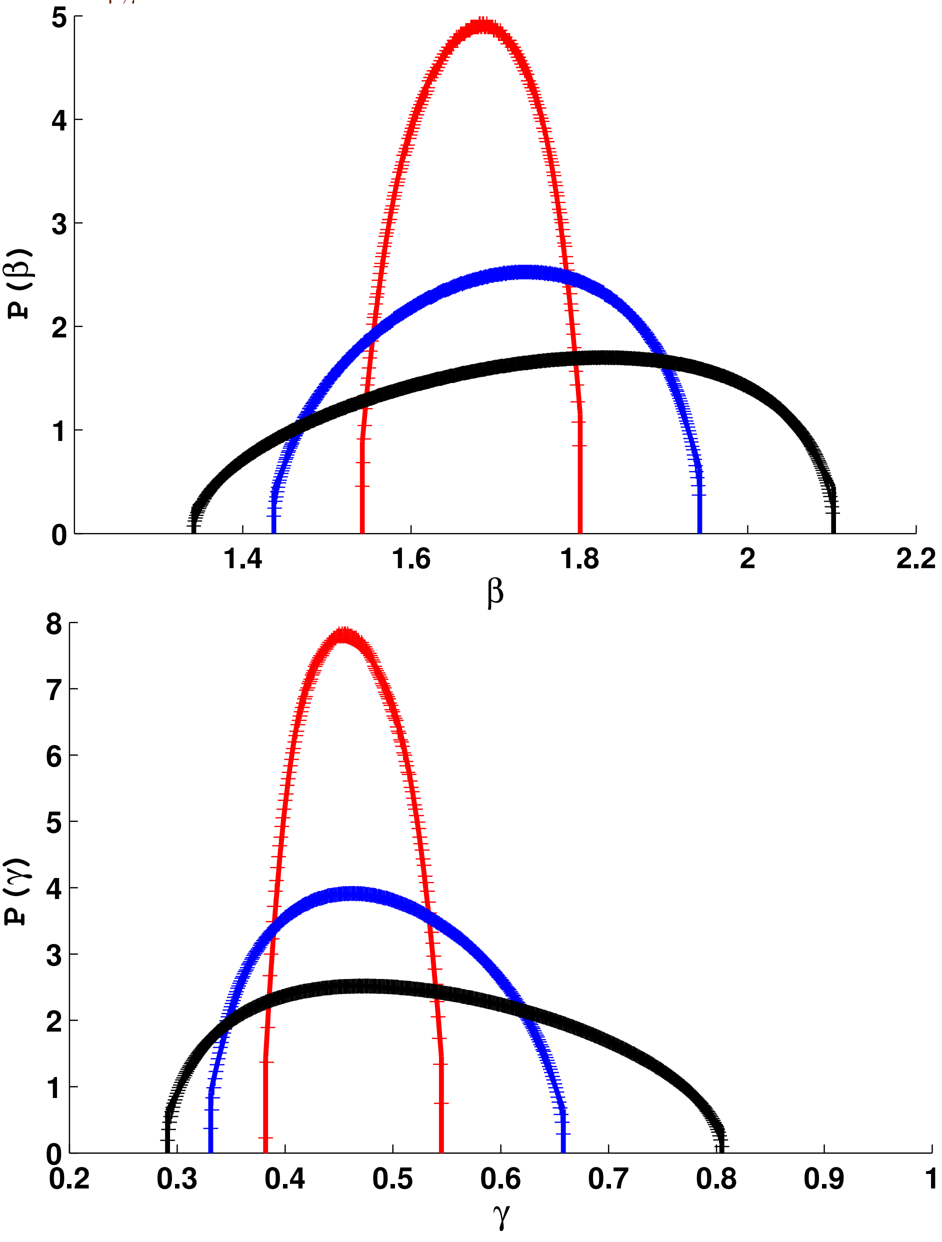
evaluated to produce Figure 2 are redundant. A simple algorithm is instead used to find the boundary of the closed shape. All points within the closed shape are plausible values (based upon a given error threshold) and all points outside the shape have an error greater than the threshold. Only determining the boundary of the closed shape provides a significant speed increase over producing the entire heat map.

Probability density functions for β and γ are determined using only the information about the boundary of the plausible values. For simplicity, β and γ are assumed to have independent distributions. While this is not strictly true, it is an acceptable assumption due to the physical interpretations of β and γ. The rate at which an individual recovers from an infection, $1/\gamma$, is independent of the rate at which they can infect susceptibles, βS.

To determine the probability density function for β, for each plausible value of β (each value of β that has at least one corresponding γ value that is inside the boundary), the number of corresponding γ values that are inside the boundary of the closed shape are counted. The number of corresponding γ values for each plausible β value are plotted as a histogram. The histogram is then approximated by a curve and the area under the curve is calculated numerically. After normalising so that the area under the curve is one, this curve gives an approximation of the probability density function for β at the given error threshold. A similar process is repeated to find the probability density function for γ. Figure 3 shows the calculated probability density for each of the three error thresholds. The plots show that when the error threshold is low, the probability distribution is quite narrow, whereas when the error threshold is increased, the probability distribution becomes wider.

As the probability distributions do not resemble any of the familiar probability distributions, they are approximated by polynomials using MATLAB's `polyfit`. It is found that fifth order polynomials provided a good fit to the curves.

Figure 3: Probability density functions for: (top) β; and (bottom) γ; at different error thresholds. Red has $E_{\beta,\gamma} < 0.15$, blue has $E_{\beta,\gamma} < 0.25$, and black has $E_{\beta,\gamma} < 0.35$.



3.4 Applying the stochastic Galerkin method to the boarding school data

To apply the stochastic Galerkin method, orthogonal polynomials whose weight functions match the probability density functions of the uncertain parameters are required. As the probability distributions for β and γ are non-standard and are approximated using fifth order polynomials, the orthogonal polynomials are unlikely to be known and must to be derived.

To calculate the orthogonal polynomials that have $w(\xi_1)$ as their weight function, the zero order polynomial is assumed to be $\Psi_0(\xi_1) = 1$. The first order polynomial is then given by $\Psi_1(\xi_1) = c_1\xi_1 + c_0$ where the constants c_1 and c_0 are chosen such that

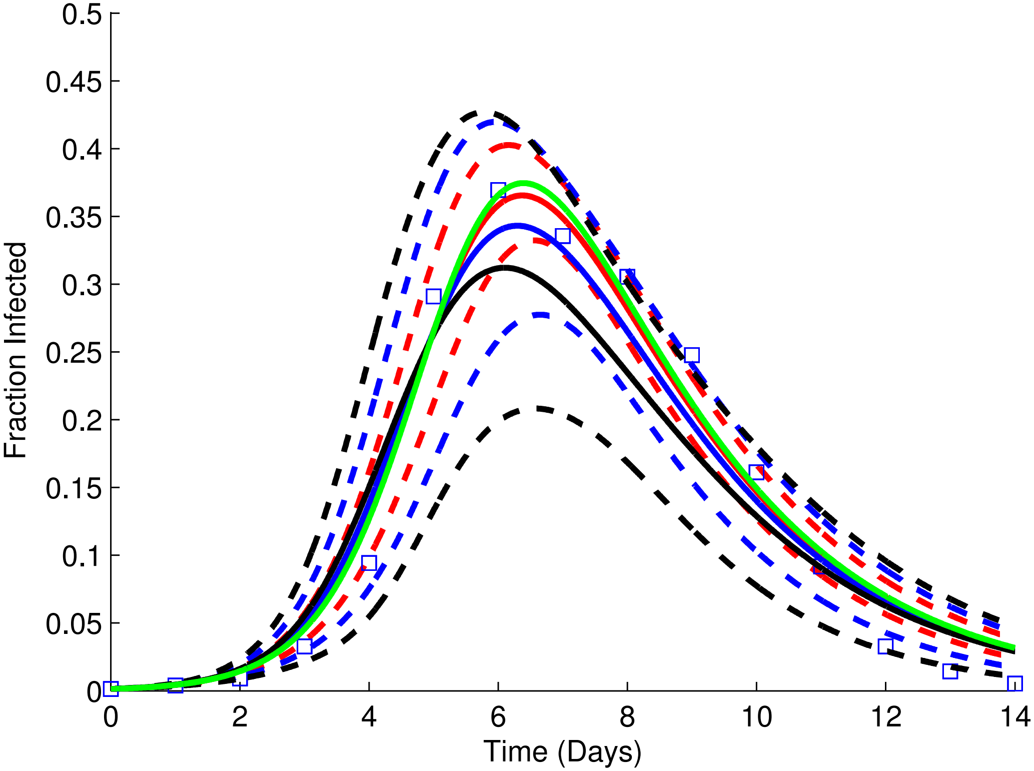
$$\langle \Psi_0(\xi_1), \Psi_1(\xi_1) \rangle = 0 \quad \text{and} \quad \langle \Psi_1(\xi_1), \Psi_1(\xi_1) \rangle = 1 .$$

The coefficients of the higher order orthogonal polynomials are derived through a similar process, ensuring that the inner product with any lower order orthogonal polynomial is zero and the inner product with itself is one.

Once the probability distributions of β and γ are approximated and the associated orthogonal polynomials are determined, the stochastic Galerkin method is applied (as shown in Section 2.1) to find the mean solution for each of the three error thresholds as well as their variances. A third order expansion ($P = 3$) is used and the results are shown in Figure 4. A third order expansion is chosen as it provides accurate results while only needing a small amount of time to derive the orthogonal polynomials and the $2\binom{P+2}{2}$ deterministic differential equations from the stochastic Galerkin method.

Figure 4 shows that the mean solution using $E_{\beta,\gamma} < 0.15$ is very similar to the 'best fit'. This is because the error threshold is only slightly larger than the smallest error possible, so the plausible values of β and γ are very similar to the 'best' values. As the error threshold increases, the mean peak of the epidemic decreases. The three mean solutions are very similar, with the exception of the peaks.

Figure 4: Stochastic Galerkin solutions to the SIR model with different error thresholds. Blue squares are known data points. Green is the ‘best fit’. Red has an error less than 0.15, blue has an error less than 0.25, black has an error less than 0.35. Solid lines are the mean solution and dashed lines are ± one standard deviation from the mean.



For each of the three error thresholds, most of the data points lie within one standard deviation of the mean, with the exception of the last three data points. Even with the largest error threshold, with $E_{\beta,\gamma} < 0.35$, only one of these last three data points is found within one standard deviation of the mean.

In Figure 4, the upper dashed lines (mean plus one standard deviation for each of the error thresholds) all have approximately the same peak whereas the lower dashed lines (mean minus one standard deviation) all have significantly different peaks. Whereas the 'best' values of β and γ only give the most likely scenario, Figure 4 gives confidence intervals for what might occur if the disease spread to the wider community.

4 Conclusion

This article looked at the stochastic Galerkin method and how it is applied to a simple data set from an influenza outbreak.

Rather than simply finding the 'best' values for the parameters of an epidemic based upon a given error formula, ranges of plausible values were instead considered. It was found that these ranges of plausible values formed simple closed shapes in a contour plot, eliminating the need for random sampling to find the ranges of plausible values. By finding the border of this shape, all plausible values are quickly determined for a given error threshold.

Using only the border of the shapes in the contour plot, probability distributions for the parameters are found. As the probability distributions are approximated by polynomials, it is unlikely that the associated orthogonal polynomials are known and therefore will need to be derived before the stochastic Galerkin method is applied. The mean solutions of all the plausible values, as well as their variances, are determined from the stochastic Galerkin expansion.

In the future, we aim to extend this work by assuming that the parameters are no longer independent. This would make determining the probability density functions more complicated but it is hoped that this will provide more accurate results.

References

[1] B. M. Chen-Charpentier and D. Stanescu. “Epidemic models with random coefficients”. In: *Math. Comput. Model.* 52.7–8 (2010). doi:10.1016/j.mcm.2010.01.014, pp. 1004–1010 (cit. on pp. C161, C162, C163).

[2] B. M. Chen-Charpentier et al. “Some recommendations for applying gPC (generalized polynomial chaos) to modeling: An analysis through the Airy random differential equation”. In: *Appl. Math. Comput.* 219.9 (2013). doi:10.1016/j.amc.2012.11.007, pp. 4208–4218 (cit. on p. C163).

[3] D. B. Harman and P. R. Johnston. “Applying the stochastic Galerkin method to epidemic models with uncertainty in the parameters”. In: *Math. Biosci.* 277 (2016). doi:10.1016/j.mbs.2016.03.012, pp. 25–37 (cit. on p. C162).

[4] H. W. Hethcote. “The Mathematics of Infectious Diseases”. In: *SIAM Rev.* 42.4 (2000). doi:10.1137/S0036144500371907, pp. 599–653 (cit. on p. C162).

[5] R. I. Hickson and M. G. Roberts. “How population heterogeneity in susceptibility and infectivity influences epidemic dynamics”. In: *J. Theor. Biol.* 350.0 (2014). doi:10.1016/j.jtbi.2014.01.014, pp. 70–80 (cit. on p. C162).

[6] “Influenza in a boarding school”. In: *Br. Med. J.* 1 (1978). doi:10.1136/bmj.1.6112.586, p. 586 (cit. on pp. C165, C166).

[7] W. O. Kermack and A. G. McKendrick. “A Contribution to the Mathematical Theory of Epidemics”. In: *Proc. R. Soc. Lon. Ser. A* 115.772 (1927). doi:10.1098/rspa.1927.0118, pp. 700–721 (cit. on pp. C161, C162).

[8] M. G. Roberts. “A Two-Strain Epidemic Model With Uncertainty In The Interaction”. In: *ANZIAM J.* 54 (2012). doi:10.1017/S1446181112000326, pp. 108–115 (cit. on p. C162).

[9] M. G. Roberts. “Epidemic models with uncertainty in the reproduction number”. In: *J. Math. Biol.* 66.7 (2013). doi:10.1007/s00285-012-0540-y, pp. 1463–1474 (cit. on p. C162).

[10] F. Santonja and B. Chen-Charpentier. “Uncertainty quantification in simulations of epidemics using polynomial chaos”. In: *Comput. Math. Method. Med.* 2012 (2012). doi:10.1155/2012/742086, p. 742086 (cit. on p. C162).

[11] B. Sudret. “Global sensitivity analysis using polynomial chaos expansions”. In: *Reliab. Eng. Syst. Safe.* 93.7 (2008). doi:10.1016/j.ress.2007.04.002, pp. 964–979 (cit. on p. C162).

[12] D. Xiu. *Numerical Methods for Stochastic Computations: A Spectral Method Approach.* http://press.princeton.edu/titles/9229.html. Princeton University Press, 2010 (cit. on p. C165).

Author addresses

1. **D. B. Harman**, School of Natural Sciences, Griffith University, Queensland 4111, Australia.
mailto:david.harman@griffithuni.edu.au

2. **P. R. Johnston**, School of Natural Sciences, Griffith University, Queensland 4111, Australia.
mailto:p.johnston@griffith.edu.au