Romanian Reports in Physics **72**, 112 (2020)

STOCHASTIC SIMULATION OF EPIDEMICS USING THE MAXIMUM ENTROPY PRINCIPLE AND GENERALIZED POLYNOMIAL CHAOS EXPANSIONS

JULIA CALATAYUD, MARC JORNET

Institut Universitari de Matemàtica Multidisciplinar, Building 8G, access C, 2nd floor, Universitat Politècnica de València, Camí de Vera s/n, 46022, Valencia, Spain
Email: *jucagre@doctor.upv.es*,
Corresponding author's Email: *marcjornet31@gmail.com*

Received January 4, 2020

Abstract. The evolution of epidemics can be better understood using compartmental models based on differential equations. Accounting for modeling and data errors, the transmission parameters of the model must be regarded as random variables. The maximum entropy principle infers consistent probability distributions for the parameters, by maximizing the ignorance on their density functions while not violating physical principles. Once the parameter estimation problem is solved, a non-intrusive approach based on generalized polynomial chaos expansions reconstructs the stochastic solution in terms of the random parameters. This allows for uncertainty quantification to obtain robust predictions for the epidemic. Also, a variance-based sensitivity analysis can be conducted to determine the parameter having the highest impact on the model output.

Key words: epidemic model, uncertainty quantification, maximum entropy principle, generalized polynomial chaos expansion.

1. INTRODUCTION

Mathematical models are an important tool to analyze the dynamics of epidemics. The epidemic is divided into compartments according to the disease stage, where the individuals flow between the compartments as time passes. The models are based on systems of differential equations in which homogeneous mixing is assumed [1–4].

Models involve parameters. If these input parameters have an interpretation and their values can be set experimentally, we have a forward model to describe and forecast the main features of the phenomenon under study. However, in general, parameters must be adjusted from collected data on the response variable. Parameter estimation is an inverse problem.

In the deterministic case, minimization procedures provide the optimal values for the input coefficients. However, deterministic models do not take into account inaccuracies due to the complexity of the phenomenon under study, lack of informa-

tion, missed data, errors, etc. Thus, it makes sense to consider that the transmission parameters vary randomly on a probability space. In this case, the inverse problem consists in inferring probability distributions for the parameters.

The maximum entropy principle derives probability distributions for the parameters consistently. Subject to constraints on the support and the statistical moments of the parameters, the entropy (ignorance) on the density functions described by the Shannon measure is maximized using Lagrange multipliers [5, 6].

Once the probability distributions for the random parameters are set, uncertainty quantification consists in studying the propagation of uncertainty from the inputs to the model output (forward problem). Generalized polynomial chaos (gPC) expand the model solution in terms of orthogonal polynomials with respect to the input parameters [7, 8]. For example, if the parameters follow a Normal, Uniform or Exponential distribution, the corresponding family of orthogonal polynomials is Hermite, Legendre or Laguerre, respectively. The non-intrusive approach determines the coefficients of the gPC expansion *via* integration rules, and the existing codes for solving the original model are not modified. This method improves the efficiency and accuracy of Monte Carlo simulation strategies, which are based on calculating sample statistics from generated realizations [9].

We apply this methodology to an obesity model [10] (region of Valencia, Spain). According to the Body Mass Index size, the adulthood population is divided into compartments: normal weight, overweight and obese. A system of differential equations describes the growth rate of the obese population. From the deterministic model [10], we propose probability distributions for the parameters relying on the maximum entropy principle. Then uncertainty quantification is conducted *a posteriori* by employing gPC expansions. A variance-based sensitivity analysis is carried out, to determine the flow between compartments that has the highest impact on the growth of obesity prevalence.

2. MATHEMATICAL MODEL

Obesity can be considered as a worldwide epidemic [11]. It is not only a serious health concern, but also a public economic problem [12]. Hence effective policies and control measures are necessary in order to mitigate the augment of excess weight.

In [10], a mathematical model for the evolution of obesity in the region of Valencia, Spain, was proposed. The region of Valencia is an autonomous region and historical nationality located in eastern Mediterranean Spain, with an area of 23 255 km^2 and a population of 4–5 million inhabitants for the last twenty years. The model is based upon a system of ordinary differential equations with input parameters, where obesity is "transmitted" by social pressure and contacts that entail

unhealthy habits [13, 14]. Obesity is regarded as a contagious disease of social transmission. The population is divided into three subpopulations according to the Body Mass Index (BMI) size. The formula for BMI is $\text{BMI} = \text{weight}/\text{height}^2$, where the weight is measured in kilograms and the height in meters: normal weight individuals have BMI less than 25, overweight individuals have BMI in the range $[25, 30)$, and obese individuals have it greater than or equal to 30. Homogeneous mixing between the individuals and constant population size are assumed [1].

Let $N(t)$, $S(t)$ and $O(t)$ denote the proportion of normal weight, overweight and obese individuals who are 24–65 years old in the region of Valencia, Spain, at week $t \geq 0$ since year 2000. The dynamics of these three functions is explained *via* the following model of nonlinear ordinary differential equations:

$$\begin{cases} N'(t) = \mu N^0 - \mu N(t) - \beta N(t)\,[S(t) + O(t)] + \rho S(t), & t \geq 0, \\ S'(t) = \mu S^0 + \beta N(t)\,[S(t) + O(t)] - (\mu + \rho + \gamma) S(t) + \epsilon O(t), & t \geq 0, \\ O'(t) = \mu O^0 + \gamma S(t) - (\mu + \epsilon) O(t), & t \geq 0, \\ N(0) = N_0, \\ S(0) = S_0, \\ O(0) = O_0. \end{cases} \tag{1}$$

Here, $\mu > 0$ is the average stay time in the system, $\beta > 0$ is the force of "infection" (transmission rate because of social pressure), $\rho > 0$ is the rate at which an overweight individual becomes normal weight, $\gamma > 0$ is the rate at which an overweight person becomes obese, and $\epsilon > 0$ is the rate at which an obese adult becomes overweight. Their units are weeks^{-1}. On the other hand, N^0, S^0 and O^0 are the proportion of normal weight, overweight and obese individuals, respectively, coming from the 23-year old group. The sum $N(t) + S(t) + O(t)$ is equal to 1.

In [10], the initial conditions were set from health surveys:

$$N_0 = 0.522,\ S_0 = 0.362,\ O_0 = 0.116.$$

That is, 52.2%, 36.2% and 11.6% of the inhabitants aged 24–65 years old in the region of Valencia were normal weight, overweight and obese adults, respectively, in the year 2000. The parameters of the model were estimated by health surveys, technical reports and a least-squares fitting, giving as a result:

$$\begin{gathered} \hat{\beta} = 0.00085,\ \hat{\mu} = 0.000469,\ \hat{\gamma} = 0.0003,\ \hat{\epsilon} = 0.000004,\ \hat{\rho} = 0.000035, \\ \hat{N}^0 = 0.704,\ \hat{S}^0 = 0.25,\ \hat{O}^0 = 0.046. \end{gathered} \tag{2}$$

This gives rise to a deterministic model that can predict the incidence of obesity *via* pointwise estimates.

3. MAXIMUM ENTROPY PRINCIPLE

Given a parameter θ from the model, we wish to infer a probability distribution for θ that does not violate the prior information about it. The Shannon entropy of θ is defined as

$$\mathcal{S}(\theta) = -\int_a^b f_\theta(\theta)\log f_\theta(\theta)\,\mathrm{d}\theta, \tag{3}$$

where f_θ is the probability density function of θ and $[a,b]$ denotes its support (of course, it could be $a=-\infty$ and/or $b=\infty$). Suppose that we have prior information about θ:

$$\mathbb{E}[\theta^k] = \int_a^b \theta^k f_\theta(\theta)\,\mathrm{d}\theta = f^k,\ 1\leq k\leq m. \tag{4}$$

Here $\mathbb{E}$ is the expectation operator and f^k denotes the k-th statistical moment of θ. For instance, f^1 is the mean of θ. The maximum entropy principle maximizes (3) subject to (4). See [5, 6]. The maximum is attained when

$$f_\theta(\theta) = \mathbb{1}_{[a,b]}(\theta)\exp\left(-\lambda_0 - \sum_{i=1}^m \lambda_k\theta^k\right), \tag{5}$$

for certain Lagrange constants $\lambda_0,\ldots,\lambda_m\in\mathbb{R}$. These constants are determined by solving the nonlinear system of equations that appears when substituting (5) into (4).

When the support of θ is $[0,\infty)$ and only information about the mean f^1 is available ($m=1$), the density (5) becomes the density function of the Exponential(λ) distribution, $f_\theta(\theta)=\lambda\mathrm{e}^{-\lambda\theta}$, $\theta\geq 0$, where the rate parameter is $\lambda = 1/f^1$.

In model (1), we consider the transmission parameters between compartments, β, γ, ϵ and ρ, as independent random variables. These parameters are the most complex from the model and, moreover, variations of them by policy makers determine the future evolution of the epidemic. In this manner, we are incorporating randomness into the model while keeping it in its simplest form. From the parameter estimates (2), we infer the probability distributions

$$\begin{gathered}\beta\sim\text{Exponential}(\hat\beta^{-1}),\ \ \gamma\sim\text{Exponential}(\hat\gamma^{-1}),\\ \epsilon\sim\text{Exponential}(\hat\epsilon^{-1}),\ \ \rho\sim\text{Exponential}(\hat\rho^{-1}).\end{gathered} \tag{6}$$

4. GENERALIZED POLYNOMIAL CHAOS EXPANSIONS

Given the probability distributions (6), the stochastic solution to (1) formed by $N(t)$, $S(t)$ and $O(t)$, is expanded in terms of tensor Laguerre polynomials [7]. We consider four families of Laguerre polynomials (in increasing polynomial degree), $\{L_i^j(z)\}_{i=0}^\infty$, $j=1,2,3,4$, which are orthogonal with respect to the density functions

of β, γ, ϵ and ρ (weighting functions), respectively. Consider the orthogonal polynomials for degrees $0 \leq i \leq p$. The tensor product construction is

$$L_i(z_1, z_2, z_3, z_4) = L_{i_1}^1(z_1) L_{i_2}^2(z_2) L_{i_3}^3(z_3) L_{i_4}^4(z_4),$$

where i is associated with the multi-index (i_1, i_2, i_3, i_4) bijectively, $i_1, i_2, i_3, i_4 \geq 0$, $i_1 + i_2 + i_3 + i_4 \leq p$. The index i ranges from 1 to $P = (p+4)!/(p!4!)$. The family of multivariate Laguerre polynomials $\{L_i(\beta, \gamma, \epsilon, \rho)\}_{i=1}^P$ is orthogonal with respect to the joint density function of the random vector $(\beta, \gamma, \epsilon, \rho)$ (weighting function), due to independence. When $p, P \to \infty$, we have the mean square gPC expansions

$$N(t) = \sum_{i=1}^{\infty} \tilde{N}_i(t) L_i(\beta, \gamma, \epsilon, \rho),$$

$$S(t) = \sum_{i=1}^{\infty} \tilde{S}_i(t) L_i(\beta, \gamma, \epsilon, \rho),$$

$$O(t) = \sum_{i=1}^{\infty} \tilde{O}_i(t) L_i(\beta, \gamma, \epsilon, \rho),$$

where

$$\tilde{N}_i(t) = \frac{\mathbb{E}[N(t) L_i(\beta, \gamma, \epsilon, \rho)]}{\mathbb{E}[L_i(\beta, \gamma, \epsilon, \rho)^2]},$$

$$\tilde{S}_i(t) = \frac{\mathbb{E}[S(t) L_i(\beta, \gamma, \epsilon, \rho)]}{\mathbb{E}[L_i(\beta, \gamma, \epsilon, \rho)^2]},$$

$$\tilde{O}_i(t) = \frac{\mathbb{E}[O(t) L_i(\beta, \gamma, \epsilon, \rho)]}{\mathbb{E}[L_i(\beta, \gamma, \epsilon, \rho)^2]},$$

are the Fourier coefficients. As $N(t)$, $S(t)$ and $O(t)$ are not known, the expectations of the denominators of $\tilde{N}_i(t)$, $\tilde{S}_i(t)$ and $\tilde{O}_i(t)$ are approximated *via* tensor Gauss-Laguerre quadrature integration rules, each one of degree Q and exact for polynomials of degree $2Q-1$. This requires to solve (1) deterministically for each integration node β, γ, ϵ and ρ, by means of numerical methods (Q^4 resolutions in all). This is a non-intrusive approach, as the deterministic solvers for the governing model (1) are employed to solve the stochastic problem.

Since $N(t)$, $S(t)$ and $O(t)$ are C^∞ with respect to β, γ, ϵ and ρ, for each $t \geq 0$, the convergence of the gPC expansions is exponential with p, and the convergence of the tensor Gauss-Laguerre quadrature is exponential with Q.

The gPC expansions and their orthogonality properties allow for approximating the mean and the variance of $N(t)$, $S(t)$ and $O(t)$, by truncating the infinite-term series of the relations

$$\mathbb{E}[N(t)] = \tilde{N}_1(t), \quad \mathbb{V}[N(t)] = \sum_{i=2}^{\infty} \left(\tilde{N}_i(t)\right)^2 \mathbb{E}[L_i(\beta, \gamma, \epsilon, \rho)^2],$$

$$\mathbb{E}[S(t)] = \tilde{S}_1(t), \quad \mathbb{V}[S(t)] = \sum_{i=2}^{\infty} \left(\tilde{S}_i(t)\right)^2 \mathbb{E}[L_i(\beta,\gamma,\epsilon,\rho)^2],$$
$$\mathbb{E}[O(t)] = \tilde{O}_1(t), \quad \mathbb{V}[O(t)] = \sum_{i=2}^{\infty} \left(\tilde{O}_i(t)\right)^2 \mathbb{E}[L_i(\beta,\gamma,\epsilon,\rho)^2].$$

5. UNCERTAINTY QUANTIFICATION

In this section, we report the numerical results obtained for the obese subpopulation by employing Laguerre chaos. The Laguerre families have been truncated at degree $p = 3$, so the final gPC expansion consists of $P = (p+4)!/(p!4!) = 35$ terms. The Gauss-Laguerre quadrature rule for the gPC coefficients is the tensor product construction of four univariate rules of degree $Q = 10$. It was checked that these orders are sufficient to reach good accuracy. In Figure 1, we plot the mean of $O(t)$ until week $t = 800$ (approximately 15 years), together with a confidence interval of the form $\mathbb{E}[O(t)] \pm \sqrt{\mathbb{V}[O(t)]}$. The prevalence of the epidemic is predicted not only with pointwise estimates, as a deterministic model would do, but also with probability bands, which permits a more robust and faithful description of the dynamics. We observe that the number of obese adults presents an increasing pattern.

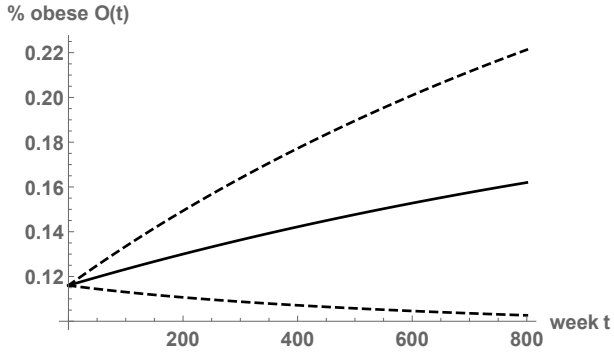

Fig. 1 – Estimated prevalence of obesity in the adulthood Valencian population since year 2000. The solid line reports the mean and the dashed lines represent the confidence interval.

Figures 2 and 3 report the evolution of the normal weight and overweight subpopulations, using the mean values and confidence intervals (again, mean plus minus standard deviation). In all three subpopulations, the confidence intervals widen as we distance from the initial condition, as the uncertainty gets bigger. The overweight population grows, while the normal weight population is worryingly decreasing. This may be due to the bad nutritional and physical habits in the society along the last years.

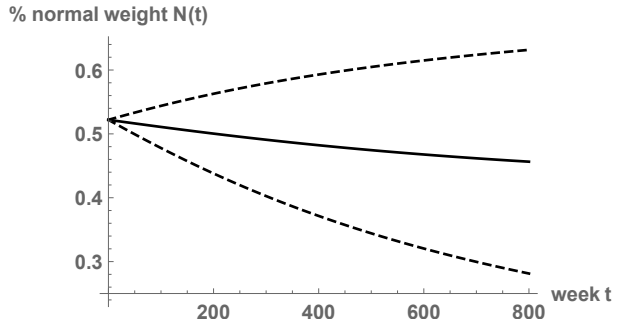

Fig. 2 – Estimated prevalence of normal weight state in the adulthood Valencian population since year 2000. The solid line reports the mean and the dashed lines represent the confidence interval.

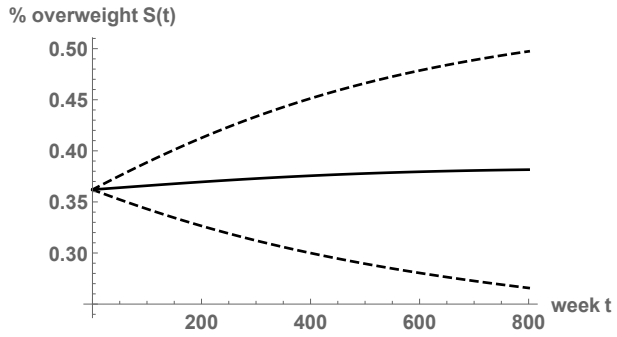

Fig. 3 – Estimated prevalence of overweight state in the adulthood Valencian population since year 2000. The solid line reports the mean and the dashed lines represent the confidence interval.

In conclusion, effective policies must be carried out in order to invert this alarming trend. It is possible to further analyze which strategy is the optimal one for tackling and eventually reducing the spread of the epidemic, by conducting a variance-based sensitivity analysis employing Sobol' indices [15, 16]. Let us consider the obese subpopulation, which is the most problematic from the point of view of health and economy. Given a transmission parameter θ, its Sobol' index is defined as

$$\mathcal{I}_\theta(t) = \frac{\mathbb{E}\left[(\mathbb{E}[O(t)] - \mathbb{E}[O(t)|\theta])^2\right]}{\mathbb{V}[O(t)]} = \frac{\mathbb{V}[\mathbb{E}[O(t)|\theta]]}{\mathbb{V}[O(t)]}.$$

The numerator measures how the expectation of $O(t)$ changes when conditioning to θ, that is, the influence of θ on the output. The denominator scales the Sobol' coefficient so that it lies within $[0, 1]$. Values near 1 mean larger impact of θ on the output, therefore the transit between compartments corresponding to θ should be the one controlled by health authorities. The Sobol' index is very easy to compute once the gPC expansion of $O(t)$ is available:

$$\mathcal{I}_\theta(t) = \frac{\sum_{i=2}^{\infty}\{(\tilde{O}_i(t))^2\mathbb{E}[L_i(\beta,\gamma,\epsilon,\rho)^2] : L_i \text{ only depends on } \theta\}}{\sum_{i=2}^{\infty}(\tilde{O}_i(t))^2\mathbb{E}[L_i(\beta,\gamma,\epsilon,\rho)^2]}.$$

In practice, the infinite-term series are truncated to order P. Figures 4 and 5 report the Sobol' indices of γ, on the one hand, and of β, ϵ and ρ, on the other hand. The parameters β, ϵ and ρ have very small Sobol' indices (especially ϵ and ρ), while γ has its Sobol' index near 1 along all the time domain. The coefficient γ models the flow from overweight to obese states. Hence prevention strategies are more important than treatment strategies to control adulthood obesity.

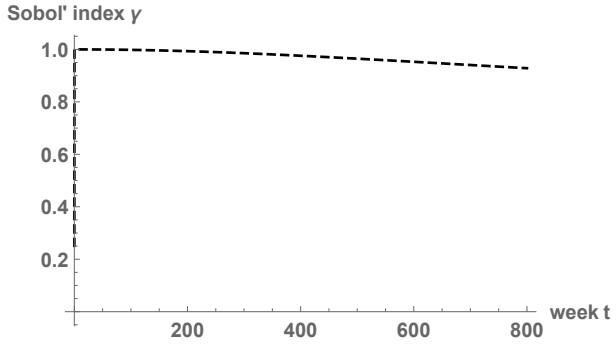

Fig. 4 – Sobol' index of the transmission parameter γ.

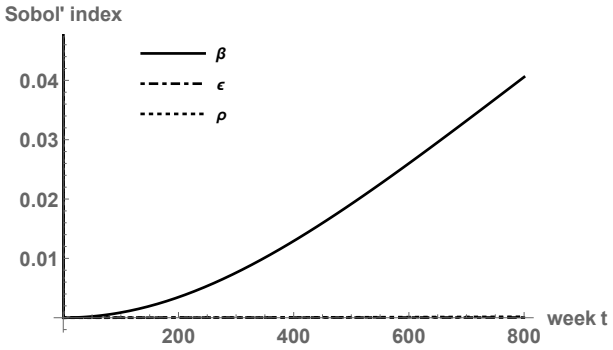

Fig. 5 – Sobol' indices of the transmission parameters β (solid line), ϵ (dotted dashed line) and ρ (dotted line).

6. CONCLUSIONS

In this paper, we have shown how the maximum entropy principle and gPC expansions can be combined to address the problem of uncertainty quantification for the dynamics of an epidemic. The transmission parameters are considered as independent random variables, whose probability distributions are inferred by maximizing the Shannon entropy measure subject to constraints. Once this step is completed, the stochastic solution is expanded in the mean square sense in terms of orthogonal polynomials, to derive fast and accurate approximations of the mean and the variance and to conduct variance-based sensitivity analyses efficiently. The methodology has been applied to a model for the growth of adulthood obesity in the region of Valencia, Spain, based on a compartmental system of ordinary differential equations. To the best of our knowledge, this work is one of the first contributions to the mathematical modeling of epidemics in which the maximum entropy method and gPC expansions have been combined for uncertainty quantification.

REFERENCES

1. F. Brauer, in *Mathematical Epidemiology*, pp. 19–79 (Springer, 2008).
2. Z. Ma, J. Li, *Dynamical Modeling and Analysis of Epidemics* (World Scientific, 2009).
3. L. Acedo, J.-A. Moraño, F.-J. Santonja, R.-J. Villanueva, Physica A: Statistical Mechanics and its Applications **450**, 278–286, DOI:10.1016/j.physa.2015.12.153 (2016).
4. L. Acedo, J. Díez-Domingo, J.-A. Moraño, R.-J. Villanueva, Epidemiology & Infection **138**(6), 853–860, DOI:10.1017/S0950268809991373 (2010).
5. F. Udwadia, SIAM Review **31**(1), 103–109, DOI:10.1137/1031004 (1989).
6. F. Dorini, R. Sampaio, Journal of Applied Mechanics **79**(5), DOI:10.1115/1.4006453 (2012).
7. D. Xiu, *Numerical Methods for Stochastic Computations: A Spectral Method Approach* (Princeton University Press, 2010).

8. D. Xiu, Communications in Computational Physics **2**(2), 293–309 (2007).
9. G. Fishman, *Monte Carlo: Concepts, Algorithms, and Applications* (Springer Science & Business Media, 2013).
10. F.-J. Santonja, R.-J. Villanueva, L. Jódar, G. González-Parra, Mathematical and Computer Modelling of Dynamical Systems **16**(1), 23–34, DOI:10.1080/13873951003590149 (2010).
11. P. James, R. Leach, E. Kalamara, M. Shayeghi, Obesity Research **9**(S11), 228S–233S (2001).
12. E. prospectivo Delphi, Madrid: Gabinete de estudios Bernard Krief (1999).
13. N. Christakis, J. Fowler, New England Journal of Medicine **357**(4), 370–379 (2007).
14. D. Blanchflower, B. Van Landeghem, A. Oswald, Journal of the European Economic Association **7**(2-3), 528–538 (2009).
15. I. Sobol', Mathematics and Computers in Simulation **55**(1-3), 271–280 (2001).
16. A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, S. Tarantola, *Global Sensitivity Analysis: The Primer* (John Wiley & Sons, 2008).