J Nonlinear Sci (2018) 28:371–393
https://doi.org/10.1007/s00332-017-9411-4

Journal of Nonlinear Science

The Stochastic Quasi-chemical Model for Bacterial Growth: Variational Bayesian Parameter Update

Panagiotis Tsilifis[1] · William J. Browning[2] · Thomas E. Wood[2] · Paul K. Newton[3] · Roger G. Ghanem[1]

Received: 4 April 2017 / Accepted: 11 August 2017 / Published online: 24 August 2017

Abstract We develop Bayesian methodologies for constructing and estimating a stochastic quasi-chemical model (QCM) for bacterial growth. The deterministic QCM, described as a nonlinear system of ODEs, is treated as a dynamical system with random parameters, and a variational approach is used to approximate their probability distributions and explore the propagation of uncertainty through the model. The approach consists of approximating the parameters' posterior distribution by a probability measure chosen from a parametric family, through minimization of their Kullback–Leibler divergence.

Communicated by Charles R. Doering.

✉ Paul K. Newton
newton@usc.edu

Panagiotis Tsilifis
tsilifis@usc.edu

William J. Browning
wjbrowning@applmath.com

Thomas E. Wood
tewood@applmath.com

Roger G. Ghanem
ghanem@usc.edu

1 Department of Civil Engineering, University of Southern California, Los Angeles, CA 90089, USA

2 Applied Mathematics Inc., 1622 Route 12, Gales Ferry, CT 06335, USA

3 Department of Aerospace and Mechanical Engineering and Mathematics, University of Southern California, Los Angeles, CA 90089, USA

Keywords Bayes rule · Kullback–Leibler divergence · Evidence lower bound · Quasi-chemical model · Gradient-based optimization

1 Introduction

The mathematical modeling of bacterial growth has been an important analytical tool for the food microbiology community due to its potential for achieving cost savings in food product development as well as food safety and hazard analysis. Its mathematical characterization has generally dealt with coarse-scale observables of population growth rate (Gompertz 1825; Schnute 1981; Baranyi et al. 1993; Baranyi and Roberts 1994; Ricker 1979; Buchanan 1922; Buchanan et al. 1997; Whiting 1993; McMeekin et al. 1997). This approach is natural, but has two essential limitations. First, it is primarily suitable for describing the growth kinetics of a population until an asymptotic maximum is reached and is not equipped to describe the decline and death of bacteria. As a result, isolating growth from inactivation kinetics and using an additional distinct model for the death phase (Whiting et al. 1996) become a necessary additional step. Second, these models are not typically informed by underlying chemical reactions and cannot, therefore, provide insight into the chemical mechanisms involved in the dynamic processes under investigation. This is important for understanding and mitigating safety hazards and associated risks. The quasi-chemical model (QCM) (Doona et al. 2005; Taub et al. 2003) that was developed initially for the study of the growth/death kinetics of *Staphylococcus aureus* in bread (Taub et al. 2000) manages to fully address the first drawback mentioned above by considering that the bacteria concentrations satisfy an autonomous nonlinear dynamical system that can successfully capture the death dynamics of the population for certain parameter values and partially addresses the second issue by introducing an intermediate antagonistic metabolite that acts as an intercellular signaling molecule, referred to as *quorum sensing* (Goldbeter 1997). The model has shown promising potential in fitting microbial data obtained from cultures grown in a wide range of environmental conditions of water activity (Aw), pH and temperature and adapts well to several variations of these intrinsic parameters as well as to additional pathogenic microorganisms apart from S. aureus.

Typically the model calibration process of the QCM (and other bacteria models) is performed using ordinary nonlinear least squares. Despite the reasonable results obtained by point estimation, it can be argued that the deterministic setting has certain limitations. Although the usual t- and F-tests have been used in nonlinear models for hypothesis testing and construction of confidence intervals (Zwietering et al. 1990), their accuracy is limited as they are theoretically valid only for the linear case. It is clear that such prediction intervals for the model parameters and the bacteria population would greatly benefit from a fully stochastic setting that allows exploring the probability distributions of the parameters either within a frequentist (Banks and Bihari 2001) or Bayesian framework (Tarantola 2005). Another benefit of the probabilistic approach is that it allows for various interpretations of the induced model uncertainty in ways that can uncover non-obvious dependencies between the parameters or even

a hierarchical structure between the model parameters and possible hidden variables that can be thought of as hyper-parameters.

This is the approach followed in this paper. The first step is to treat the model parameters as random variables and identify their probability distributions, taking into account available observations of the system solution. We consider a Bayesian formalism of the calibration problem on the QCM that allows exploration of the posterior distribution of the parameters conditioned on observations of the population concentrations at discrete time instances. We apply a variational approach that was first introduced in Gershman et al. (2012) that provides efficient exploration of the posterior distribution by approximating it with a distribution from a previously defined parametric family and using gradient information of the system solution. Similar approaches have been developed also in the context of stochastic differential equations (SDEs) (Vrettas et al. 2011) and more generally on arbitrary stochastic processes in discrete and continuous time (Ye et al. 2015) where it is of interest to estimate posterior distribution over possible states given discrete panel data. One of the main advantages of the method is that, unlike Markov chain Monte Carlo methods (MCMC) (Robert and Casella 2013) or other variational approaches (Pinski et al. 2015), the number of forward evaluations required to solve the proposed optimization problem is significantly lower (see also Tsilifis et al. 2016 for a comparison with MCMC). In addition, unlike the above-mentioned works on variational approaches for SDEs, our method consists of developing schemes particularly for approximating the posterior with a Gaussian mixture, thus providing the freedom to choose a suitable number of components in order to capture complex distributions with possible multimodal structure, and has been shown to perform effectively on such cases (Chen et al. 2015). This essentially addresses the issue of ill-posedness of the inverse problem and the non-uniqueness of its solution which is present in our case where several datasets either consist of only a small number of observations or exhibit significant fluctuations (see Stuart 2010 for a detailed review of advantages of the Bayesian approach). Our Bayesian approach differs from previous probabilistic methods that have been applied to the QCM (Doona et al. 2012) where deterministic calibration was performed using as data Monte Carlo samples synthesized as model outputs with additive noise in order to investigate the corresponding sensitivity of the parameter estimates. Thus in contradistinction to this latter approach where a probability is imposed on the observed data, Bayesian approaches, including ours, postulate a prior model on the parameters, which is then shaped by the inference procedure. The next step is to enrich the complexity of the model by incorporating time-dependent white noise in the model parameters. We think of this new representation as a way to capture possible fluctuations in the observations that cannot be explained through additive measurement noise. This formulation results in deriving the stochastic differential equation (SDE) analogue of the QCM which is presented in a separate publication.

The paper is structured as follows. First we introduce the QCM and its main characteristics. We then present the variational approach used for the Bayesian inversion. The method is then demonstrated on various datasets previously presented in Taub et al. (2003) that consist of bacteria counts grown under different environmental conditions.

2 The Quasi-chemical Model

The quasi-chemical model Doona et al. (2005) and Taub et al. (2003) was introduced in order to describe growth of bacterial populations in food and is based on a hypothetical four-step kinetic reaction scheme deriving the following system of nonlinear ordinary differential equations,

$$\begin{aligned}
\frac{d[M]}{\mathrm{d}t} &= -k_1[M] \\
\frac{d[M^*]}{\mathrm{d}t} &= k_1[M] + (k_2 - k_4)[M^*] - hk_3[M^*][A] \\
\frac{d[A]}{\mathrm{d}t} &= k_2[M^*] - hk_3[M^*][A] \\
\frac{d[D]}{\mathrm{d}t} &= k_4[M^*] + hk_3[M^*][A],
\end{aligned} \tag{1}$$

where $[M]$, $[M^*]$, $[A]$ and $[D]$ are the concentrations of lag-phase cells, exponential growth phase cells, an antagonistic metabolite chemical and dead cells, respectively, $k_1, \ldots, k_4$ are nonnegative parameters and h is a scaling factor, here taken as $h = 10^{-9}$. In bacterial growth simulations, the initial conditions ($t = 0$) always admit a value I ($\approx 10^3 - 10^4$) for $[M]$ which is the inoculum level, while the rest are 0.

The main mathematical properties of (1) can be found in Ross et al. (2005). Specifically, under the parameter constraints $k_i \geq 0$ and the initial conditions that we consider here to be always $(I, 0, 0, 0)$, the system has only one critical point at $[M] = [M^*] = 0$ which is attained only as $t \to \infty$. Taking also into account the relationship

$$\frac{d[A]}{\mathrm{d}t} = k_2[M^*]e^{-\phi(t)}, \tag{2}$$

where $\phi(t) = hk_3 \int_0^t [M^*](s)\mathrm{d}s$, the critical point values of $[M]$, $[M^*]$ imply that $[A] = 0$ and subsequently $[D] = 0$. Equation (2) also suggests that $\frac{d[A]}{\mathrm{d}t} \geq 0$ and therefore $[A]$ never decreases. Moreover, we have

$$[A] = \frac{k_2}{hk_3}\left[1 - e^{-\phi(t)}\right] \leq \frac{k_2}{hk_3}, \tag{3}$$

which gives an upper bound for the concentration $[A]$ of the antagonist.

3 Variational Bayesian Inference

3.1 Bayesian Inference

We are interested in inferring the unknown parameters $\boldsymbol{\kappa} = (k_1, \ldots, k_4)$ of system (1) given experimental data that consist of discrete observations of the logarithm of the number of living cells of the population

$$U(t) = [M](t) + [M^*](t). \tag{4}$$

The true values are subject to additional random measurement noise ϵ_i at time t_i; thus, we assume that our observations are the realizations

$$y_i = \log_{10} U(t_i, \boldsymbol{\kappa}) + \epsilon_i, \qquad i = 1, \ldots, N \tag{5}$$

where we highlight the dependence of $\log_{10} U(t)$ on the uncertain parameters $\boldsymbol{\kappa}$ and we denote with $\mathbf{y} = (y_1, \ldots, y_N)^T$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_N)^T$ the set of all observations and the additive noises, respectively, in vector form. The additive noise ϵ_i for each observation will be taken here to be independent and identically distributed Gaussian random variables that are independent of the parameters $\boldsymbol{\kappa}$, that is, $\epsilon_i \sim \mathcal{N}(0, \sigma)$ or in vector form $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma \mathbf{I}_N)$ and $p(\boldsymbol{\kappa}, \sigma) = p(\boldsymbol{\kappa}) p(\sigma)$. Throughout our numerical examples below, it will be convenient to allow the parameters taking values over the real line and for that we will further use the parameterization $\boldsymbol{\xi} = \log \boldsymbol{\kappa}$ and $\omega = \frac{1}{2} \log \sigma$ and (with no loss of generality) rewrite in compact form

$$\mathbf{y} = \mathcal{G}(\boldsymbol{\xi}) + \boldsymbol{\epsilon}, \tag{6}$$

where

$$\mathcal{G}(\boldsymbol{\xi}) = \left(\log_{10} U(t_1, \boldsymbol{\xi}), \ldots, \log_{10} U(t_N, \boldsymbol{\xi})\right)^T. \tag{7}$$

The Bayesian paradigm consists of considering that $\boldsymbol{\xi}$ follows a prior distribution $p(\boldsymbol{\xi})$ which after collecting the observations $\mathbf{y}$ is updated using Bayes' rule to

$$p(\boldsymbol{\xi}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\xi}) p(\boldsymbol{\xi})}{p(\mathbf{y})} = \frac{p(\mathbf{y}|\boldsymbol{\xi}) p(\boldsymbol{\xi})}{\int p(\mathbf{y}|\boldsymbol{\xi}) p(\boldsymbol{\xi}) \mathrm{d}\boldsymbol{\xi}}. \tag{8}$$

To allow for inference on the Gaussian measurement noise, we can also write

$$p(\boldsymbol{\xi}, \omega|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\xi}, \omega) p(\boldsymbol{\xi}) p(\omega)}{p(\mathbf{y})} = \frac{p(\mathbf{y}|\boldsymbol{\xi}, \omega) p(\boldsymbol{\xi}) p(\omega)}{\int\int p(\mathbf{y}|\boldsymbol{\xi}, \omega) p(\boldsymbol{\xi}) p(\omega) \mathrm{d}\boldsymbol{\xi} \mathrm{d}\omega}, \tag{9}$$

and denote with $\boldsymbol{\theta} \in \mathbb{R}^k$ the augmented set of parameters $\boldsymbol{\theta} = (\boldsymbol{\xi}, \omega)$.

Typically, the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$ is explored via Markov chain Monte Carlo (MCMC) techniques (Robert and Casella 2013) which enable sampling from a Markov chain that has $p(\boldsymbol{\theta}|\mathbf{y})$ as its invariant distribution, and therefore, the generated samples can be thought of as (asymptotically) drawn from that distribution. In order for that assumption to be accurate, it is often required to draw thousands of samples until the generated chain approaches its invariant density. The speed of convergence of the chain depends on the nature of the true posterior and on the generating procedure adapted. For instance, several variations of the Metropolis–Hastings (Metropolis et al. 1953; Hastings 1970; Haario et al. 2001; Roberts and Rosenthal 1998) or even Metropolis-within-Gibbs (Chaspari et al. 2016) algorithm exist in the literature which provide different acceptance rates and eventually faster or slower convergence. However, in all but the simplest cases, several thousands of samples are required, which

means that the forward model must run equally many times, resulting in large or even unaffordable computational costs. In order to avoid excessive simulations and reduce the computational cost of exploring the posterior distribution, we consider here a variational formulation of the problem that has previously appeared in Gershman et al. (2012) and Tsilifis et al. (2016) and is based on approximating $p(\boldsymbol{\theta}|\mathbf{y})$ by a known probability distribution chosen from a known family.

Let $\mathcal{Q}$ be a family of probability densities, parameterized by $\boldsymbol{\lambda}$, that is, $\mathcal{Q} = \{q(\boldsymbol{\theta}|\boldsymbol{\lambda}),\ \boldsymbol{\lambda} \in \mathbb{R}^d\}$. We are interested in identifying the element of $\mathcal{Q}$ that is as "close" as possible to the true posterior $p(\boldsymbol{\theta}|\mathbf{y})$ defined above. To quantify the "distance" between two probability distributions or their densities, we employ the Kullback–Leibler (KL) divergence Kullback and Leibler (1951)

$$\mathrm{KL}\left[q(\boldsymbol{\theta}|\boldsymbol{\lambda})||p(\boldsymbol{\theta}|\mathbf{y})\right] = \int_{\mathbb{R}^k} \log\left[\frac{q(\boldsymbol{\theta}|\boldsymbol{\lambda})}{p(\boldsymbol{\theta}|\mathbf{y})}\right] q(\boldsymbol{\theta}|\boldsymbol{\lambda}) d\boldsymbol{\theta}. \tag{10}$$

The KL divergence does not define a metric since it does not satisfy the triangular inequality; however, it is often used in related context since it is always nonnegative and $\mathrm{KL}\,[q||p] = 0 \iff q = p$ a.s. In addition, $\mathrm{KL}\,[q||p] \to 0$ implies that $q \to p$ in total variation, a result that guarantees proximity between two probability measures. Intuitively here, the KL divergence can be thought of as the "information loss" of approximating the true posterior $p(\boldsymbol{\theta}|\mathbf{y})$ by $q(\boldsymbol{\theta}|\boldsymbol{\lambda})$. Similarly, a KL divergence between the prior and the posterior distributions can be thought of as the "information gain" of collecting data $\mathbf{y}$ and has been used extensively for experimental design purposes (Chaloner and Verdinelli 1995; Tsilifis et al. 2017). It is clear from the above that minimizing the KL divergence provides a constraint on approximations of the true posterior, and therefore, we can state the optimization problem of finding $\boldsymbol{\lambda}^*$ such that

$$q(\boldsymbol{\theta}|\boldsymbol{\lambda}^*) = \arg\min_{\boldsymbol{\lambda}\in\mathbb{R}^d} \mathrm{KL}\left[q(\boldsymbol{\theta}|\boldsymbol{\lambda})||p(\boldsymbol{\theta}|\mathbf{y})\right]. \tag{11}$$

Trivially, if the true posterior $p(\boldsymbol{\theta}|\mathbf{y})$ is in $\mathcal{Q}$ then one should expect the result of the optimization to be 0; otherwise, it will be a strictly positive value. In practice, it is not possible to evaluate the above expression since $p(\boldsymbol{\theta}|\mathbf{y})$ is not known. For that, we formulate in the next subsection an equivalent optimization problem that is possible to solve, by deriving a lower bound of the evidence $p(\mathbf{y})$.

3.2 The Evidence Lower Bound (ELBO)

First we observe that by substituting Bayes' rule (9) in (10) we can write the relation

$$\log p(\mathbf{y}) = \mathcal{F}[q] + \mathrm{KL}\left[q(\boldsymbol{\theta}|\boldsymbol{\lambda})||p(\boldsymbol{\theta}|\mathbf{y})\right], \tag{12}$$

where

$$\begin{aligned}\mathcal{F}[q] &= \int \log\left[\frac{p(\mathbf{y},\boldsymbol{\theta})}{q(\boldsymbol{\theta}|\boldsymbol{\lambda})}\right] q(\boldsymbol{\theta}|\boldsymbol{\lambda}) d\boldsymbol{\theta} \\ &= \mathcal{H}[q] + \int \log p(\mathbf{y},\boldsymbol{\theta}) q(\boldsymbol{\theta}|\boldsymbol{\lambda}) d\boldsymbol{\theta},\end{aligned} \tag{13}$$

with

$$\mathcal{H}[q] = -\int q(\boldsymbol{\theta}|\boldsymbol{\lambda}) \log q(\boldsymbol{\theta}|\boldsymbol{\lambda}) d\boldsymbol{\theta} \tag{14}$$

being the entropy of $q(\boldsymbol{\theta}|\boldsymbol{\lambda})$. Observing that the left-hand side of (12) is independent of $q(\boldsymbol{\theta}|\boldsymbol{\lambda})$, we see that minimizing $KL[q||p]$ is equivalent to maximizing the quantity $\mathcal{F}[q]$ which due to positivity of $KL[q||p]$ always satisfies

$$\mathcal{F}[q] \leq \log p(\mathbf{y}). \tag{15}$$

We refer $\mathcal{F}[q]$ as *evidence lower bound (ELBO)*. Maximization of $\mathcal{F}[q]$ is a feasible task since it does not depend on the posterior or the evidence distributions but only on the joint distribution $p(\mathbf{y}, \boldsymbol{\theta}) = p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ which in general is known.

3.3 Approximating Schemes

In order to evaluate the integrals in (13), we need to employ some numerical method. While numerical integration or Monte Carlo sampling provides accurate sampling strategies, when used within an optimization algorithm, however, they quickly become computationally prohibitive. To address this issue, we develop approximating schemes for (13) that are valid when the approximating family of distributions $\mathcal{Q}$ is the family of Gaussian mixtures. Specifically, we assume that

$$\mathcal{Q}_L = \left\{ q(\boldsymbol{\theta}|\boldsymbol{\lambda}) \,\middle|\, q(\boldsymbol{\theta}|\boldsymbol{\lambda}) = \frac{1}{L} \sum_{i=1} \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \right\}. \tag{16}$$

Here the optimization problem needs to be solved with respect to the parameters $\boldsymbol{\lambda} = \{\boldsymbol{\mu}_i\}_{i=1}^L \cup \{\boldsymbol{\Sigma}_i\}_{i=1}^L$, where L, the number of Gaussian components, will be assumed to be fixed and of course $\boldsymbol{\Sigma}_i$ must be restricted to be positive definite. In fact, for the purpose of inferring the parameters of the QCM in our numerical implementations below, $\boldsymbol{\Sigma}$ will also be assumed to be diagonal which will further simplify the calculations. The latter implies that the parameters will be *a posteriori* independent, which is not always the case, but we put this constraint here for the sake of simplicity.

First we derive a lower bound for $\mathcal{H}$ by using Jensen's inequality (see Huber et al. 2008 for details) to obtain

$$\mathcal{H}[q] \geq \mathcal{H}_0[q] \tag{17}$$

where

$$\mathcal{H}_0[q] = -\frac{1}{L} \sum_{i=1}^{L} \ln q_i \tag{18}$$

with

$$q_i = \frac{1}{L} \sum_{j=1}^{L} \mathcal{N}(\boldsymbol{\mu}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j) \tag{19}$$

and by replacing $\mathcal{H}[q]$ with $\mathcal{H}_0[q]$ in Eq. (13) we derive a lower bound for the ELBO. Next, we define $\mathcal{L}[q]$ to be the second term in (13),

$$\mathcal{L}[q] = \int \log p(\mathbf{y}, \boldsymbol{\theta}) q(\boldsymbol{\theta}|\boldsymbol{\lambda}) d\boldsymbol{\theta} \tag{20}$$

which by substituting $q(\boldsymbol{\theta}|\boldsymbol{\lambda}) = \frac{1}{L}\sum_{i=1}^{L} \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ gives

$$\mathcal{L}[q] = \frac{1}{L}\sum_{i=1}^{L} \int \log p(\mathbf{y}, \boldsymbol{\theta}) \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) d\boldsymbol{\theta} \tag{21}$$

and each integral term can be approximated by taking a second-order Taylor expansion of $\log p(\mathbf{y}, \boldsymbol{\theta})$ about $\boldsymbol{\theta} = \boldsymbol{\mu}_i$. This gives

$$\begin{aligned} \log p(\mathbf{y}, \boldsymbol{\theta}) \approx \log p(\mathbf{y}, \boldsymbol{\mu}_i) + \nabla_{\boldsymbol{\theta}} \log p(\mathbf{y}, \boldsymbol{\mu}_i)(\boldsymbol{\theta} - \boldsymbol{\mu}_i) \\ + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}_i)^T \nabla^2_{\boldsymbol{\theta}} \log p(\mathbf{y}, \boldsymbol{\mu}_i)(\boldsymbol{\theta} - \boldsymbol{\mu}_i), \end{aligned} \tag{22}$$

where $\nabla_{\boldsymbol{\theta}}$ and $\nabla^2_{\boldsymbol{\theta}}$ denotes the Jacobian and Hessian with respect to $\boldsymbol{\theta}$. Upon substitution of (22) into (20) and observing that the expectation of the first-order term (Jacobian) vanishes, we take the approximation

$$\mathcal{L}_2[q] = \mathcal{L}_0[q] + \frac{1}{2L}\sum_{i=1}^{L} \mathrm{Tr}\left[\boldsymbol{\Sigma}_i \nabla^2_{\boldsymbol{\theta}} \log p(\mathbf{y}, \boldsymbol{\mu}_i)\right], \tag{23}$$

where $\mathcal{L}_0[q]$ is the term resulting from the zeroth-order approximation in (22)

$$\mathcal{L}_0[q] = \frac{1}{L}\sum_{i=1}^{L} \log p(\mathbf{y}, \boldsymbol{\mu}_i). \tag{24}$$

This approximation is also referred to as the multivariate delta method for moments (Bickel and Doksum 2015). At last, Eqs. (18) and (23), when combined, give us an approximation of the ELBO

$$\mathcal{F}_2[q] = \mathcal{H}_0[q] + \mathcal{L}_2[q] \tag{25}$$

that can be used in the optimization algorithm to compute a local maximum of (13).

The quality of the above approximation can be intuitively assessed by observing first that the $\mathcal{L}_0[q]$ term essentially encourages placing the components of the mixture distribution in areas where data were observed, thereby making the observations "highly probable", while the entropy term $\mathcal{H}_0[q]$ penalizes those locations. At last, the second term of $\mathcal{L}_2[q]$ captures the local curvature (Hessian) of the posterior that attempts to maximize the volume (broaden the variance $\boldsymbol{\Sigma}_i$) around samples with high probability by extending the tails (areas with high curvature—highly concave density

function). Overall, assuming that a proper number of components in the Gaussian mixture posterior are chosen, such that a possibly broad and multimodal posterior can be accurately described, the Taylor approximation on the log-joint density function contains all the information we typically need: the centers of high probably data and the curvature (tail) behavior.

Before using $\mathcal{F}_2[q]$ in our optimization algorithm, it is necessary to clarify some computational details. First, $\mathcal{L}_2[q]$ requires knowledge of the Hessian of $\log p(\mathbf{y}, \boldsymbol{\mu}_i)$. Therefore, when $\mathcal{F}_2[q]$ is being optimized with respect to $\boldsymbol{\mu}_i$ by using a gradient-based optimization algorithm, at least the third-order derivatives of $\log p(\mathbf{y}, \boldsymbol{\mu}_i)$ will need to be known, or in other words, the third-order derivatives of $\mathcal{G}(\boldsymbol{\kappa})$ defined in (6). To avoid the computational burden of computing third-order derivatives, which in a general setting might not even be feasible, we only use the approximation $\mathcal{L}_0[q]$ when optimizing with respect to $\boldsymbol{\mu}_i$. This way only the first-order derivatives of our forward model are used. Although this might seem as a poor approximation, in practice it achieves the main goal which is to place the mixture components at areas of high probability. On the other hand, when optimizing with respect to $\boldsymbol{\Sigma}_i$, as indicated by the form of (23), the dependence on $\boldsymbol{\Sigma}_i$ is only through the second term. That requires the second-order derivatives of $\mathcal{G}(\boldsymbol{\kappa})$ which, although non-trivial, is comparable to computing the gradient (taking into account that the trace of the product of two matrices in the expression requires only the diagonal elements); therefore, it is typically feasible and the full approximation $\mathcal{L}_2[q]$ can be used. Taking this into account, we adopt an optimization algorithm that interchanges the use of $\mathcal{L}_0[q]$ and $\mathcal{L}_2[q]$ when successively optimizing with respect to $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$, respectively, until a certain tolerance level is achieved. Computation of $\mathcal{H}_0[q]$ and its derivatives is straightforward. Details of the computation of $\log p(\mathbf{y}, \boldsymbol{\theta})$ along with its first- and second-order derivatives are provided in "Appendix A."

At last, we note that we have restricted ourselves in estimating the posterior with mixtures of equally weighted Gaussian kernels. One could easily extend the above derivations to Gaussian mixtures with arbitrary weights, that is,

$$q(\boldsymbol{\theta}|\boldsymbol{\lambda}) = \sum_{i=1} w_i \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{26}$$

with $\sum_{i=1}^{L} w_i = 1$. Although such a generalization would seemingly allow for further flexibility on successfully capturing the true posterior, similar results can be achieved with our current setting after observing that any weight w_i can be estimated by a rational number $w_i \approx l_i/L = \sum_{j=1}^{l_i} 1/L$; therefore, the corresponding weighted Gaussian term in the mixture in fact corresponds to a sum of equally weighted Gaussians. This resembles the core idea behind Gaussian kernel density estimators (Silverman 1986). A trade-off in terms of computational efficiency is also present here as in our setting, a larger number L will be required for the posterior approximation which will result in a high-dimensional optimization problem. On the other hand, the arbitrarily weighted mixtures require additional computation of gradients of the ELBO with respect to the weights. Detailed comparison of the performance of the two parametric families falls beyond the scope of our work.

4 Numerical Results

In our numerical simulations below, we are interested in inferring $\boldsymbol{\theta} = (\log \boldsymbol{\kappa}, \omega)$ based on available experimental data that consist of the kinetics profile $U(t)$ (CFU/mL) as a function of time over intervals ranging from [0–7 days] up to [0–70 days]. The various datasets correspond to combinations of different water activity (Aw), pH and temperature T as well as small variations in the initial conditions which result in different growth curves and therefore different sets of parameter values. First we consider the case where the posterior is approximated by single Gaussians. That is, we choose $L = 1$ as the number of components in the elements of the family of Gaussian mixtures and later we explore the case where $L > 1$ for a particular reference dataset at $T = 25\,^{\circ}\mathrm{C}$. Furthermore, in order to reduce the dimensionality of the objective function to be optimized, we also take $\boldsymbol{\Sigma}$ to be diagonal and the total dimension of the optimization problem becomes $d = L(\#\boldsymbol{\mu} + \#\,\mathrm{diag}\,\boldsymbol{\Sigma}) = 10L$. Each optimization step is carried out using the L-BFGS-B algorithm (Byrd et al. 1995) that can perform bound constrained optimization.

4.1 The Role of the Prior

At this point, it is important to mention that the solution of the optimization problem described above or in other words, the variational solution of the inverse problem depends on the choice of the prior $p(\boldsymbol{\theta})$. This can be seen directly by expanding the term $p(\mathbf{y}, \boldsymbol{\theta}) = p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ in $\mathcal{F}[q]$. Intuitively, the prior must be "wide" enough so that the mean of the posterior can be detected within the support of the prior with a relatively high probability. In all our computations, we assume a priori that all parameters are independent; therefore, the prior can be factorized as $p(\boldsymbol{\theta}) = \left(\prod_{i=1}^{4} p(\xi_i)\right) p(\omega)$. In all datasets, we have chosen all components of the prior to be Gaussian distributions. Table 1 summarizes the choice of the Gaussian for each parameter and each dataset. The mean values of $\xi_i, i = 1, \ldots, 4$ for Aw = 0.79 and 0.84 are in fact the estimates obtained by fitting the model using nonlinear least squares. The main reason that motivates this choice is that, as was demonstrated in Banks et al. (2016) and observed in our own analysis, these datasets do not sufficiently inform on all parameters. The implications within the present Bayesian formulation is that, for objective functions exhibiting multiple local maxima, a wide prior reflecting limited prior knowledge would steer the algorithm toward one of these local maxima. Intuitively, the prior can be strengthened through deterministic parameter estimation (e.g., least squares), concentrating the prior around a smaller area which biased toward the Bayesian posterior. Specifically, in the case Aw $= 0.79$ it was also observed that a small discrepancy from the mean provided poor estimates of the posterior and it was necessary to further decrease the variance of the prior. In fact, the posterior variances (Table 2) converge to zero and the posteriors become essentially Dirac (point) distributions. The situation for the datasets Aw $= 0.87$ and 0.91 has been less challenging, and a choice of standard normal was proved to be sufficient for all parameters of the model.

Table 1 Prior distributions for each parameter and each dataset

	$\log \kappa_1$	$\log \kappa_2$	$\log \kappa_3$	$\log \kappa_4$	$\log \sigma$
Dataset 1					
Aw = 0.79	$\mathcal{N}(0.039, 0.05)$	$\mathcal{N}(5.47, 0.05)$	$\mathcal{N}(5.75, 0.05)$	$\mathcal{N}(5.47, 0.05)$	$\mathcal{N}(-1, 0.05)$
Aw = 0.84	$\mathcal{N}(-0.99, 1)$	$\mathcal{N}(3.70, 1)$	$\mathcal{N}(5.23, 1)$	$\mathcal{N}(3.69, 1)$	$\mathcal{N}(-1, 1)$
Aw = 0.87	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(-1, 1)$
Aw = 0.91	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(-1, 1)$
Dataset 2					
$T = 15$	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(-1, 1)$
$T = 20$	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(-1, 1)$
$T = 25$	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(-1, 1)$
$T = 30$	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(-1, 1)$
$T = 35$	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(-1, 1)$
$T = 40$	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(-1, 1)$

Table 2 Dataset 1 (pH = 5.4, $T = 35\ ^\circ$C)

	$\log \kappa_1$	$\log \kappa_2$	$\log \kappa_3$	$\log \kappa_4$	SSR
Aw = 0.79	0.0420 ± 0.00	5.4587 ± 0.00	5.7588 ± 0.00	5.4563 ± 0.00	0.0385
(NLS)	0.4762 ± 2.5748	5.7278 ± 13.2534	5.3447 ± 12.8739	5.7260 ± 13.2534	0.0388
Aw = 0.84	-0.9908 ± 0.2219	3.7172 ± 0.0025	5.2166 ± 0.2249	3.7071 ± 0.0025	1.0103
(NLS)	0.9942 ± 0.00	4.5824 ± 9.6803	4.3445 ± 9.4727	4.5773 ± 9.6803	1.0102
Aw = 0.87	-0.6269 ± 0.5471	0.8939 ± 0.0254	1.4446 ± 0.6360	0.3671 ± 0.0421	2.2046
(NLS)	-1.4696 ± 0.00	0.5709 ± 1.1756	2.2721 ± 3.3506	-0.4308 ± 1.2208	2.0800
Aw = 0.91	0.2032 ± 0.6332	1.5118 ± 0.0565	1.2220 ± 0.1977	-0.3174 ± 0.0441	0.6566
(NLS)	0.2390 ± 0.2231	1.5085 ± 0.00	1.2669 ± 0.00	-0.3285 ± 0.00	0.6550

Estimated means and std's of the Gaussian variational posteriors. Right column: sum of squared residuals

4.2 Inversion Using a Single Gaussian

The results of the optimization including the estimated values of the posterior parameters and the squared sum of residuals (SSR) are summarized in Table 2 for the case of varying Aw (and fixed pH $= 5.4$ and $T = 35\,^\circ$C), referred to as Dataset 1 and in Table 3 for the case of varying temperature T (and fixed Aw $= 0.9$ and pH $= 5.23$), referred to as Dataset 2. To allow comparison with a deterministic approach, we also display the parameter values computed using a nonlinear least squares approach (NLS), found in Browning (2016). One can see that in several cases the parameter estimates found for each method have very small discrepancy, while in other cases the values are completely different. However, the SSR values agree for all cases, indicating that the fit is more or less equally good. The disagreement in the parameter values there-

Table 3 Dataset 2 (Aw = 0.9, pH = 5.23)

	$\log \kappa_1$	$\log \kappa_2$	$\log \kappa_3$	$\log \kappa_4$	SSR
$T = 15$	-1.5244 ± 0.4055	-0.0384 ± 0.0092	0.0029 ± 1.9714	0.0449 ± 0.008552	0.2536
(NLS)	-1.8325 ± 3.2756	3.6227 ± 13.4559	-0.4155 ± 9.4254	3.6256 ± 13.4559	0.4610
$T = 20$	-1.9141 ± 0.3840	-0.1968 ± 0.0111	-1.0193 ± 0.3859	-0.4592 ± 0.0145	1.1327
(NLS)	-2.2072 ± 0.00	1.7137 ± 5.4467	-2.9957 ± 0.6729	1.6789 ± 5.4467	0.0115
$T = 25$	-0.4976 ± 0.1344	0.3189 ± 0.0086	0.4005 ± 0.1063	-0.7676 ± 0.0254	0.5371
(NLS)	-0.5447 ± 0.00	0.3001 ± 0.00	0.4574 ± 0.00	-0.8209 ± 0.00	0.4790
$T = 30$	0.0589 ± 0.3456	1.1051 ± 0.0455	0.3679 ± 0.2787	-0.2323 ± 0.1458	1.9249
(NLS)	-0.5621 ± 0.00	1.0919 ± 0.00	1.2029 ± 0.00	-0.6539 ± 0.00	1.34
$T = 35$	0.0989 ± 0.3871	1.4900 ± 0.0755	0.6457 ± 0.3032	-0.0113 ± 0.1348	6.4204
(NLS)	-1.1394 ± 0.00	1.5971 ± 0.00	2.6071 ± 0.00	-0.5447 ± 0.00	1.9100
$T = 40$	-0.3912 ± 0.5439	1.5802 ± 0.0461	2.6079 ± 0.3978	0.6162 ± 0.1344	1.7290
(NLS)	-1.0788 ± 0.00	1.4951 ± 0.00	3.3854 ± 2.8558	0.1133 ± 0.00	1.55

Estimated means and std's of the Gaussian variational posteriors. Right column: sum of squared residuals

fore indicates the existence of multiple solutions which in fact supports the value of a probabilistic approach, that is, to impose a probability distribution on the set of admissible parameter values. Furthermore, as observed in several cases, the standard deviation computed with the NLS method is significantly larger than what is found using the Bayesian approach, especially in the datasets with lower Aw or temperature values that consist of fewer data points and describe essentially only death dynamics (see for instance Aw = 0.79, Aw = 0.84 for Dataset 1 and $T = 15$ °C from Dataset 2). Figures 1 and 2, corresponding to Dataset 1 and Dataset 2, respectively, show the uncertainty in the concentrations as a function of time t depicted with 95% credible intervals and confirm the very good (from at least a qualitative perspective) agreement with the experimental measurements in all cases. The sensitivities and the residuals evaluated at the posterior means of the parameter values are shown in the middle column and are used below for parameter ranking. One interesting characteristic to note here is the fact that some of the estimated values exhibit a relatively large standard deviation, for instance in Dataset 1 $\log \kappa_1$ for Aw = 0.91, $\log \kappa_3$ for Aw = 0.87 and again $\log \kappa_3$ in several cases for Dataset 2. At first, the wide posterior might indicate that the particular parameter's value does not measurably affect the model prediction. This can be verified by inspecting the corresponding sensitivity plot and noting that the curve decays quickly to zero, and therefore, the model output is constant with respect to that particular parameter. The same conclusion of course can be drawn by observing that the confidence bands that we obtain after propagating the uncertainty through the model are still relatively tight in most cases. This feature here serves also as our motivation for exploring posterior approximations with more than one Gaussian component as we see below for a specific dataset. More complex structures of the true posterior are thus revealed. A more detailed analysis of each case separately is beyond the scope of this work.

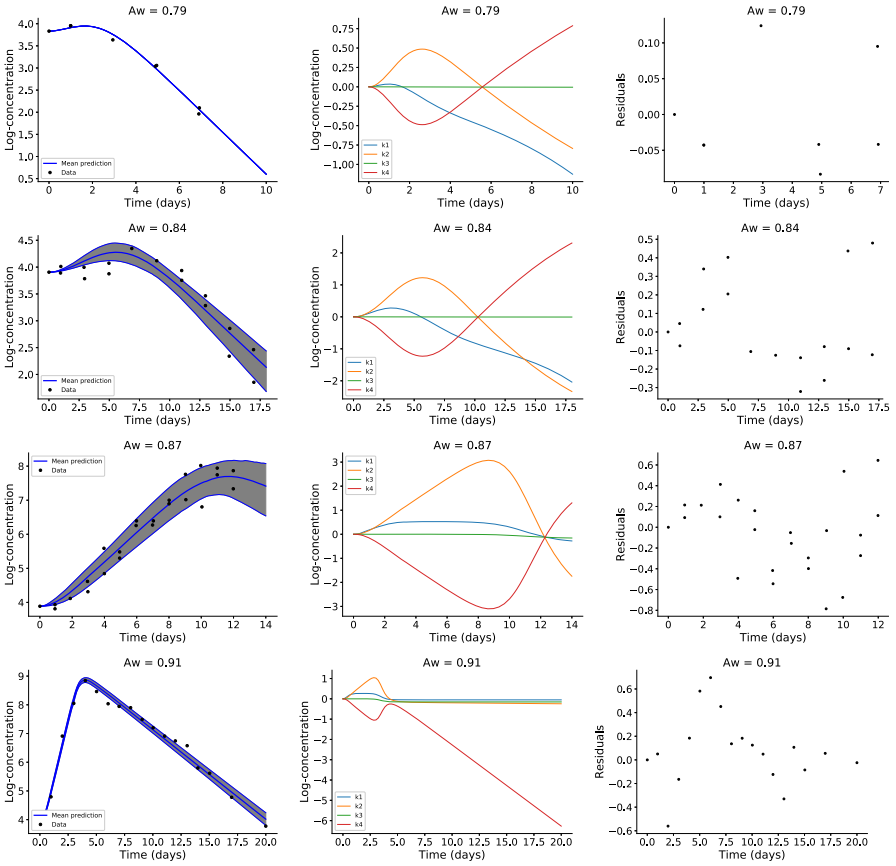


Fig. 1 Dataset 1 (pH = 5.4, $T = 35\,^{\circ}\mathrm{C}$). *Left column* model calibration for each dataset. Fit of the predictive mean and confidence bands. *Middle column* plots of sensitivities evaluated at the posterior mean. *Right column* plot of the residuals

4.3 Parameter Ranking

We next perform a model comparison between the cases where 1, 2, 3 or all 4 parameters are estimated where in each case the remaining parameters are considered known, assuming a fixed value. As was motivated by Banks et al. (2016), we first perform a parameter ranking via a rank-revealing Q-R decomposition of the sensitivity matrix $F(\boldsymbol{\kappa})$, and then, we proceed by estimating the parameters of each model. We perform this procedure specifically for the datasets Aw = 0.79 and 0.84 where it was seen that our variational method required very specific prior knowledge to achieve convergence, indicating that the available dataset does not support accurate parameter estimation. For a comparison of the models, we use two criteria. First, the Bayesian information criterion (BIC) (see Bishop 2006 Ch. 4.4.1), defined as

$$BIC = \log p(\mathbf{y}|\boldsymbol{\kappa}^*) - \frac{1}{2} d \log N \tag{27}$$

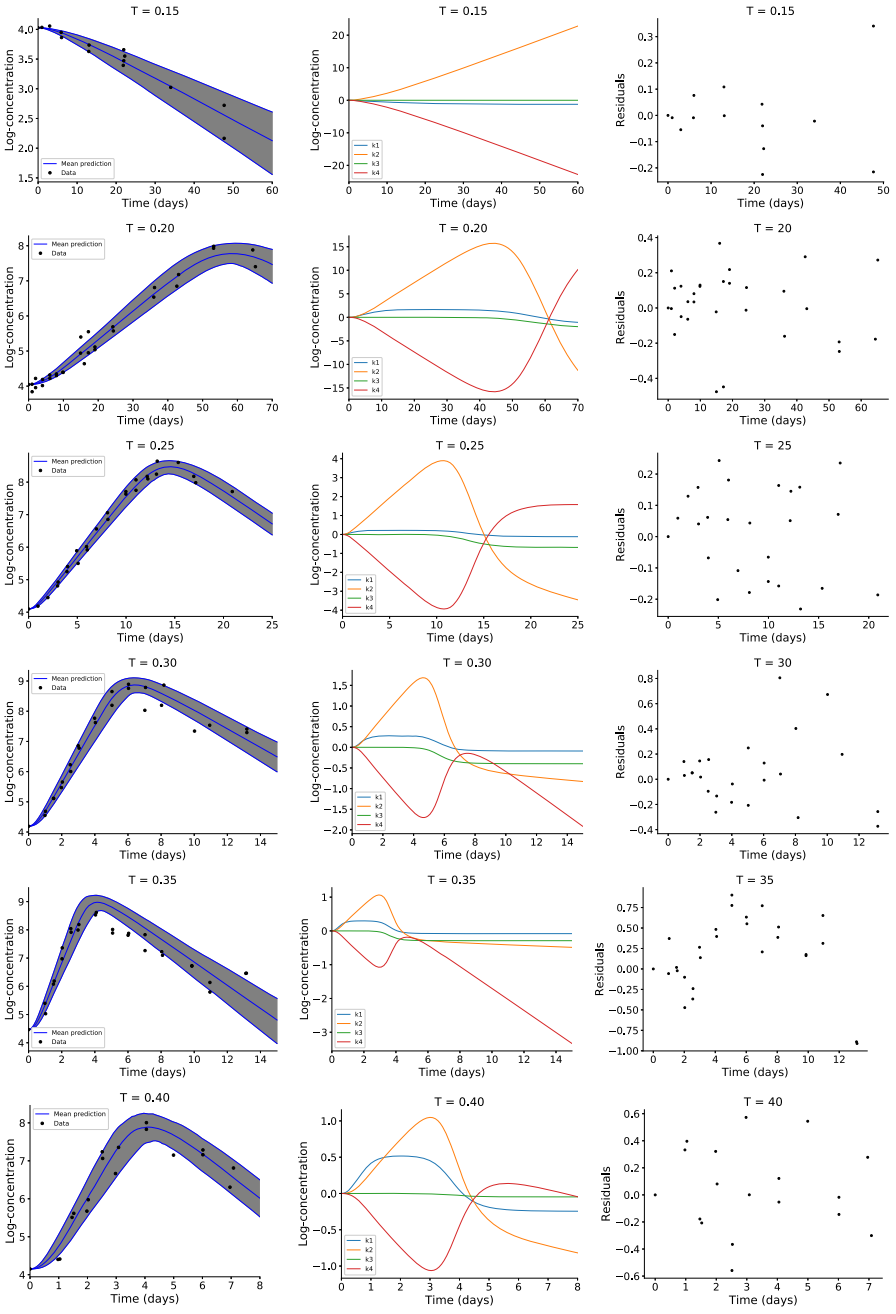


Fig. 2 Dataset 2 (Aw = 0.9, pH = 5.23). *Left column* model calibration for each dataset. Fit of the predictive mean and confidence bands. *Middle column* plots of sensitivities evaluated at the posterior mean. *Right column* plot of the residuals

Table 4 Aw = 0.79

	$\log\kappa_1$	$(\log\kappa_1, \log\kappa_4)$	$(\log\kappa_1,\log\kappa_4,\log\kappa_2)$	$(\log\kappa_1,\log\kappa_4,\log\kappa_2,\log\kappa_3)$
Value	0.0639	(0.0410, 5.4564)	(0.0683,5.4435,5.4457)	(0.0420,5.4563,5.4587,5.7588)
$\mathcal{F}[q]$	−3.5387	−27.4319	−36.6993	−45.9994
BIC	1.0226	0.1850	−0.8449	−2.0355

Table 5 Aw = 0.84

	$\log\kappa_2$	$(\log\kappa_2, \log\kappa_1)$	$(\log\kappa_2,\log\kappa_1,\log\kappa_4)$	$(\log\kappa_2,\log\kappa_1,\log\kappa_4,\log\kappa_3)$
Value	4.9359	(3.7172,−0.9908)	(3.7176,−0.9912,3.7074)	(0.1284,−0.8726,−0.1018, 0.0001)
$\mathcal{F}[q]$	−29.3267	−16.6511	−26.6623	−30.0136
BIC	−25.1637	−4.3785	−5.7933	−7.2117

where $p(\mathbf{y}|\boldsymbol{\kappa})$ is the log likelihood evaluated at the *maximum a posteriori* (MAP) estimate $\boldsymbol{\kappa}^*$, d is the number of estimated parameters and N is the number of observations. Here we clarify that the value for $\boldsymbol{\kappa}^*$ we use in our evaluations is the mean of the variational posterior $q(\boldsymbol{\kappa})$ which is Gaussian; therefore, it is the MAP estimate if we assume that q is the true posterior (or at least a very good approximation). The BIC is known to be valid in cases where $N >> d$. In our case, this is not true since we only have very few observations. In addition to BIC, we present the values of $\mathcal{F}[q]$ as a second criterion for model comparison. This is motivated by the fact that the BIC was essentially proposed as a criterion that serves as an estimate of the evidence $p(\mathbf{y})$. Recall that in our optimization problem, $\mathcal{F}[q]$ is a lower bound of the evidence so it can give us an additional idea about the true value of the evidence. In general, highest values of BIC tend to suggest a preferable model and the sharpest the increase in its value, the strongest the evidence that the model with the largest value is closest to the truth. The computed values are shown in Tables 4 and 5. For the case Aw = 0.79, we observe that the values of both $\mathcal{F}[q]$ and BIC are decreasing as we keep adding parameters in the model and the sharpest decrease occurs between the one- or two-parameter models, thus indicating that the dataset supports accurate estimation of only 1 parameter. For the case where Aw = 0.84, the two-parameter model achieves the largest values for BIC and $\mathcal{F}[q]$, followed by the three-parameter model.

4.4 Inversion Using a Mixture of Gaussians

As mentioned previously, the posterior approximations we have obtained so far are the optimal solutions of our optimization taking into account the constraint we have imposed by fixing the approximating family to be the one consisting of single Gaussian distributions. We are now exploring the case of employing Gaussian mixtures with $L > 1$ components as the approximating family and compare our results with the $L = 1$ case. Recall that we are solving the optimization problem defined in (11) or equivalently

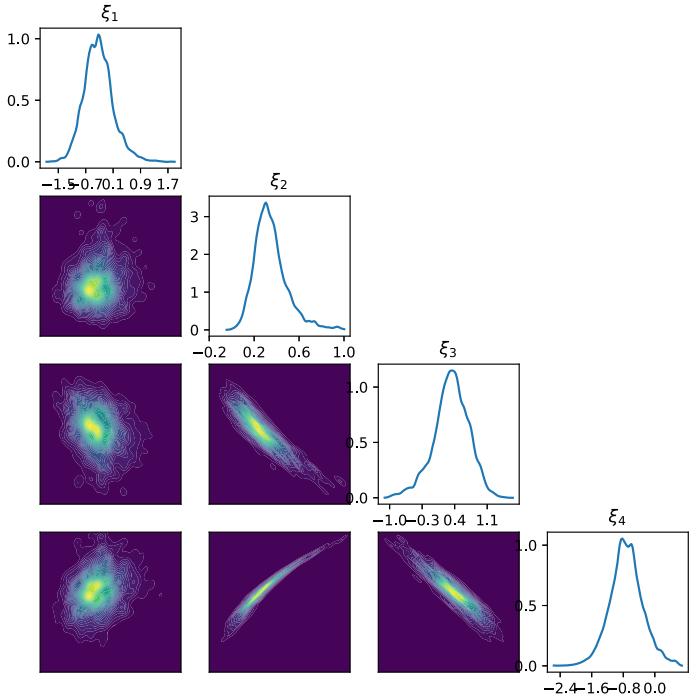


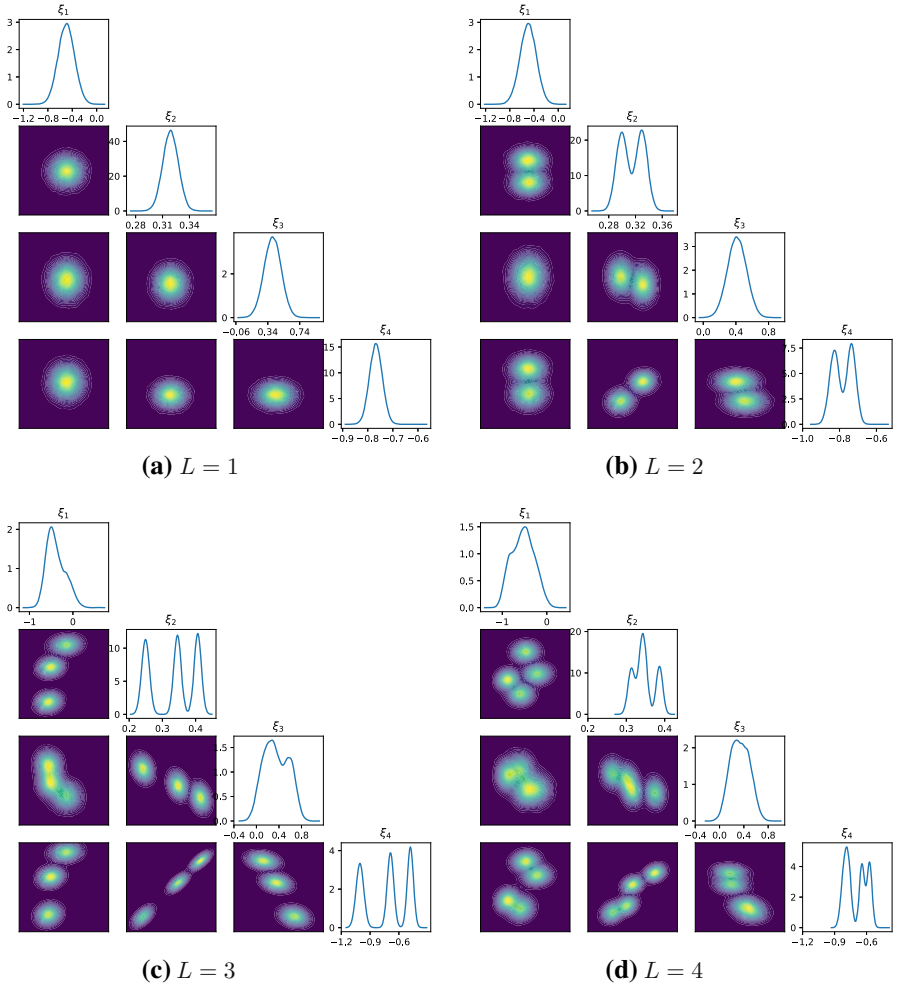
Fig. 3 Empirical marginal and pairwise joint densities corresponding to the MCMC posterior samples

$$q^* = \arg \min_{q \in \mathcal{Q}_L} \mathrm{KL}\,[q||p]\,, \tag{28}$$

where we took $L = 1$. By denoting the above solution q_L^* to indicate that is found in $\mathcal{Q}_L$ and observing that $\mathcal{Q}_L \subset \mathcal{Q}_{L+1}$ for $L \geq 1$, it is obvious that

$$\min_{q \in \mathcal{Q}_L} \mathrm{KL}\,[q||p] \geq \min_{q \in \mathcal{Q}_{L+1}} \mathrm{KL}\,[q||p] \tag{29}$$

which indicates that the solution that can be found in $\mathcal{Q}_{L+1}$ should be preferable over the one in $\mathcal{Q}_L$ if it achieves a smaller KL value (or a larger $\mathcal{F}[q_{L+1}^*]$ value). We solve the optimization problem for $L > 1$ in order to investigate whether a better solution can be found in wider families of Gaussian mixtures and we present the results below. For simplicity, we perform our simulations only on one set of observations, namely the case $\mathrm{Aw} = 0.9$, $\mathrm{pH} = 5.23$ and $T = 25\,^{\circ}\mathrm{C}$ and for $L = 2, 3, 4$. Note that the dimensionality of the optimization problem is $10L$ and therefore increases arithmetically as we increase L. To allow further judgement of the obtained posterior distribution, we have also computed an empirical posterior approximation based on MCMC samples generated using the adaptive Metropolis–Hastings algorithm (Haario et al. 2001). In particular, we generated $1.2 \cdot 10^5$ samples with a 20000-sample burn-in



(a) $L = 1$ **(b)** $L = 2$

(c) $L = 3$ **(d)** $L = 4$

Fig. 4 Empirical marginal densities of the ξ_i's and the empirical 2-d joint densities for $L = 1, 2, 3, 4$

period and a thinning strategy that keeps 1 out of 5 samples resulting in $2 \cdot 10^4$ samples. The implementation was carried out using the python package PyMC[1].

The empirical marginal and pairwise joint posteriors for MCMC are shown in Fig. 3 while Fig. 4 shows the densities of the approximating posteriors found using the variational approach with $L > 1$ along with the one that was found in the previous subsection ($L = 1$). One can observe that by increasing the components of the mixture, the structure of the true posterior is slowly revealed, verifying that the $L = 1$ case, even though satisfactory, might not by an accurate approximation in many of the cases. Overall we observe that the MCMC posteriors appear to be broader than the Gaussian mixtures. This can indicated that more Gaussian terms are necessary to accurately

[1] https://github.com/pymc-devs/pymc.

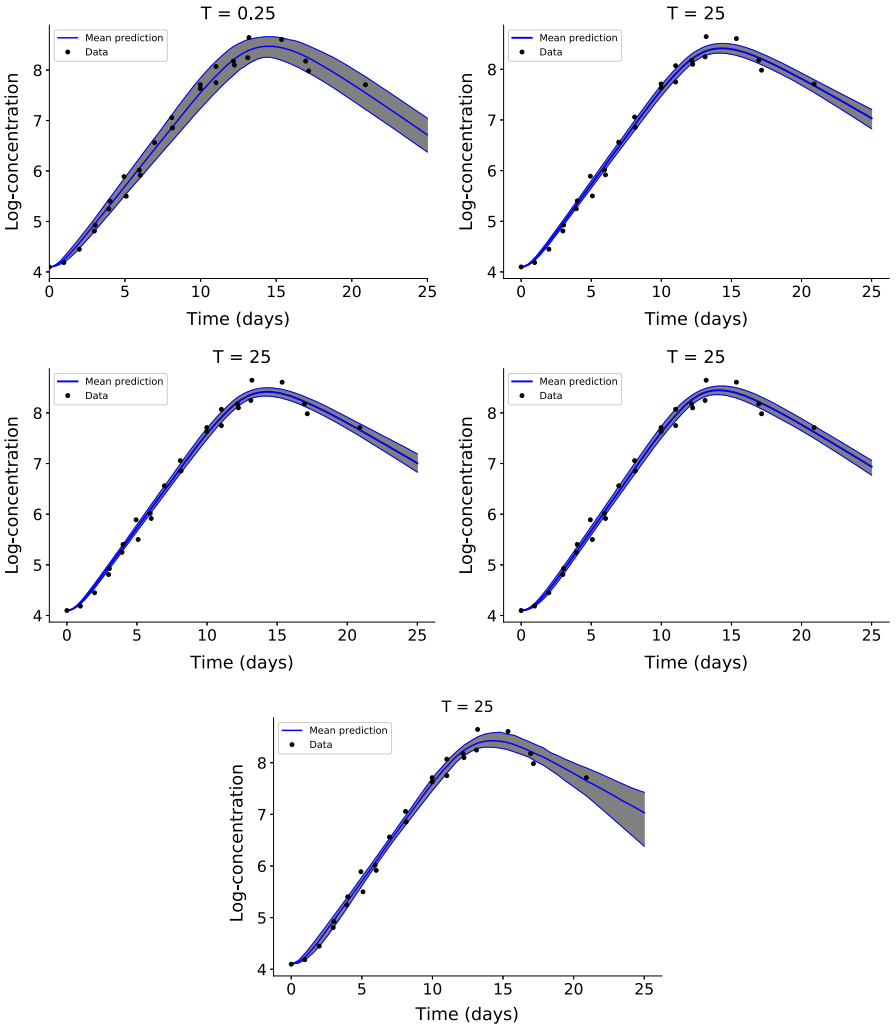


Fig. 5 Posterior uncertainty propagation for $L = 1, 2, 3, 4$ (first two rows) and MCMC (bottom)

approximate the posterior. The joint densities of (ξ_2, ξ_3), (ξ_2, ξ_4) and (ξ_3, ξ_4) confirm our previous guess that the Gaussian components are placed at the right locations that reveal the true shape of the posterior as we keep increasing their number. In general, however, the MCMC posterior should not be necessarily taken as the reference point in our case. That means that the qualitative disagreement in the joint densities of $(\xi_1, \xi_i), \imath = 2, 3, 4$ might not indicate that our Gaussian mixture approximation is poor. This is due to the fact that even MCMC algorithms can typically fail to detect possible multimodal structure. In practice, once the chain visits one of the modes it is unlikely that it will ever escape from it. To deal with such effects, one needs to employ more sophisticated MCMC algorithms such as the Metropolis-adjusted Langevin algorithm (MALA) (Roberts and Rosenthal 1998) which falls beyond the

Table 6 $\mathcal{F}[q^*]$ values as a function of L

L	$\mathcal{F}[q^*]$
1	−11.8953
2	−12.1081
3	−11.1929
4	−9.8656

scope of our analysis. The propagation results are shown in Fig. 5 and as we can see the confidence bands slowly become narrower as the number of component increases and are in excellent agreement with those obtained by MCMC. Overall we conclude that the variational approach tends to slightly underestimate the uncertainty. Finally the $\mathcal{F}[q^*]$ values obtained are shown in Table 6 where indeed it is verified that the $L = 4$ case achieves the largest value (that is the lowest KL value), suggesting a preferable solution over the rest.

5 Conclusions

We have thus far presented a novel Bayesian approach for calibrating the stochastic quasi-chemical model for bacterial growth. The approach consists of treating the model parameters as random variables admitting a prior probability density which is updated to its posterior using Bayes' rule. Estimation of the posterior, or more precisely its approximation, was performed using a variational method that finds the Gaussian (mixture) density that is closest to the true posterior by minimizing their KL distance.

Experimental data associated with the QCM are sparse both in quantity and in its parameter coverage. Credible model-based predictions both within and outside the observed parameter range is facilitated by rationally constructed probabilistic models of the parameters, and the ability to push these models forward through computational models. While the adopted Bayesian approach provided a rational formalism for describing probabilistic content, our proposed variational procedure was critical for enabling efficient sampling from the posterior and thus for integrating the probabilistic calibration step within a credible predictive engine.

In addition to the sensitivity estimates provided in the paper, a probabilistic characterization permits a rational risk assessment and the development of mitigation strategies that would not be possible otherwise. The confidence with which our proposed procedure permits the evaluation of the posterior densities lends further credence to their use in such decision-making.

Acknowledgements The authors gratefully acknowledge support from US Army Research Office Contract W911NF-14-C-0151.

Appendix A: Gradient Computation of the Log-joint Distribution $\log p(\mathbf{y}, \boldsymbol{\theta})$

In order to perform gradient-based optimization of the ELBO approximation $\mathcal{F}_2[q]$ with respect to $\{\boldsymbol{\mu}\}_i, \{\boldsymbol{\Sigma}_i\}_i, i = 1, \ldots, L$, we need to compute its gradient vector with entries

$$\frac{\partial}{\partial \zeta}\mathcal{F}_r[q] = \frac{\partial}{\partial \zeta}\mathcal{H}_0[q] + \frac{\partial}{\partial \zeta}\mathcal{L}_r[q] \tag{A.1}$$

where $\zeta = (\boldsymbol{\mu}_i)_j$, $(\boldsymbol{\Sigma}_i)_{jk}$, for $i = 1, \ldots, L$, $j, k = 1, \ldots, d$ and $r = 0, 2$. Below we provide the details in computing the gradient of $\mathcal{L}_r[q]$, $r = 0, 2$.

A.1: Gradient of $\mathcal{L}_r[q]$

For convenience, we set $J(\boldsymbol{\theta}) := \log p(\mathbf{y}, \boldsymbol{\theta})$. Then for $r = 0$, the derivatives of $\mathcal{L}_0[q]$ with respect to $\zeta = (\boldsymbol{\mu}_i)_j$ are

$$\frac{\partial}{\partial \zeta}\mathcal{L}_0[q] = \frac{1}{L}\frac{\partial}{\partial \theta_j}J(\boldsymbol{\mu}_i) \tag{A.2}$$

and with respect to $\zeta = (\boldsymbol{\Sigma})_{jk}$ are

$$\frac{\partial}{\partial \zeta}\mathcal{L}_0[q] = 0. \tag{A.3}$$

For $r = 2$ and $\zeta = (\boldsymbol{\Sigma}_i)_{jk}$, we get

$$\frac{\partial}{\partial \zeta}\mathcal{L}_2[q] = \frac{1}{2L}\frac{\partial^2}{\partial \theta_j \partial \theta_k}J(\boldsymbol{\mu}_i). \tag{A.4}$$

As mentioned above, the derivatives of $\mathcal{L}_2[q]$ with respect to $(\boldsymbol{\mu}_i)_j$ are not used in our optimization scheme and therefore are not computed here.

A.2: Derivatives of $J(\boldsymbol{\theta})$

First we rewrite $\boldsymbol{\theta} = (\boldsymbol{\xi}, \omega)$ and expand

$$J(\boldsymbol{\theta}) = J(\boldsymbol{\xi}, \omega) = \log p(\mathbf{y}|\mathcal{G}(\boldsymbol{\xi}), \omega) + \log p(\boldsymbol{\xi}) + \log p(\omega). \tag{A.5}$$

Throughout our numerical examples, we work with an isotropic Gaussian likelihood ($\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma \mathbf{I}_N)$); therefore, we set

$$L(\mathcal{G}(\boldsymbol{\xi}), \omega; \mathbf{y}) := \log p(\mathbf{y}|\mathcal{G}(\boldsymbol{\xi}), \omega) = \log \mathcal{N}(\mathbf{y}|\mathcal{G}(\boldsymbol{\xi}), e^{2\omega}\mathbf{I}_N), \tag{A.6}$$

and using the chain rule, we have

$$\frac{\partial J}{\partial \xi_j} = \sum_{s=1}^{N}\frac{\partial L}{\partial \mathcal{G}_s}\frac{\partial \mathcal{G}_s}{\partial \xi_j} + \frac{1}{p(\boldsymbol{\xi})}\frac{\partial p(\boldsymbol{\xi})}{\partial \xi_j} \tag{A.7}$$

$$\frac{\partial J}{\partial \omega} = \frac{\partial L}{\partial \omega} + \frac{1}{p(\omega)}\frac{dp(\omega)}{d\omega} \tag{A.8}$$

$$\frac{\partial^2 J}{\partial \xi_j \partial \xi_k} = \sum_{s,t=1}^{N} \frac{\partial^2 L}{\partial \mathcal{G}_s \partial \mathcal{G}_t} \frac{\partial \mathcal{G}_s}{\partial \xi_j} \frac{\partial \mathcal{G}_t}{\partial \xi_k} + \sum_{s=1}^{N} \frac{\partial L}{\partial \mathcal{G}_s} \frac{\partial^2 \mathcal{G}_s}{\partial \xi_j \partial \xi_k} + \frac{1}{p(\boldsymbol{\xi})} \frac{\partial^2 p(\boldsymbol{\xi})}{\partial \xi_j \partial \xi_k} - \frac{1}{p(\boldsymbol{\xi})^2} \frac{\partial p(\boldsymbol{\xi})}{\partial \xi_j} \frac{\partial p(\boldsymbol{\xi})}{\partial \xi_k} \tag{A.9}$$

$$\frac{\partial^2 J}{\partial \omega^2} = \frac{\partial^2 L}{\partial \omega^2} + \frac{1}{p(\omega)} \frac{d^2 p(\omega)}{d\omega^2} - \frac{1}{p(\omega)^2} \frac{dp(\omega)}{d\omega} \tag{A.10}$$

$$\frac{\partial^2 J}{\partial \xi_j \partial \omega} = \sum_{s=1}^{N} \frac{\partial^2 L}{\partial \mathcal{G}_s \partial \omega} \frac{\partial \mathcal{G}_s}{\partial \xi_j} \tag{A.11}$$

In the above expressions, it becomes clear that the Jacobian and Hessian of the forward model $\mathcal{G}(\boldsymbol{\xi})$ need to be computed. As mentioned in our application, the covariance matrices of the Gaussian mixtures components are taken to be diagonal which implies that only the diagonal elements of the Hessian of $\mathcal{G}(\boldsymbol{\xi})$ are necessary.

A.3: Log-Likelihood Derivatives

The derivatives of the log-likelihood function required for the expression in the previous subsection are given as follows:

$$\frac{\partial L}{\partial \mathcal{G}_s} = e^{-2\omega}(y_s - \mathcal{G}_s(\boldsymbol{\xi})) \tag{A.12}$$

$$\frac{\partial L}{\partial \omega} = e^{-\omega}\left(||\mathbf{y} - \mathcal{G}(\boldsymbol{\xi})||_2^2 e^{-2\theta} - k + 1\right) \tag{A.13}$$

$$\frac{\partial^2 L}{\partial \omega^2} = e^{-\omega}\left(k - 1 - 3||\mathbf{y} - \mathcal{G}(\boldsymbol{\xi})||_2^2 e^{-2\omega}\right) \tag{A.14}$$

$$\frac{\partial^2 L}{\partial \mathcal{G}_s \partial \mathcal{G}_t} = -e^{-2\omega} \tag{A.15}$$

$$\frac{\partial^2 L}{\partial \mathcal{G}_s \partial \omega} = -2e^{-3\omega}\left(y_s - \mathcal{G}_s(\boldsymbol{\xi})\right). \tag{A.16}$$

A.4: Derivatives of the Quasi-chemical Model

For the sake of generality and due to the presence of a nonlinear term in the quasi-chemical model, we present the general derivation of the system of ODEs satisfied by the derivatives of a solution $\mathbf{u}(t;\boldsymbol{\xi})$ of the QCM with respect to its parameters. Assume $\mathbf{u}(t;\boldsymbol{\xi})$ satisfies

$$\dot{\mathbf{u}} = \mathbf{g}(\mathbf{u}, t; \boldsymbol{\xi}) \tag{A.17}$$

$$\mathbf{u}(0) = \mathbf{u}_0, \tag{A.18}$$

where $\boldsymbol{\xi} \in \mathbb{R}^4$ are parameters and the initial condition is fixed and independent of $\boldsymbol{\xi}$. By simply differentiating the above system of equations, one can derive the following initial value problem satisfied by $v_{ij} = \partial u_i / \partial \xi_j$:

$$\dot{v}_{ij} = \sum_{s=1}^{4} \frac{\partial g_i}{\partial u_s} v_{sj} + \frac{\partial g_i}{\partial \xi_j} \tag{A.19}$$

$$v_{ij}(0) = 0. \tag{A.20}$$

Similarly, for the second derivatives $w_{ijk} = \partial^2 u_i / (\partial \xi_j \partial \xi_k)$ we get

$$\dot{w}_{ijk} = \sum_{s=1}^{4} \frac{\partial g_i}{\partial u_s} w_{ijk} + \sum_{s,t=1}^{4} \frac{\partial^2 g}{\partial u_s \partial u_t} v_{sj} v_{tk} + \frac{\partial^2 g_i}{\partial \xi_j \partial \xi_k} \tag{A.21}$$

$$w_{ijk} = 0. \tag{A.22}$$

In practice, during numerical implementation one need to first solve (A.17) and then solve (A.19) using the solution of the former as forcing. At last, (A.21) can be solved by using both the QCM solution and its gradient as forcing.

References

Banks, H., Bihari, K.: Modelling and estimating uncertainty in parameter estimation. Inverse Prob. **17**, 95 (2001)

Banks, H., Browning, W., Catenacci, J., Wood, T.: Analysis of a Quasi-chemical Kinetic Food Chemistry Model. Center for Research in Scientific Computation Technical Report CRSC-TR16-05. NC State University, Raleigh, NC (2016)

Baranyi, J., Roberts, T.: A dynamic approach to predicting bacterial growth in food. Int. J. Food Microbiol. **23**, 277–294 (1994)

Baranyi, J., Roberts, T., McClure, P.: A non-autonomous differential equation to model bacterial growth. Food Microbiol. **10**, 43–59 (1993)

Bickel, P., Doksum, K.: Mathematical Statistics: Basic Ideas and Selected Topics, vol. 2. CRC Press, Boca Raton (2015)

Bishop, C.: Pattern Recognition and Machine Learning, Information Science and Statistics. Springer, New York (2006)

Browning, W.J.: Near real-time quantification of stochastic model parameters. Tech. rep., prepared by Applied Mathematics Inc., Small Business Technology Transfer, Phase II Final Report, Army STTR Topic A13A-009 (28 September 2016)

Buchanan, R.: Predictive microbiology: Mathematical modeling of microbial growth in foods. In: ACS Symposium Series-American Chemical Society, (1992)

Buchanan, R., Whiting, R., Damert, W.: When is simple good enough: a comparison of the Gompertz. Baranyi, and three-phase linear models for fitting bacterial growth curves. Food Microbiol. **14**, 313–326 (1997)

Byrd, R., Lu, P., Nocedal, J., Zhu, C.: A limited memory algorithm for bound constrained optimization. SIAM J. Sci. Comput. **16**, 1190–1208 (1995)

Chaloner, K., Verdinelli, I.: Bayesian experimental design: a review. Stat. Sci. **10**(3), 273–304 (1995)

Chaspari, T., Tsiartas, A., Tsilifis, P., Narayanan, S.: Markov chain monte carlo inference of parametric dictionaries for sparse bayesian approximations. IEEE Trans. Signal Process. **64**, 3077–3092 (2016)

Chen, P., Zabaras, N., Bilionis, I.: Uncertainty propagation using infinite mixture of gaussian processes and variational bayesian inference. J. Comput. Phys. **284**, 291–333 (2015)

Doona, C., Feeherry, F., Ross, E.: A quasi-chemical model for the growth and death of microorganisms in foods by non-thermal and high-pressure processing. Int. J. Food Microbiol. **100**, 21–32 (2005)

Doona, C., Feeherry, F., Ross, E., Kustin, K.: Inactivation kinetics of listeria monocytogenes by highpressure processing: pressure and temperature variation. J. Food Sci. **77**, M458–M465 (2012)
Gershman, S., Hoffman, M., Blei, D.: Nonparametric variational inference. In: International Conference on Machine Learning (2012)
Goldbeter, A.: Biochemical Oscillations and Cellular Rhythms: The Molecular Bases of Periodic and Chaotic Behaviour. Cambridge University Press, Cambridge (1997)
Gompertz, B.: On the nature of the function expressive of the law of human mortality and on a new mode of determining the value of life contingencies. Philos. Trans. R. Soc. Lond. **115**, 513–583 (1825)
Haario, H., Saksman, E., Tamminen, J.: An adaptive Metropolis algorithm. Bernoulli **7**, 223–242 (2001)
Hastings, W.: Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**, 97–109 (1970)
Huber, M., Bailey, T., Durrant-Whyte, H., Hanebeck, U.: On entropy approximation for Gaussian mixture random vectors, In: IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems, MFI 2008, (pp. 181–188). IEEE, (2008)
Kullback, S., Leibler, R.: On information and sufficiency. Ann. Math. Stat. **22**, 79–86 (1951)
McMeekin, T., Brown, J., Krist, K., Miles, D., Neumeyer, K., Nichols, D., Olley, J., Presser, K., Ratkowsky, D., Ross, T., Salter, M.: Quantitative microbiology: a basis for food safety. Emerg. Infect. Dis. **3**, 541 (1997)
Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., Teller, E.: Equations of state calculations by fast computing machines. J. Chem. Phys. **21**, 1087–1091 (1953)
Pinski, F., Simpson, G., Stuart, A., Weber, H.: Algorithms for Kullback–Leibler approximation of probability measures in infinite dimensions. SIAM J. Sci. Comput. **37**, A2733–A2757 (2015)
Ricker, W.: Growth rates and models. Fish Physiol. **8**, 677–743 (1979)
Robert, C., Casella, G.: Monte Carlo Statistical Methods. Springer, Berlin (2013)
Roberts, G., Rosenthal, J.: Optimal scaling of discrete approximations to Langevin diffusions. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **60**, 255–268 (1998)
Ross, E., Taub, I., Doona, C., Feeherry, F., Kustin, K.: The mathematical properties of the quasi-chemical model for microorganism growth-death kinetics in foods. Int. J. Food Microbiol. **99**, 157–171 (2005)
Schnute, J.: A versatile growth model with statistically stable parameters. Can. J. Fish. Aquat. Sci. **38**, 1128–1140 (1981)
Silverman, B.: Density estimation for statistics and data analysis, vol. 26. CRC Press, Boca Raton (1986)
Stuart, A.: Inverse problems: a bayesian perspective. Acta Numer. **19**, 451–559 (2010)
Tarantola, A.: Inverse Problem Theory and Methods for Model Parameter Estimation. SIAM, Philadelphia (2005)
Taub, I., Ross, E., Feeherry, F.: Model for predicting the growth and death of pathogenic organisms. In: Van Impe, J.F.M., Gernaerts, K. (eds.) Proceedings of the Third International Conference on Predictive Modeling in Foods (2000)
Taub, I., Feeherry, F., Ross, E., Kustin, K., Doona, C.: A quasi-chemical kinetics model for the growth and death of Staphylococcus aureus in intermediate moisture bread. J. Food Sci. **68**, 2530–2537 (2003)
Tsilifis, P., Bilionis, I., Katsounaros, I., Zabaras, N.: Computationally efficient variational approximations for bayesian inverse problems. J. Verif. Valid. Uncertain. Quantif. **1**, 031004 (2016)
Tsilifis, P., Ghanem, R., Hajali, P.: Efficient bayesian experimentation using an expected information gain lower bound. SIAM/ASA J. Uncertain. Quantif. **5**, 30–62 (2017)
Vrettas, M., Cornford, D., Opper, M.: Estimating parameters in stochastic systems: a variational bayesian approach. Phys. D **240**, 1877–1900 (2011)
Whiting, R.: Modeling bacterial survival in unfavorable environments. J. Ind. Microbiol. **12**, 240–246 (1993)
Whiting, R., Sackitey, S., Calderone, S., Morely, K., Phillips, J.: Model for the survival of Staphylococcus aureus in nongrowth environments. Int. J. Food Microbiol. **31**, 231–243 (1996)
Ye, J., Rey, D., Kadakia, N., Eldridge, M., Morone, U., Rozdeba, P., Abarbanel, H., Quinn, J.: Systematic variational method for statistical nonlinear state and parameter estimation. Phys. Rev. E **92**, 052901 (2015)
Zwietering, M., Jongenburger, I., Rombouts, F., Van't Riet, K.: Modeling of the bacterial growth curve. Appl. Environ. Microbiol. **56**, 1875–1881 (1990)