Ensemble Modeling of Biological Systems

David Swigon

Abstract. Mathematical modeling of biological systems must cope with difficulties that are rarely present in traditional fields of applied mathematics, such as a large number of components involved, extreme complexity and variety of interactions, and the lack of reproducible and consistent data. These difficulties may be overcome by models of new type, termed ensemble models, which allow for the parameters and the model structure to vary and thereby describe a population of all models that are consistent with biological knowledge about the system. Ensemble models are identified and their parameters are estimated using Bayesian techniques, and the models are subsequently used to provide probabilistic predictions about the system behavior and and its response to changes in conditions. We here survey the basic methodology of ensemble modeling and its applications to biological systems.

Keywords. ensemble modeling, Bayesian inference, parameter estimation, system identification.

AMS classification. 92B05,62F15,93B30,65C40,91-08.

1 Introduction

Mathematical modeling of biological systems has intensified considerably in recent years, especially with the advent of experimental techniques such as microarray analysis of gene expression profiles, whole genome sequencing, or high throughput flow cytometry and ELISA biochemical assays, that are capable of providing a wealth of new biological data. There are important differences in the character of biological system models, when compared to those in traditional fields of applied mathematics, which require new approaches to model development, parameter estimation and model prediction. [44, 52]

The first difference is in the modeling approach. Traditional model development relies on a strategy of reduction: the search for a small number of general laws governing the behavior of the system or a decomposition of the system into simple components governed by few interactions, and the design of experiments in which those laws or interactions can be characterized. Biological modeling has not been universally successful with this approach, primarily because in many situations the decomposition of the system into the simplest laws is tedious and impractical, especially when complexities in the behavior extend over multiple length- and time-scales.[74] Many genetic

D.S. acknowledges support by NSF grant DMS0739261 and NIH grants R01-GM83602 and R01-DC008290

network models have hundreds of components and thousands of reactions, and it is nearly impossible to design experiments that would isolate each such reaction so that its kinetic rate constants can be measured with sufficient accuracy . Furthermore, the results of such experiments, performed *in vitro*, may not represent the true interaction *in vivo*, for that reaction may be influenced by other reactions and molecules in ways yet unknown.

The second difference is in the amount and quality of experimental data. Reductionist approach requires large amount of low dimensional data to provide statistical confidence in estimates of parameters. In biological modeling, the data may be available only in limited quantities or in large quantities (obtained by high throughput techniques) but over multiple dimensions.[42] More importantly, the data may not be available for the same subject (organism or cell), but as a collection of measurements obtained for a population of systems that were subject to identical stimulus. In the case of time-dependent (longitudinal) data, it is frequently impossible to use the same subject for two time-point measurements because data collection methods require its destruction (e.g., animal sacrifice). The data at different time-points thus come from different subjects and it is not guaranteed that these subjects represent identical copies of the same system. Although at the molecular level the biological processes in such subjects are likely to be governed by chemical and biochemical laws with similar parameters, there may be genetic differences between cells or organisms which can lead to discrepancies in interactions that multiply into large differences in parameters describing macroscopically observable phenomena.

The combination of these difficulties calls for the development of a spectrum of phenomenological models tailored to specific situations, and the use of parameter/model inference techniques which result in distributions of parameters rather than specific values, describing parameter variability over a population. Such models have been called ensemble models, and they are understood not to represent the behavior of an individual subject but that of a whole population. The models are probabilistic in principle, because of the underlying distribution of parameters, but not necessarily stochastic - that term is reserved for models that describe interactions and laws that are probabilistic in individual interactions (such as binding and unbinding of molecules). Indeed, majority of ensemble models in the literature are deterministic, in that the evolution of the system, once its parameters are fixed, is free of random effects. The best approach to formulating ensemble models appears to be using Bayesian inference. The use of Bayesian techniques in biology thus brings a new meaning to the probability densities produced by Bayesian computation. Because data used for parameter estimation comes from a population of systems, the distributions can be thought of as distributions of parameters within the population.

This paper contains the description and properties of ensemble models, gives examples of the use of such models in studies of biological systems, and outlines both conceptual and technical open problems in this new, important, and exciting area of research. Basic introduction to parameter estimation and inverse problems can be ob-

tained in [38, 39, 37, 2] or [70]. The focus here is on modeling within the framework of ordinary differential equations; for recent review of the use of Bayesian methods in bioinformatics and computational system biology, focused primarily on sequence analysis, microarray data analysis, protein bioinformatics, and network inference, see [22, 77].

2 Background

For simplicity, consider a model formulated as an initial-value problem for a system of ordinary differential equations, in which the time evolution of a vector of state variables $\mathbf{x}$ depends on a vector of parameters $\mathbf{p}$ (such as kinetic rate constants) and a vector of inputs $\mathbf{u}$. The observations about the system are made by measuring output variables $\mathbf{y}$ which are functions of the state variables, and may depend on the parameters and inputs as well.

$$\dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t), \mathbf{p}), \qquad \mathbf{x}(t) = \mathbf{x}_0 \tag{2.1}$$

$$\mathbf{y}(t) = \mathbf{g}(\mathbf{x}(t), \mathbf{u}(t), \mathbf{p}) \tag{2.2}$$

The function $\mathbf{u}(.)$ may be employed to control the dynamics of the system, or to perturb its dynamics in order to study its behavior. In majority of biological models $\mathbf{u}$ is fixed and the dynamics of the system is studied by changing the initial conditions and/or parameters. We assume that the model (2.2) is mathematically well posed and has a unique solution $(\mathbf{x}(t), \mathbf{y}(t))$ for any $\mathbf{p}$, $\mathbf{x}_0$, and admissible function $\mathbf{u}(.)$. (This condition can be relaxed by restricting attention to sub-domains of the parameter, state, and input spaces in which existence and uniqueness holds.) In a classical setting, one assumes that (i) the system under consideration is described by a model with a unique parameter set $\mathbf{p}$ and a trajectory $\mathbf{y}(t; \mathbf{p})$ and (ii) the observed values $\overline{\mathbf{y}}^i$ of the output variables at time points t_i are normally distributed random variables with mean $\mathbf{y}(t_i; \mathbf{p})$ and variance $\boldsymbol{\sigma}_i^2$ due to independent random measurement noise. In that case the probability density for $\overline{\mathbf{y}}^i$ is

$$L_i(\overline{\mathbf{y}}^i|\mathbf{p}) = \prod_j (2\pi\sigma_{i,j}^2)^{-\frac{1}{2}} e^{-\frac{(y_j(t_i;\mathbf{p}) - \overline{y}_j^i)^2}{2\sigma_{i,j}^2}} \tag{2.3}$$

where j ranges over the components of $\mathbf{y}$.

We assume that the data Q provided about the system consist of the observed values $\overline{\mathbf{y}}^i$ and their uncertainties $\boldsymbol{\sigma}_i^2$. If measurement errors are independent, the likelihood of observing the full set of data for a model with parameters $\mathbf{p}$ is

$$L(Q|\mathbf{p}) = \prod_i L_i(\overline{\mathbf{y}}^i|\mathbf{p}) \tag{2.4}$$

In practice, the values of the parameters are not known and are to be estimated from the data Q. A *cost function* (or *objective function*) $E(\mathbf{p}, Q)$ is constructed to reflect

the agreement between the model with parameters $\mathbf{p}$ and the data Q. It quantifies the difference between measured and model-predicted values of the output at each time point, and depends on the source of the discrepancy between model output and measured data. Traditionally $E(\mathbf{p}, Q)$ is taken to be the negative logarithm of the likelihood $L(Q|\mathbf{p})$, i.e.,

$$E(\mathbf{p}, Q) = -\log L(Q|\mathbf{p}) = E_0 + \sum_i \sum_j \frac{(y_j(t_i; \mathbf{p}) - \overline{y}_j^i)^2}{2\sigma_{i,j}^2} \tag{2.5}$$

More general distributions of the measurement error would result in different forms of (2.5). Furthermore, in addition to the information included in the measured data, one can include in $E(\mathbf{p}, Q)$ terms that account for observed maximum or minimum values of outputs over the duration of experiment, expected timing of peak values of output variables, etc. If multiple runs of the experiment are done with different initial conditions and/or input functions $\mathbf{u}$, then the function E in (2.5) should include additional summation over all such conditions.

Traditional parameter fitting identifies the parameters of the observed system with the parameter set $\mathbf{p}^*$ that minimizes $E(\mathbf{p}, Q)$ given the data Q. When the form (2.5) is used, the vector $\mathbf{p}^*$ is called the *maximum likelihood estimate* of the parameters. The vector $\mathbf{p}^*$ exists for any reasonable choice of the function E, but it is unique only if the model is *structurally identifiable*, i.e., if there are no two distinct parameter sets that would lead to systems with identical dynamics. Multiple types of identifiability have been defined and various algebraic criteria and transform methods for testing identifiability have been developed, see, e.g., the review [54].

If the existence and uniqueness of the minimizer $\mathbf{p}^*$ is guaranteed, one can proceed with its computation using function minimization algorithms, such as the Levenberg-Marquardt scheme [43]. Such methods may require the computation of the trajectory sensitivity gradient $\mathbf{S}(t) = \partial \mathbf{x}(t)/\partial \mathbf{p}$ which can be obtained for the system (2.2) by simultaneous integration of the associated ODE problem:[27]

$$\dot{\mathbf{S}}(t) = \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \mathbf{S}(t) + \frac{\partial \mathbf{f}}{\partial \mathbf{p}}, \qquad \mathbf{S}(t) = \mathbf{0} \tag{2.6}$$

For stiff or chaotic systems the trajectory is extremely sensitive to the initial value and parameter values. Multiple shooting algorithm removes this difficulty by converting the initial value problem to multiple boundary value problems that are solved for segments of the full trajectory.[73, 9, 3] Kalman filter technique is another popular method for estimating a unique parameter set. It is a stepwise algorithm based on predicting the trajectory $\mathbf{x}(t + \Delta t)$ using a model estimated from the data up to time t, and then correcting the parameters of the model using the data $\mathbf{y}(t + \Delta t)$. It has been shown that if the model is linear and the data are subject to Gaussian noise, the parameters of the model converge to the maximum likelihood estimate $\mathbf{p}^*$. There is a large amount of literature on extensions of the Kalman filter method to nonlinear systems.[75]

Note that the identifiability of a model is a structural property that is independent of the data. Sontag has shown that to determine all r parameters of an identifiable system, one needs at most $2r + 1$ judiciously chosen measurements.[66] Thus, for sparse enough or poorly chosen data the minimizer $\mathbf{p}^*$ may not be unique, even if the system is identifiable. One can still proceed with minimization algorithm for such an underdetermined problem, but the resulting value of $\mathbf{p}^*$ will depend on the initial guess for $\mathbf{p}$ used in the minimization algorithm. By varying initial guesses one can obtain a collection of parameter sets that represent the set of optimal parameter values for the system; these values lie on a lower-dimensional manifold in the parameter space defined as $E(\mathbf{p}, Q) = E(\mathbf{p}^*, Q)$. A model with such collection of parameters has been called ensemble model in some literature [17, 49]. Alternatively, one can try to enforce uniqueness of the minimizer by adding to $E(\mathbf{p}, Q)$ an artificial term that controls the magnitude of $\mathbf{p}$, such as $\sum_i p_i^2$ or $\sum_i (\log p_i)^2$ (which controls both the large and small values of $\mathbf{p}$).

An issue of prime importance in parameter estimation is parameter sensitivity of the model, i.e., the dependence of objective value on a change in parameters. The sensitivity of a model (given data Q) can be assessed by analyzing the quadratic expansion of $E(\mathbf{p}, Q)$ about the minimizer $\mathbf{p}^*$:

$$E(\mathbf{p}, Q) = E_0 + \frac{1}{2}(\mathbf{p} - \mathbf{p}^*)^T \mathbf{H} (\mathbf{p} - \mathbf{p}^*) + O(||\mathbf{p} - \mathbf{p}^*||^3) \tag{2.7}$$

The eigenvectors of $\mathbf{H}$ determine the principal linear combinations of parameters; the largest eigenvalue of the Hessian $\mathbf{H}$ correspond to the stiff parameter combinations, i.e., a linear combinations of parameters (specified by the corresponding eigenvectors) along which a small change results in a large increase in the cost function. These represent the directions in which the system is sensitive to parameter changes. Small eigenvalues correspond to the soft directions along which the change in $E(\mathbf{p}, Q)$ is relatively small. The sensitivity can be decomposed into three components: sensitivity of the output to the trajectory $\mathbf{G}(t) = \partial \mathbf{g}(\mathbf{x}(t), \mathbf{u}(t), \mathbf{p}) / \partial \mathbf{x}$ computed for the trajectory $\mathbf{x}(t)$, the sensitivity of the trajectory to parameters $\mathbf{S}(t)$, and the sensitivity of the objective function $E(\mathbf{p}, Q) \equiv \tilde{E}(\mathbf{y}(.; \mathbf{p}))$ to the output $\mathbf{y}$:

$$\mathbf{H} = \sum_i \mathbf{S}(t_i)^T \mathbf{G}(t_i)^T \frac{\partial^2 \tilde{E}}{\partial \mathbf{y}^2}(t_i) \mathbf{G}(t_i) \mathbf{S}(t_i) \tag{2.8}$$

(If weighted Euclidean norm (2.5) is used to compute $E(\mathbf{p}, Q)$ then $\partial^2 \tilde{E}(t_i) / \partial \mathbf{y}^2$ is a diagonal matrix with entries $1/(4\sigma_{i,j}^2)$.) In classical parameter fitting, the sensitivity $\mathbf{H}$ is used to characterize the accuracy with which model parameters can be determined. Along the soft direction the accuracy is low, while along the stiff direction the accuracy is high. For large number of ODE models inspired by biological systems, most parameters contribute to the eigenvector of some soft direction which results in apparent parameter insensitivity.[29]

In addition to the traditional parameter fitting techniques, there are methods based on theory of probability. Bayesian parameter inference [38, 39, 37, 56, 22, 14] provides, for a given model and a set of data, not a unique set of parameters, but a parameter distribution describing the probability (some authors say plausibility) that the model has a parameter set $\mathbf{p}$ given the data Q. This distribution is characterized by the *posterior density* $\rho(\mathbf{p}|Q)$. The application of Bayesian inference requires that the likelihood $L(Q|\mathbf{p})$ of observing the data Q for a model with parameters $\mathbf{p}$ is specified. In the classical setting described above one would take $L(Q|\mathbf{p}) = \exp(-E(\mathbf{p}, Q))$ with $E(\mathbf{p}, Q)$ defined as in (2.5). The posterior density is related to $L(Q|\mathbf{p})$ by

$$\rho(\mathbf{p}|Q) = \frac{L(Q|\mathbf{p})\theta(\mathbf{p})}{\int L(Q|\mathbf{p})\theta(\mathbf{p})d\mathbf{p}} \tag{2.9}$$

where $\theta(\mathbf{p})$, the *prior density*, reflects all information known about the distribution of the parameters before the data are taken into account.

The prior density is generally based on probable ranges of parameters obtained from biological literature, but it can also include heuristic terms that account for qualitative criteria known to be satisfied by the system, such as the number and stability of equilibrium points in the system, long term behavior of the trajectory, or the presence of oscillations in the trajectory.[33] The choice of the prior distribution plays an important role in Bayesian parameter estimation [11, 50], although that role is diminished the more data is available for parameter estimation. Some researchers argue for non-informative (also called objective) priors which are flat relative to the likelihood function, but such priors are difficult to construct [59], while others argue for the use of subjective priors that express our belief in likely ranges of parameter values [8]. Jeffreys [39] discusses a uniform prior based on Fisher information matrix that does not change much over the region in which the likelihood is significant and does not assume large values outside that range. In biological ODE models it is a common practice to employ for kinetic rate constants priors that are uniform in the log space over biologically reasonable parameter ranges [13]. Such a choice does not affect the principles of Bayesian inference, only the sampling metric on the parameter space and the prior distributions.

The posterior distribution $\rho(\mathbf{p}|Q)$ can be characterized by its mean $\bar{\mathbf{p}}$ and a covariance matrix $\mathbf{C}$, which, in the case of a quadratic cost function (2.7), is equal to $\mathbf{H}^{-1}$. For more general distributions, $\mathbf{C}^{-1}$ provides better information about global parameter sensitivity of the model than $\mathbf{H}$. This information can be obtained by eigenvalue analysis as described above.

3 Ensemble model

Ensemble modeling replaces the uniquely parametrized model (2.2) with a collection of models. In the simplest case, the models in the ensemble all have identical structure

(i.e., identical functions $\mathbf{f}, \mathbf{g}$) but different parameter values. The ensemble is then characterized by the probability density function $\rho(\mathbf{p})$. The most common, but not necessarily exclusive, interpretation of an ensemble model is one in which the model represents the response of a population of individuals and the parameter distribution represents the variability of parameters within that population. The solution of the ensemble version of the model (2.2) for any fixed $\mathbf{x}_0$ and $\mathbf{u}(.)$ is a special type of a stochastic process for which each sample trajectory $\mathbf{x}(t;\mathbf{p})$ is a deterministic solution of (2.2) and the probability of the trajectory is equal to $\rho(\mathbf{p})$. The output $\hat{\mathbf{y}}(t)$ of the ensemble model (2.2) is a time-dependent random variable with probability density

$$p_t(\hat{\mathbf{y}})d\hat{\mathbf{y}} = \int_{\Omega} \rho(\mathbf{p})d\mathbf{p} \tag{3.1}$$

where $\mathbf{\Omega} = \{\mathbf{p}|\hat{\mathbf{y}} < \mathbf{y}(t;\mathbf{p}) \leq \hat{\mathbf{y}} + d\hat{\mathbf{y}}\}$. The likelihood of observing the value $\bar{\mathbf{y}}^i$ as an output of the ensemble model at time t_i is therefore

$$L_i(\bar{\mathbf{y}}^i) = p_{t_i}(\bar{\mathbf{y}}^i) \tag{3.2}$$

In this case however the densities $p_t(.)$ are not independent and hence, in general, the likelihood of observing the data, $L(Q)$, is not equal to $\prod_i L_i(\bar{\mathbf{y}}^i)$ (cf. (2.4)).

Both the classical model with noisy measurement and the ensemble model result in probability distributions for data and therefore may be difficult to distinguish based on the observations alone. One of the main differences between the two models is that in the case of the classical model the mean output values $< \mathbf{y}^i >= \int \zeta L_i(\zeta|\mathbf{p})d\zeta$, with L_i as in (2.3), all lie on a single trajectory of (2.2), i.e., $< \mathbf{y}^i >= \mathbf{y}(t_i;\mathbf{p})$ for all i, while in the case of the ensemble model there is no single trajectory of (2.2) containing the mean values $< \mathbf{y}^i >= \int \zeta L_i(\zeta)d\zeta$ with L_i as in (3.2) .

The ensemble model resembles a classical model with parameters estimated using Bayesian inference. Both are characterized by a distribution over the space of parameters, but there is an important distinction. The posterior distribution estimate for a classical model reflects the likelihood of a particular parameter value given the available data, while the parameter distribution of the ensemble model, $\rho(\mathbf{p})$, describes the frequency of occurrence of a particular parameter set in the ensemble and is clearly independent of the data. The posterior distribution depends on the data and contains additional information, such as the sensitivity of the model to parameters and the sensitivity of the cost function to the data. The posterior density will converge to unique parameter values if one collects sufficiently large amount of data so as to average out the measurement error. On the other hand, no amount of data can average out the effect of the parameter distribution, it can only improve the accuracy of its estimate. For the moment, however, it appears that Bayesian inference is the only readily available tool for estimating the probability $\rho(\mathbf{p})$, it is being used in that way with the hope that with sufficient amount of data the influence of other factors will diminish, and the Bayesian estimate $\rho(\mathbf{p}|Q)$ will converge to $\rho(\mathbf{p})$ just like it converges to a localized density in the case of a classical model with a unique parameter value.

Brown and Sethna [13] have made several interesting observations about parameter distributions inferred for biological ensemble models with large numbers of parameters. They found that cost functions for such models have $\mathbf{H}$ (or $\mathbf{C}^{-1}$) with eigenvalues distributed almost uniformly (on the logarithmic scale) over broad range of magnitudes, no matter how detailed are the data used to fit the model. They also noted that the eigenvectors corresponding to the eigenvalues generally include nontrivial components along multiple parameters. They termed such models "sloppy", which is meant to indicate that both the structure and parameter values of the models cannot not be determined with certainty, however, the behavior predicted by the ensemble models was very well characterized. They also discuss the selection of optimal model using Bayesian techniques.

Gutenkunst et al. [29] extended the work of [13] by analyzing a collection of biological models and finding that they all have the spectral distribution characteristic of sloppy models. They found that fitting even to large amount of data leaves many parameters poorly determined. They also found that, contrary to the expectation, if a random perturbation is made to the parameters of the model, of magnitude much smaller than the variance given by the largest eigenvalue, then the fit of the model is worsened significantly. Thus, having poorly determined parameters does not mean that the parameters can take on any value, because their values are tightly correlated. They suggest to shift the focus of investigation from parameter estimation onto ensemble model predictions. Within the Bayesian framework, the posterior likelihood $\pi(R|Q)$ of observing additional output R of the model given existing data Q obeys[20]

$$\pi(R|Q) = \int p(R, \mathbf{p}|Q) d\mathbf{p} = \int L(R|\mathbf{p}) \rho(\mathbf{p}|Q) d\mathbf{p} \tag{3.3}$$

Of course, in ensemble modeling one is not restricted to using just one model structure. In the case of a finite number of distinct models $M_1, M_2, ..., M_m$, the likelihood of the model i given data Q is given by the marginal posterior probability $P(M_i|Q)$:

$$P(M_i|Q) = \frac{g(Q|M_i)P(M_i)}{\sum_i g(Q|M_i)P(M_i)} \tag{3.4}$$

where $P(M_i)$ is the prior probability expressing our belief in the accuracy of the model i, and $g(Q|M_i)$ can be written as

$$g(Q|M_i) = \int L(Q|\mathbf{p}, M_i) \theta(\mathbf{p}|M_i) d\mathbf{p} \tag{3.5}$$

Ensemble model can be extended to include models of different types via weighted averaging with weights corresponding to their posterior probabilities.[15, 23]

4 Computational techniques

The posterior probability density $\rho(\mathbf{p}|Q)$ can be computed by a variety of methods. The most efficient and generally applicable method appears to be Markov Chain Monte

Carlo (MCMC) sampling [24, 20], which is based on Metropolis-Hastings algorithm [53, 31]. Originally, MCMC method was designed in computational physics to sample the Boltzmann distribution $\rho(\mathbf{p}) \propto \exp(-\Phi(\mathbf{p}))$ for the states of a system with potential $\Phi(\mathbf{p})$, but the method can be used to sample any distribution. MCMC constructs a collection of points $\mathbf{p}^1, \mathbf{p}^2, ...$ as a trajectory of a Markov chain constructed so that its limiting distribution is equal to $\rho(\mathbf{p}|Q)$. The algorithm, here adapted to the computation of the Bayesian posterior density (2.9) , works as follows:

(i) Initialize $\mathbf{p}^k$, $k = 1$

(ii) Sample $\hat{\mathbf{p}}$ from a proposal distribution $q(\hat{\mathbf{p}}|\mathbf{p}^k)$

(iii) Solve the ODE system (2.2) and compute $\rho(\hat{\mathbf{p}}|Q) = L(Q|\hat{\mathbf{p}})\theta(\hat{\mathbf{p}})$

(iv) Set $\mathbf{p}^{k+1} = \hat{\mathbf{p}}$ with probability $P = \min\left\{1, \frac{\rho(\hat{\mathbf{p}}|Q)q(\mathbf{p}^k|\hat{\mathbf{p}})}{\rho(\mathbf{p}^k|Q)q(\hat{\mathbf{p}}|\mathbf{p}^k)}\right\}$
or $\mathbf{p}^{k+1} = \mathbf{p}^k$ with probability $1 - P$

(v) Increment k by 1, go to (ii)

The choice of the proposal distribution q does not affect the limiting distribution for $\mathbf{p}^k$ as long as q is symmetric, i.e., as long as $q(\mathbf{a}|\mathbf{b}) = q(\mathbf{b}|\mathbf{a})$. If q is symmetric, then MCMC algorithm implies that $\hat{\mathbf{p}}$ is accepted with certainty when $\rho(\hat{\mathbf{p}}|Q) \geq \rho(\mathbf{p}^k|Q)$, and that there is chance that $\hat{\mathbf{p}}$ will be accepted even if $\rho(\hat{\mathbf{p}}|Q) < \rho(\mathbf{p}^k|Q)$, but that chance decreases with decreasing $\rho(\hat{\mathbf{p}}|Q)$. This mechanism allows the chain to escape from local maxima of $\rho(\mathbf{p}|Q)$. The variance of the proposal distribution q should be chosen so that the acceptance probability P in step (iv) is on average about 25%. Higher variance leads to larger proportion of rejections and a waste of computing time, while lower variance results in small distances between $\hat{\mathbf{p}}$ and $\mathbf{p}^k$ and leads to inefficient sampling of the parameter space. The distribution $q(\hat{\mathbf{p}}|\mathbf{p}^k)$ is usually chosen to be a multivariate Gaussian distribution with mean $\mathbf{p}^k$.

As $k \to \infty$, the distribution of points $\mathbf{p}^k$ approaches $\rho(\mathbf{p}|Q)$. A sample of points $\mathbf{p}^1, ..., \mathbf{p}^m$ can be used to estimate the ensemble average of any trajectory-dependent quantity G as

$$< G(t) > \cong m^{-1} \sum_{i=1}^{m} G(\mathbf{x}(t; \mathbf{p}^i)) \tag{4.1}$$

and the percentile value $P_X(G)$ as the smallest number that is larger than $X\%$ of values of $\{G(\mathbf{x}(t; \mathbf{p}^i))\}_{i=1}^m$.

There are technical issues that need to be addressed when applying MCMC to Bayesian computation [65]. The most important problem is how to decide whether the Markov chain generated by MCMC has converged to the stationary distribution. A number of statistical tests can be utilized test to the convergence and mixing of the chain [16]. A commonly used test of Gelman and Rubin [21] compares the mean values of parameters of two chains running in parallel. While coding of MCMC and convergence tests may be a hurdle, it can be avoided by using one of several packages

for the analysis of biological dynamical systems that include Bayesian inference, for example, ABC-SysBio [48], Synbioss [34], BioBayes [76], or KINSOLVER [1].

MCMC techniques can be generalized to include multiple model types. One possibility is to enlarge parameter space to include the model as an additional unknown and allow jumps between the models (*reversible jump* MCMCs, or RJMCMCs for short). A prior over the model space must be specified, but with a good choice of the jumps the number of the models need not be specified in advance. [26, 64]. Alternative approach is the *birth and death* MCMC (BDMCMC) where the time between jumps to a model of different dimension is determined by a rate constant, and moves between models are always accepted - the probability of a model is determined by the length of time MCMC spends at that model. [67]

Greater efficiency of MCMC computations can be achieved using the technique of *parallel tempering*, which results in a more thorough exploration of the parameter space, and ultimately faster convergence of the chain.[30] Unlike MCMC, which consists of a single Markov chain, parallel tempering algorithm generates samples $\mathbf{p}^{k,i}$ of multiple Markov chains with limiting probabilities $\rho_i(\mathbf{p}) \propto \exp(-\beta_i \Phi(\mathbf{p})) \propto \rho(\mathbf{p})^{\beta_i}$ evaluated for different values β_i of a new parameter β. The origin of this technique is again in statistical physics where Φ represents the potential of the system and β is proportional to the reciprocal of the temperature at which the system is observed. At low β (i.e., high temperature), the potential differences between any two states fo the system are smaller and hence the chain can have a bigger step size and explore a larger region of the state space. At high β (i.e., low temperature), the distribution becomes more focused in small regions [19]. The chains are evolved independently and at regular intervals the parameter sets $\mathbf{p}^{k,i}$ and $\mathbf{p}^{k,i+1}$ for two neighboring values β_i and β_{i+1} are swapped with probability

$$\hat{P} = \min\left\{1, \left(\frac{\rho(\mathbf{p}^{k,i+1}|Q)}{\rho(\mathbf{p}^{k,i}|Q)}\right)^{\beta_i - \beta_{i+1}}\right\} \tag{4.2}$$

Swapping of the parameter values allows for the region surrounding a newly discovered high probability parameter set to be explored by low temperature chain. The modified algorithm becomes

(i) Initialize $\mathbf{p}^{k,i}$ and β_i, $k = 1$, $i = 1, ..., C$.

(ii) For each $i = 1, ..., C$
- Sample $\hat{\mathbf{p}}^i$ from a proposal distribution $q_i(\hat{\mathbf{p}}^i|\mathbf{p}^{k,i})$
- Solve the ODE system (2.2) with $\mathbf{p} = \hat{\mathbf{p}}^i$ and compute $\rho(\hat{\mathbf{p}}^i|Q) = L(Q|\hat{\mathbf{p}}^i)\theta(\hat{\mathbf{p}}^i)$
- Set $\mathbf{p}^{k+1,i} = \hat{\mathbf{p}}^i$ with probability $P = \min\left\{1, \frac{\rho(\hat{\mathbf{p}}^i|Q)^{\beta_i} q_i(\mathbf{p}^{k,i}|\hat{\mathbf{p}}^i)}{\rho(\mathbf{p}^{k,i}|Q)^{\beta_i} q_i(\hat{\mathbf{p}}^i|\mathbf{p}^{k,i})}\right\}$ or $\mathbf{p}^{k+1,i} = \mathbf{p}^{k,i}$ with probability $1 - P$

(iii) For each $i = 1, ..., C-1$ swap $\mathbf{p}^{k,i}$ and $\mathbf{p}^{k,i+1}$ with probability $\hat{P}$ given by (4.2)

(iv) Increment k by 1, go to (ii)

The main benefit of the parallel tempering algorithm is to improve mixing by allowing the low temperature chain (the primary chain of interest) to escape energy barriers and converge faster to a stationary probability distribution. The values of β_i should be chosen so that $1 = \beta_1 > \beta_2 > ... > \beta_C$. In computational physics literature there is an ongoing discussion about the optimal strategy for choosing β^i for various problem types.[58, 46, 41, 72]. A rule of thumb is that β_i should be spaced so that the swapping probability is on average about 20%. Too closely spaced β_i result in an inefficient sampling of the space by the highest temperature chain. Too widely spaced β_i result in an insufficient communication between the chains. For any choice of β_i, the variances of the proposal distributions q_i must be adjusted so that the acceptance of the proposed value in the step (ii) is 25% on average for each chain. Only the chain $\mathbf{p}^{k,1}$ corresponding to $\beta_1 = 1$ samples the posterior distribution $\rho(\mathbf{p}|Q)$.

MCMC with the included parameter β can be used to find the maximum likelihood parameter set $\mathbf{p}^*$ by using a simulated annealing algorithm [7], in which the value of β is gradually increased during the simulation and hence the limiting distribution $\rho(\mathbf{p})$ becomes narrower and localized near $\mathbf{p}^*$.

Approximate Bayesian computation (ABC) has become popular recently due to its computational efficiency, especially in cases in which the likelihood function cannot be written out explicitly (for example, when no standard deviations are given for the observed data or because we require precise agreement between the simulated and observed data). [62, 51, 71, 5] ABC presents an alternative representation of the posterior distribution $\rho(\mathbf{p}|Q)$. The output of the ABC algorithm is a sample from the distribution $\rho(\mathbf{p}|d(Q^*, Q) \leq \epsilon)$ where $d(Q^*, Q)$ denotes the distance between the experimental data Q and the simulated data Q^* corresponding to the parameter $\mathbf{p}$. When ϵ is small, the distribution $\rho(\mathbf{p}|d(Q^*, Q) \leq \epsilon)$ is a good approximation to the distribution $\rho(\mathbf{p}|Q)$. Toni et al. [71] proposed a novel approximate Bayesian computation (ABC) method for evaluating posterior distributions for ODE models based on sequential Monte-Carlo (SMC) method, in which a fixed number of sampled parameter values are propagated though a sequence of intermediate distribution until they represent a sample from the target distribution. They show that ABC SMC performs better than the traditional ABC approach. Busetto and Buhmann [14] proposed a method for Bayesian parameter estimation based on new stable resampling technique for sequential Monte Carlo algorithm. They argue that this technique overcomes some drawbacks of classical SMC methods such as lack of stability and sample degeneracy.

5 Application to viral infection dynamics

In this section we give a simple example of construction and utilization of an ensemble model. Consider the following basic nonlinear model of acute viral infection:[55]

$$\dot{V} = pI - cV \tag{5.1}$$

$$\dot{H} = -\beta HV \tag{5.2}$$

$$\dot{I} = \beta HV - \delta I \tag{5.3}$$

where V is the concentration of viable virus particles, H is the number of uninfected target cells, and I is the number of infected cells. The virus particles interact with uninfected target cells which become infected at a rate βHV (here the parameter β is not to be confused with the inverse temperatures β_i). Free virus particles are cleared at a rate of c per day. The infected cells increase viral concentration at a rate of p per cell and die at a rate of δ per day. The initial conditions for the model are $(V, H, I)(0) = (H_0, V_0, 0)$. Baccam et al. [4] discuss the application of this model to modeling influenza A virus infection and calibrate it with virus titer data for individual human subjects.

The model can be readily extended to an ensemble model by assuming that the parameters are taken from a distribution - such a model then can be applied to cases in which the data are collected and combined for a group of subjects, such as in the study of the dynamics of virus infection in humans by Hayden et al. [32]. The study provides virus titer measurements from nasal wash expressed in TCID_{50} per ml of the wash for a group of 26 human volunteers inoculated intranasally with influenza A/Texas/91 (H1N1). The combined data are shown in Table 1.

t	$\log_{10} V$	σ
days	$\log_{10}(\text{TCID}_{50}/\text{ml})$	
1	0.95	0.62
2	2.67	0.94
3	2.67	0.94
4	1.90	0.67
5	0.81	0.52
6	0.65	0.42
7	0.48	0.47
8	0.30	0.21

Table 1. Averaged virus titer data for a group human volunteers inoculated intranasally with influenza A/Texas/91 (H1N1) [32].

Since the initial level of the virus, V_0, is not known in that study, it is included

in the list of parameters characterizing the model. The intial number H_0 of uninfected target cells is estimated at 4×10^8. The posterior distribution for $\mathbf{p} = (\log_{10} V_0, \log_{10} p, \log_{10} c, \log_{10} \beta, \log_{10} \delta)$ can be obtained using MCMC algorithm with parallel tempering as described in Section 4, by sampling 3 chains. For the prior distribution $\theta(\mathbf{p})$ we chose a product of uniform distributions for $\mathbf{p}$ of width 2 centered at the base line values $\bar{\mathbf{p}} = \log_{10}(0.25\,\mathrm{TCID}_{50}/\mathrm{ml},\ 0.014\,(\mathrm{TCID}_{50}/\mathrm{ml})^{-1},\ 2.7 \times 10^{-5}\,(\mathrm{TCID}_{50}/\mathrm{ml})\mathrm{day}^{-1},\ 3.2\,\mathrm{day}^{-1},\ 3.2\,\mathrm{day}^{-1})$ and for the proposal distributions q_i we chose uncorrelated multivariate Gaussian distributions for $\ln \mathbf{p}$ with variance matrices $\Sigma_i = 2\epsilon_i^2 I$, i.e.,

$$\theta(\mathbf{p}) = \begin{cases} 1/2^5 & p_i \in (\bar{p_i} - 1, \bar{p_i} + 1) \\ 0 & p_i \notin (\bar{p_i} - 1, \bar{p_i} + 1) \end{cases} \tag{5.4}$$

$$q_i(\hat{\mathbf{p}}|\mathbf{p}) = (2\pi\epsilon_i^2)^{-5/2} \exp\left(-\sum_{j=1}^{5} (\hat{p}_j - p_j)^2/(2\epsilon_i^2) \right) \tag{5.5}$$

(Note that the uniform prior on $\mathbf{p}$ corresponds to the uninformative Jeffreys prior on the original positive parameters.) For the likelihood function we use $L(Q|\mathbf{p}) = \exp(-E(\mathbf{p}, Q))$ with $E(\mathbf{p}, Q)$ as in (2.5), with the output variable $y = \log_{10} V$, and with data as in Table 1. The sample size obtained in this example is 900,000 parameter sets. Reasonable acceptance and swapping ratios were obtained for $(\beta_1, \beta_2, \beta_3) = (1, 0.33, 0.11)$ and $(\epsilon_1, \epsilon_2, \epsilon_3) = (0.087, 0.15, 0.26)$.

Marginal histograms of the posterior distribution $\rho(\mathbf{p}|Q)$ show that the data contain no information about the initial value V_0 (see Fig. 1A). The coefficient of infectivity β is localized at the center of the prior range while the value of p gravitates toward the lower end of the range. There is an interesting bimodality in the marginal distributions for the two degradation rates, c and δ. Further information about the distribution can be obtained from correlation plots in Fig. 1B, which show that there is a negative correlation between $\log_{10} \beta$ and $\log_{10} p$ and that the sample can be split into two clusters in the projection onto (c, δ) plane. It appears that each cluster is focused around a local minimum of the likelihood function.

The ensemble trajectories in Fig. 1C show that the model has trouble following the prescribed data (it does not reach the peak value of V and the decay is linear on the logarithmic scale) but that the variance of virus and uninfected target cell (H) trajectories is rather low. On the other hand, the variance over the ensemble in the trajectory of infected cells (I) is rather high.

To investigate the matter further, we examine the 500 best fitting trajectories (those with the largest value of $L(Q|\mathbf{p})$). These trajectories and the histograms of parameters for these trajectories are shown in Fig. 2. Clearly, the trajectories are almost indistinguishable in V and H, but they differ significantly in I. The trajectories with large maximum value of I correspond to the histograms shown in cyan and are characterized by a well defined value of δ and poorly defined c. The trajectories with small

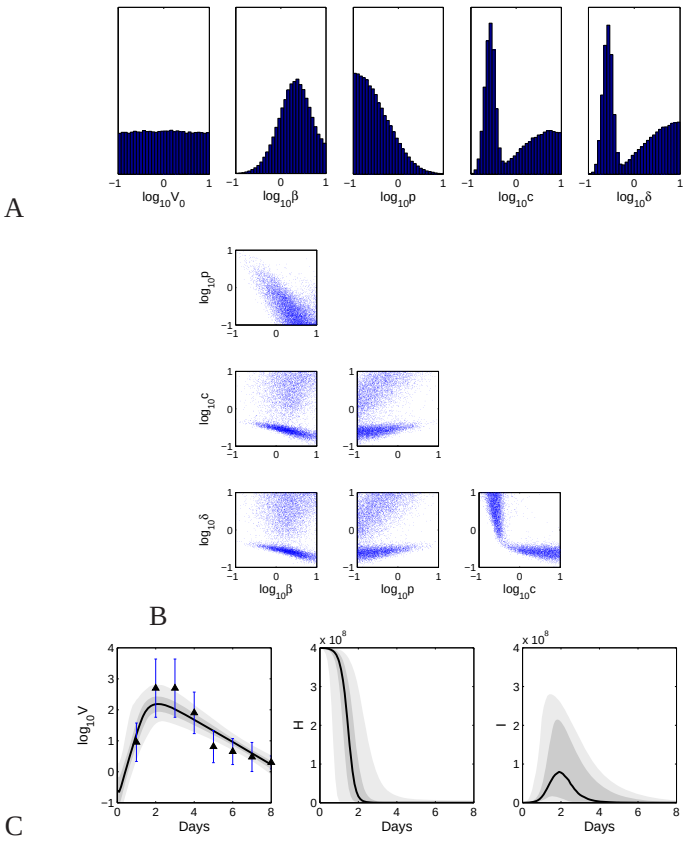

Figure 1. (A) Marginal distributions and (B) correlation plots for the posterior distribution, and (C) probabilistic trajectory prediction for the ensemble model obtained using (4.1). At each timepoint, the solid curve shows the median value of the variable over the ensemble, dark gray shows the 50-th percentile range, and light gray the 90-th percentile range. The data of [32], used to compute the posterior distribution, are shown as triangles.

maximum I correspond to the histograms shown in red and are characterized by well defined c and poorly defined δ. In the language of classical modeling, the data can be fitted equally well with multiple parameter sets that fall under two distinct categories. If the parameters were computed by maximization of the likelihood function, it is likely that one or the other optimum would have been missed.

Consider now the situation, also studied by Hayden et al. [32], in which an antiviral treatment is initiated about 30 hrs after the initial infection. Specifically, suppose that the subjects are administered a neuraminidase inhibitor, such as zanamivir or oseltamivir, which blocks the function of neuraminidase protein and prevents the virus

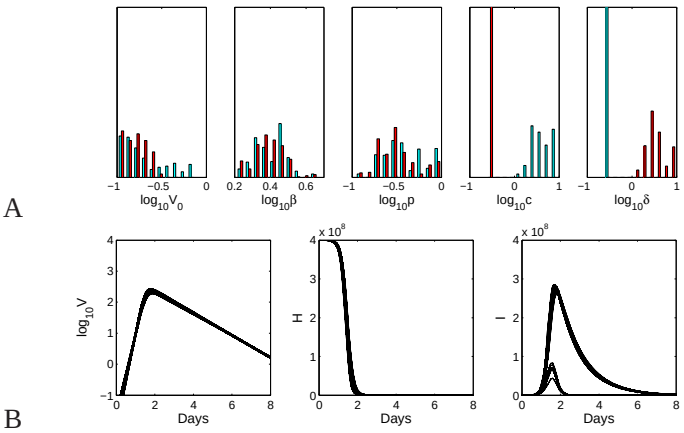

Figure 2. (A) Marginal distributions for two locally optimal clusters of parameters and (B) trajectories corresponding to the clusters. The trajectories with low maximum I are for the red distributions.

from budding from the host cell [28]. In the ensemble model this effect can be accounted for by lowering the value of the parameter p, describing the production of new virus cells, to $1/30^{\text{th}}$ of its starting value when $t > 1.25$ days for each parameter set in the sample. The resulting trajectories in Fig. 3 show significant decrease of virus levels after the administration of the treatment, with larger variance observed for the predicted virus level after treatment than the variances observed for the original model. The model predictions agree well with the observed data considering that $\log_{10} V = 0$ is the detection limit of the experiment. The other two variables also show larger variance, especially the trajectory of the uninfected target cells H. Although the trajectory of H was well constrained in the original model, the model does not allow us to make any conclusion about the number of uninfected cells after the treatment. This result, which would have been completely missed with any uniquely parametrized model, further illustrates the benefits of ensemble modeling.

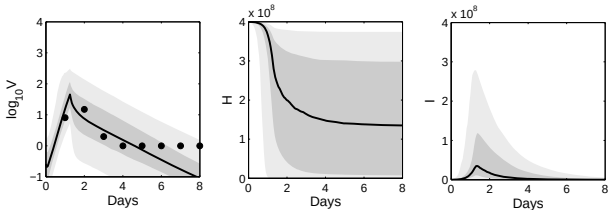

Figure 3. Prediction of the distribution of trajectories for the ensemble treated with neuraminidase inhibitor 30 hrs post infection. The data from [32] are shown as solid circles. The detection limit for $\log_{10} V$ is 0.

6 Ensemble models in biology

The number of papers utilizing ensemble modeling in biology is growing steadily, mostly in the the area of genetic and biochemical networks. For example, Battogtokh et al. [6] utilized the principles of ensemble modeling in a study of chemical reaction network for the regulation of the quinic acid (*qa*) gene cluster of *Neurospora crassa*. Their model consists of 14 differential equations for the mRNA and protein concentrations of 7 genes. The unknowns are 14 initial concentrations of both mRNA and proteins and 25 rate constants, that were fitted to the total of 42 data points obtained for mRNA concentrations. They pointed out that the poor constraining of parameters may be improved by collecting additional data for protein concentrations which were predicted by the model with broad distributions.

Putter et al. [63] used Bayesian approach to estimate parameters of HIV model of Griffith, May and Nowak. They argue for the use of Bayesian model because the data are collected from two compartments, one of which is subject to censoring, and random effects in one variable are assumed to be from beta distribution. They use a mix of informative (Gaussian in the log space) and non-informative prior distributions and estimate the pre-treatment and post-treatment reproductive ratios for the virus.

Brown et al. [12] used the ensemble technique to study genetic network modeling the action of of NGF and mitogenic epidermal growth factor (EGF) in rat pheochromocytoma (PC12) cells. They predict the influence of specific signaling modules in determining the integrated cellular response to the two growth factors. They note that only a small fraction of parameter combinations are well constrained and most parameters (rate constants) vary over huge ranges. However, the few well constrained parameters reveal critical features of the network that generates the appropriate output.

Kuepfer et al. [47] utilize ensemble modeling in the analysis of signaling networks. As opposed to the traditional approach, they vary not only the parameters of the model but also the model structure by including or excluding particular interaction terms. They develop a library of alternative dynamical models for TOR pathway of *S. cerevisiae*. They perform comparison of the models using Bayesian methods and point out that significant information about the pathway can be extracted using this procedure even with highly uncertain biological mechanisms and few quantitative experimental data.

Zenker, Rubin, and Clermont [78] present an example of the application of ensemble modeling and Bayesian inference to a disease diagnostic process. The output of the technique, based on simplified model of hypotension and patient specific clinical observation, is shown to produce a multimodal posterior density function, which peaks correspond to clinically relevant differential diagnoses. These can be constrained to a single diagnosis using additional observations from dynamical interventions.

Daun et al. [17] have employed an ensemble model to study acute inflammatory response to bacterial LPS in rats. The model they construct is not based on Bayesian inference but rather constructed by finding the maximum likelihood optimal set of

parameters starting from different initial guesses. Due to nonuniqueness of the maximum a sample of 103 points that fit the data equally well was obtained that fit the data equally well. Sensitivity analysis was used to reduce the 46 parameter space of the 8 state model to 18 parameters that was show to capture satisfactorily the essential dynamical properties of the full model.

Jayawardhana et al. [36] employed Bayesian inference to study the variability of steady states of metabolic pathways. The pathways are modeled using ODE systems, however, there is no data available about the dynamics of these pathways, only their steady states. The authors have adapted the ensemble technique to estimate the parameters distributions and steady state concentrations in the ensemble model based on the observed steady states of some state variables.

Klinke [45] uses Bayesian approach to calibrate a complex model of early EGF signaling comprised of 35 non-linear ODEs. Using experimentally determined dissociation constants he reduced the model parameters to 28 kinetic parameters and 6 initial concentrations, and determined maximum likelihood estimates for unknown parameters from experimental data using a simulated annealing optimization algorithm. The observed significant covariance between specific parameters and a broad range of variance is characteristic of a sloppy model.

Jaeger and Lambert [35] present a systematic study of Bayesian estimation of parameters for linear ODE system based on expansion of solutions of the ODE system in B-splines, which coefficients are adjusted to satisfy the system of ODEs. The benefit of such expansion, compared to the usual numerical integration, is that it is much less time consuming, independent of the inherent drawbacks associated with numerical integration, and that the posterior distribution can be given in a closed form. The B-spline approach also has its problems as poor spline fit can lead to poor parameter estimate.

Luan et al. [49] report on how ensemble models can be used to characterize response of a biological system to therapeutic interventions on the example of the response of human coagulation cascade to recombinant factor VIIa and prothrombin additions in normal and hemophilic plasma. They develop an ensemble of human coagulation models consisting of 193 state variables (protein concentrations) with 467 unknown parameters. The ensemble model in this work has not been created using Bayesian methods but rather by repeated energy minimization, which lead to different optima being found due to its nonuniqueness.

Ensemble techniques have been employed also in the study of human immune response to influenza A virus infection [61, 60] and a study of the vocal fold infection [69]. The influenza model [61, 60] consists of 20 nonlinear differential equations with 90 parameters that was calibrated with a collection of FACS, luminex, and ELISA assay data obtained for mice infected with the virus at two levels of the inoculum leading to either lethal or sublethal outcome. As each data point was collected from a different mice, ensemble modeling was essential for describing the parameter variability over the population. Using multiobjective fitting, a parameter distribution was constructed

that explained both types of trajectory paths. The study of vocal fold inflammation [69] contains a 4 variable model with 17 constants that is calibrated with sparse data. The ensemble modeling approach allowed to make probabilistic predictions about the effect of treatment strategies on the outcome of inflammation.

7 Conclusions

Notwithstanding the difficulties connected with implementation of ensemble models and parameter estimation techniques, it is clear that such models have their place in theoretical biology. They are easy to construct as straightforward extensions of existing stochastic or deterministic models, they are applicable to situations with small or extensive amount of data, and they provide probabilistically supported predictions of model behavior with clearly identified implications of all model assumptions.

Of course, in the growing area of ensemble modeling there are still many problems to be resolved and techniques to be developed. The utility and applicability of ensemble models depends on how accurately one is able to determine the parameter distribution $\rho(\mathbf{p})$ from the available data. Bayesian inference can provide an estimate of that distribution, however, the posterior distribution $\rho(\mathbf{p}|Q)$ contains not only information about the parameter variability, but also about parameter sensitivity, and the accuracy and completeness of data. Further analytical work is needed to find methods for deconvoluting the posterior distribution into individual components, provide convergence proofs, if possible, and criteria for estimating the amount of data needed to achieve accurate prediction of $\rho(\mathbf{p})$. Furthermore, since the posterior distribution includes information about parameter sensitivity, it should be possible to use that information to reduce the number of parameters needed to be estimated. This is important for computational efficiency because the number of points needed to obtain convergent MCMC sample grows exponentially with the dimension of the parameter space.[25]. More work is also needed to explore and design appropriate methods for parameter reduction of ensemble models. [68, 40]

Open problems remain also in Bayesian inference itself, such as, for example, the appropriate choice of the prior distribution. The presence of heuristic criteria in the prior distribution has proved to be important for controlling the space of admissible qualitative dynamical trajectories of the system.[33] As the parameters of nonlinear ODE systems are changed, such systems undergo bifurcations that qualitatively change their phase portraits. Such changes are usually undesirable in the ensemble model. Moreover, there are intuitive expectations about the behavior of a biology inspired model that are not captured by the available data, such as the requirement of non-explosion (existence and boundedness of trajectories for all time), or the requirement of a steady homeostatic state of an organism. Additionally, it would also be helpful to incorporate specifics of ODE solutions into Bayesian inference, specifically the dependence on noise on time, noise autocorrelation between data taken at neighboring timepoints, and possible noise correlation in multiresponse data. Exten-

sions of Bayesian approach from output error analysis to equation error or input error formulation would also be useful.

There is an alternative way to estimate parameter variability within population by using the hierarchical mixed-effects modeling approach (see, e.g., [10, 18, 57]. The authors estimate population parameters using maximum likelihood formulation and then estimate the standard deviation of the parameter variability from the data by linearizing the model about the maximum likelihood trajectory and inverse-mapping the standard deviations of the data. Such methods are convenient when variances in parameter values are small, while the key feature of Bayesian approach is that it provides global information about the parameter distribution.

Acknowledgments. Many thanks to G. Clermont and J. Rubin for numerous discussions on the subject and to S. Zenker for excellent lecture notes.

Bibliography

[1] B. Aleman-Meza, Y. Yu, H.B. Schüttler, J. Arnold and T.R. Taha, KINSOLVER: A simulator for computing large ensembles of biochemical and gene regulatory networks, *Computers & Mathematics with Applications* 57 (2009), 420–435.

[2] R.C. Aster, C.H. Thurber and B. Borchers, *Parameter estimation and inverse problems*, 90, Academic Press, 2005.

[3] E. Baake, M. Baake, HG Bock and KM Briggs, Fitting ordinary differential equations to chaotic data, *Physical Review A* 45 (1992), 5524–5529.

[4] P. Baccam, C. Beauchemin, C.A. Macken, F.G. Hayden and A.S. Perelson, Kinetics of influenza A virus infection in humans, *Journal of virology* 80 (2006), 7590.

[5] C. Barnes, D. Silk, X. Sheng and M.P.H. Stumpf, Bayesian design of synthetic biological systems, *Arxiv preprint arXiv:1103.1046* (2011).

[6] D. Battogtokh, DK Asch, ME Case, J. Arnold and H.B. Schüttler, An ensemble method for identifying regulatory circuits with special reference to the qa gene cluster of Neurospora crassa, *Proceedings of the National Academy of Sciences* 99 (2002), 16904.

[7] K.J. Beers, *Numerical methods for chemical engineering: applications in Matlab®*, Cambridge Univ Pr, 2007.

[8] J. Berger, The case for objective Bayesian analysis, *Bayesian Analysis* 1 (2006), 385–402.

[9] HG Bock, Recent advances in parameter identification for ordinary differential equations, *Progress in Scientific Computing* 2 (1983), 95–121.

[10] DM Bortz and PW Nelson, Model selection and mixed-effects modeling of HIV infection dynamics, *Bulletin of mathematical biology* 68 (2006), 2005–2025.

[11] G.E.P. Box and G.C. Tiao, *Bayesian inference in statistical analysis*, Wiley Online Library, 1973.

[12] K.S. Brown, C.C. Hill, G.A. Calero, C.R. Myers, K.H. Lee, J.P. Sethna and R.A. Cerione, The statistical mechanics of complex signaling networks: nerve growth factor signaling, *Physical Biology* 1 (2004), 184.

[13] K.S. Brown and J.P. Sethna, Statistical mechanical approaches to models with many poorly known parameters, *Physical review E* 68 (2003), 021904.

[14] A.G. Busetto and J.M. Buhmann, Stable Bayesian Parameter Estimation for Biological Dynamical Systems, in: *2009 International Conference on Computational Science and Engineering*, IEEE, pp. 148–157, 2009.

[15] P. Congdon, *Bayesian statistical modelling*, 670, Wiley, 2006.

[16] M.K. Cowles and B.P. Carlin, Markov chain Monte Carlo convergence diagnostics: a comparative review, *Journal of the American Statistical Association* (1996), 883–904.

[17] S. Daun, J. Rubin, Y. Vodovotz, A. Roy, R. Parker and G. Clermont, An ensemble of models of the acute inflammatory response to bacterial lipopolysaccharide in rats: results from parameter space reduction, *Journal of theoretical biology* 253 (2008), 843–853.

[18] M. Davidian and D.M. Giltinan, *Nonlinear models for repeated measurement data*, 62, Chapman & Hall/CRC, 1995.

[19] D.J. Earl and M.W. Deem, Parallel tempering: Theory, applications, and new perspectives, *Physical Chemistry Chemical Physics* 7 (2005), 3910–3916.

[20] D. Gamerman and H.F. Lopes, *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*, 68, Chapman & Hall/CRC, 2006.

[21] A. Gelman and D.B. Rubin, Inference from iterative simulation using multiple sequences, *Statistical science* (1992), 457–472.

[22] J. Geweke, *Contemporary Bayesian econometrics and statistics*, 537, Wiley-Interscience, 2005.

[23] ———, Bayesian model comparison and validation, *The American economic review* 97 (2007), 60–64.

[24] W.R. Gilks, S. Richardson and D.J. Spiegelhalter, *Markov chain Monte Carlo in practice*, Chapman & Hall/CRC, 1996.

[25] J. Gill, Is Partial-Dimension Convergence a Problem for Inferences from MCMC Algorithms?, *Political Analysis* 16 (2008), 153.

[26] P.J. Green, Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika* 82 (1995), 711.

[27] M. Guay and DD McLean, Optimization and sensitivity analysis for multi-response parameter estimation in systems of ordinary differential equations, *Computers and Chemical Engineering* 19 (1995), 1271–1286.

[28] L.V. Gubareva, Molecular mechanisms of influenza virus resistance to neuraminidase inhibitors, *Virus research* 103 (2004), 199–203.

[29] R.N. Gutenkunst, J.J. Waterfall, F.P. Casey, K.S. Brown, C.R. Myers and J.P. Sethna, Universally sloppy parameter sensitivities in systems biology models, *PLoS computational biology* 3 (2007), e189.

[30] U.H.E. Hansmann, Parallel tempering algorithm for conformational studies of biological molecules, *Chemical Physics Letters* 281 (1997), 140–150.

[31] W.K. Hastings, Monte Carlo sampling methods using Markov chains and their applications, *Biometrika* 57 (1970), 97.

[32] F.G. Hayden, J.J. Treanor, R.F. Betts, M. Lobo, J.D. Esinhart and E.K. Hussey, Safety and efficacy of the neuraminidase inhibitor GG167 in experimental human influenza, *JAMA: the journal of the American Medical Association* 275 (1996), 295.

[33] C. Higham, Bifurcation analysis informs Bayesian inference in the Hes1 feedback loop, *BMC systems biology* 3 (2009), 12.

[34] A.D. Hill, J.R. Tomshine, E. Weeding, V. Sotiropoulos and Y.N. Kaznessis, SynBioSS: the synthetic biology modeling suite, *Bioinformatics* 24 (2008), 2551.

[35] J. Jaeger and P. Lambert, Bayesian Generalized Profiling Estimation in Hierarchical Linear Dynamic Systems, (2010).

[36] B. Jayawardhana, D.B. Kell and M. Rattray, Bayesian inference of the sites of perturbations in metabolic pathways via Markov Chain Monte Carlo, *Bioinformatics* 24 (2008), 1191.

[37] E.T. Jaynes and G.L. Bretthorst, *Probability theory: the logic of science*, Cambridge Univ Pr, 2003.

[38] H. Jeffreys, *Scientific inference*, The University Press, 1937.

[39] ———, *Theory of Probability*, Oxford: Clarendon Press, 1961.

[40] B. Jin, Fast Bayesian approach for parameter estimation, *International Journal for Numerical Methods in Engineering* 76 (2008), 230–252.

[41] H.G. Katzgraber, S. Trebst, D.A. Huse and M. Troyer, Feedback-optimized parallel tempering Monte Carlo, *Journal of Statistical Mechanics: Theory and Experiment* 2006 (2006), P03018.

[42] D.B. Kell, Metabolomics and systems biology: making sense of the soup, *Current Opinion in Microbiology* 7 (2004), 296–307.

[43] C.T. Kelley, *Iterative methods for optimization*, 18, Society for Industrial Mathematics, 1999.

[44] H. Kitano, Systems biology: a brief overview, *Science* 295 (2002), 1662.

[45] D. Klinke, An empirical Bayesian approach for model-based inference of cellular signaling networks, *BMC bioinformatics* 10 (2009), 371.

[46] A. Kone and D.A. Kofke, Selection of temperature intervals for parallel-tempering simulations, *The Journal of chemical physics* 122 (2005), 206101.

[47] L. Kuepfer, M. Peter, U. Sauer and J. Stelling, Ensemble modeling for analysis of cell signaling dynamics, *Nature biotechnology* 25 (2007), 1001–1006.

[48] J. Liepe, C. Barnes, E. Cule, K. Erguler, P. Kirk, T. Toni and M.P.H. Stumpf, ABC-SysBio-approximate Bayesian computation in Python with GPU support, *Bioinformatics* 26 (2010), 1797.

[49] D. Luan, F. Szlam, K.A. Tanaka, P.S. Barie and J.D. Varner, Ensembles of uncertain mathematical models can identify network response to therapeutic interventions, *Mol. BioSyst.* (2010).

[50] D.J.C. MacKay, *Information theory, inference, and learning algorithms*, Cambridge Univ Pr, 2003.

[51] P. Marjoram, J. Molitor, V. Plagnol and S. Tavaré, Markov chain Monte Carlo without likelihoods, *Proceedings of the National Academy of Sciences of the United States of America* 100 (2003), 15324.

[52] F. Mazzocchi, Complementarity in biology, *EMBO reports* 11 (2010), 339.

[53] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller et al., Equation of state calculations by fast computing machines, *The journal of chemical physics* 21 (1953), 1087.

[54] H. Miao, X. Xia, A.S. Perelson and H. Wu, On identifiability of nonlinear ODE models with applications in viral dynamics, *SIAM Review: accepted* (2010).

[55] M.A. Nowak and R.M.C. May, *Virus dynamics: mathematical principles of immunology and virology*, Oxford University Press, USA, 2000.

[56] A. O'Hagan, J. Forster and M.G. Kendall, *Bayesian inference*, Arnold, 2004.

[57] J.C. Pinheiro and D.M. Bates, *Mixed-effects models in S and S-PLUS*, Springer Verlag, 2009.

[58] C. Predescu, M. Predescu and C.V. Ciobanu, The incomplete beta function law for parallel tempering sampling of classical canonical systems, *The Journal of chemical physics* 120 (2004), 4119.

[59] S.J. Press, *Subjective and objective Bayesian statistics: principles, models, and applications*, 328, LibreDigital, 2003.

[60] I. Price, *Mathematical modeling of chemical signals in inflammatory pathways*, Ph.D. thesis, University of Pittsburgh, 2011.

[61] I. Price, D. Swigon, B. Ermentrout, F. Toapanta, T. Ross and G. Clermont, Immune response to influenza A, *Respiratory Care Clinics of North America* 24 (2009), e33–e33.

[62] J.K. Pritchard, M.T. Seielstad, A. Perez-Lezaun and M.W. Feldman, Population growth of human Y chromosomes: a study of Y chromosome microsatellites., *Molecular Biology and Evolution* 16 (1999), 1791.

[63] H. Putter, SH Heisterkamp, JMA Lange and F. De Wolf, A Bayesian approach to parameter estimation in HIV dynamical models, *Statistics in Medicine* 21 (2002), 2199–2214.

[64] C.P. Robert and G. Casella, *Monte Carlo statistical methods*, Springer Verlag, 2004.

[65] A.F.M. Smith and G.O. Roberts, Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods, *Journal of the Royal Statistical Society. Series B (Methodological)* (1993), 3–23.

[66] E.D. Sontag, For differential equations with r parameters, 2r+ 1 experiments are enough for identification, *Journal of Nonlinear Science* 12 (2002), 553–583.

[67] M. Stephens, Bayesian analysis of mixture models with an unknown number of components-an alternative to reversible jump methods, *Annals of Statistics* (2000), 40–74.

[68] C. Sun and J. Hahn, Parameter reduction for stable dynamical systems based on Hankel singular values and sensitivity analysis, *Chemical engineering science* 61 (2006), 5393–5403.

[69] S. Tang, *Stochastic methods in modeling the immune response*, Ph.D. thesis, University of Pittsburgh, 2011.

[70] A. Tarantola, *Inverse problem theory and methods for model parameter estimation*, Society for Industrial Mathematics, 2005.

[71] T. Toni, D. Welch, N. Strelkowa, A. Ipsen and M.P.H. Stumpf, Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems, *Journal of the Royal Society Interface* 6 (2009), 187.

[72] S. Trebst, M. Troyer and U.H.E. Hansmann, Optimized parallel tempering simulations of proteins, *The Journal of chemical physics* 124 (2006), 174903.

[73] B. Van Domselaar and PW Hemker, *Nonlinear parameter estimation in initial value problems*, Stichting Mathematisch Centrum, 1975.

[74] M.H.V. Van Regenmortel, Reductionism and complexity in molecular biology, *EMBO reports* 5 (2004), 1016.

[75] H.U. Voss, J. Timmer and J. Kurths, Nonlinear dynamical system identification from uncertain and indirect measurements, *International Journal of Bifurcation and Chaos* 14 (2004), 1905–1933.

[76] V. Vyshemirsky and M. Girolami, BioBayes: a software package for Bayesian inference in systems biology, *Bioinformatics* 24 (2008), 1933.

[77] D.J. Wilkinson, Bayesian methods in bioinformatics and computational systems biology, *Briefings in bioinformatics* 8 (2007), 109.

[78] S. Zenker, J. Rubin and G. Clermont, From inverse problems in mathematical physiology to quantitative differential diagnoses, *PLoS computational biology* 3 (2007), e204.

Author information

David Swigon, 301 Thackeray Hall, Department of Mathematics, University of Pittsburgh, 15260, Pittsburgh, PA, U.S.A..
E-mail: swigon@pitt.edu