Bayesian parameter estimation for the Wnt pathway: an infinite mixture models approach

Konstantinos Koutroumpas, Paolo Ballarini, Irene Votsi, Paul-Henry Cournède

Bayesian parameter estimation for the Wnt pathway: An infinite mixture models approach.

Konstantinos Koutroumpas*[1], Paolo Ballarini[1], Irene Votsi[1] and Paul-Henry Cournède[1]

[1]Laboratory MICS, CentraleSupélec, University of Paris Saclay, Chatenay-Malabry, 92295, France

Abstract

Motivation: Likelihood–free methods, like Approximate Bayesian computation (ABC), have been extensively used in model–based statistical inference with intractable likelihood functions. When combined with Sequential Monte Carlo (SMC) algorithms they constitute a powerful approach for parameter estimation and model selection of mathematical models of complex biological systems. A crucial step in the ABC–SMC algorithms, significantly affecting their performance, is the propagation of a set of parameter vectors through a sequence of intermediate distributions using Markov kernels.

Results: In this paper we employ Dirichlet process mixtures (DPMs) to design optimal transition kernels and we present an ABC–SMC algorithm with DPM kernels. We illustrate the use of the proposed methodology using real data for the canonical Wnt signalling pathway. A multi–compartment model of the pathway is developed and it is compared to an existing model. The results indicate that DPMs are more efficient in the exploration of the parameter space and can significantly improve ABC–SMC performance. In comparison to alternative sampling schemes that are commonly used, the proposed approach can bring potential benefits in the estimation of complex multimodal distributions. The method is used to estimate the parameters and the initial state of two models of the Wnt pathway and it is shown that the multi–compartment model fits better the experimental data.

Availability: Python scripts for the Dirichlet Process Gaussian Muixture model and the Gibbs sampler are available at https://sites.google.com/site/kkoutroumpas/software.

Contact: konstantinos.koutroumpas@ecp.fr

1 Introduction

Mathematical modeling has become an integral part of modern biological research. With the onset of Systems Biology, mathematical models have been widely used to formally describe complex biological systems. Models enable the use of mathematical analysis and computer simulations not only to understand but also to predict and recently even control the behavior of biochemical processes (Milias-Argeitis *et al.*, 2011; Uhlendorf *et al.*, 2012). A challenging task during model building is the estimation of its parameters from experimental data. The equations describing the dynamics of a system depend on several parameters. For instance, chemical reactions are described by rate parameters, which capture how quickly a protein is synthesized or degraded or how fast a protein moves from one location to another in the cell. While some parameters can be measured using biochemical studies, chemical–kinetic models also include a substantial number of rate parameters that must be inferred from experimental data.

Parameter estimation approaches can be classified into two categories: frequentist and Bayesian. In the frequentist approach a point, or interval, estimation of model parameters is provided by maximizing the likelihood function

$$L(\boldsymbol{\theta}) = p(D^o|\boldsymbol{\theta}),$$

naturally defined as the probability that the observed data D^o have been generated by parameters $\boldsymbol{\theta}$.

Bayesian methods, on the other hand, have been proven a powerful alternative to frequentist methods, one in which parameters are treated as random variables with distributions attached to them (i.e. posterior distributions $p(\boldsymbol{\theta}|D^o)$). The overall goal in this case is to estimate the parameters' posterior distribution based on: 1) prior beliefs/knowledge on the unknown parameters, in form of the prior distribution $p(\boldsymbol{\theta})$, and 2) knowledge coming from the experimental data, in form of the likelihood function $L(\theta)$.

By relying on probability distributions, Bayesian methods do not require large sample approximations to

---

**To whom correspondence should be addressed. konstantinos.koutroumpas@ecp.fr

provide one with interval estimates but rather allow one to use prior beliefs on the unknown parameters, as expressed through the prior distribution $p(\boldsymbol{\theta})$. In other words, in a Bayesian context, the prior beliefs are updated by the observed data D^o via the likelihood function $p(D^o|\boldsymbol{\theta})$, and the posterior distribution is determined up to a normalized constant as follows

$$p(\boldsymbol{\theta}|D^o) \propto p(\boldsymbol{\theta})p(D^o|\boldsymbol{\theta}).$$

In several cases the calculation of the likelihood function $L(\boldsymbol{\theta})$ is difficult or even impossible, precluding the use of conventional likelihood–based inferential techniques. However, simulating from the likelihood is generally straightforward, even if obtaining a reliable numerical/functional representation of the model is not possible. In such settings, likelihood-free computation or approximate Bayesian computation (ABC) (Tavaré *et al.*, 1997; Beaumont *et al.*, 2002; Marjoram *et al.*, 2003; Sisson *et al.*, 2007) can be employed. ABC methods circumvent the explicit evaluation of the likelihood using a simulation–based approximation. The original ABC algorithm jointly simulates

$$\boldsymbol{\theta}' \sim p(\boldsymbol{\theta}) \quad \text{and} \quad D^s \sim p(D^s|\boldsymbol{\theta}')$$

and accept $\boldsymbol{\theta}'$ if and only if the simulated dataset D^s is close to the observed dataset.

The simple ABC scheme suffers from the same shortcamings as other rejection samplers: most of the samples are drawn from regions of parameter space, which cannot give rise to simulation outputs that resemple data. Therefore a number of computational schemes have been proposed to enhance the efficiency of ABC algorithms. ABC coupled to SMC samplers (Toni *et al.*, 2009) aim to sample sequentially from a sequence of distributions, which increasingly resemble the target posterior. The successive distributions are constructed by sampling parameters from the previous samples, and perturbing them according to a transition kernel (Del Moral *et al.*, 2006). The choice of the transition kernel has a significant impact on how accurate the approximate posterior is (Filippi *et al.*, 2013). In most cases a random–walk kernel is used to locally perturb a sample (Givens and Raftery, 1996; Filippi *et al.*, 2013). However, the choice of the transition kernel in an ABC context remains an open question, since the kernel influence is not formally understood and can be problem–dependent.

Here we revisit the problem of optimal transition kernel design using infinite mixture models. Finite mixture models have been previously used as importance densities in Adaptive Importance Sampling approaches (Cappé *et al.*, 2008; Wraith *et al.*, 2009; Cornuet *et al.*, 2012), with however a difficulty in choosing the proper number of elementary distributions in the mixture. We thus propose the use of infinite mixture models as successive important sampling distributions in the SMC framework, which will prove to bring very appealing results as detailed below. We formally describe how infinite mixture models and more precisely Dirichlet Process Mixtures (DPM) can be integrated in the ABC–SMC algorithm. We show that the ABC–SMC with DPM kernels is significantly more efficient in approximating complex, multimodal distributions and we illustrate its use in a model of the Wnt signalling pathway using real experimental data. Given the wide use of the ABC–SMC algorithm in parameter estimation and model selection and the improved performance of the proposed variant, we expect that it will significantly impact future modeling efforts in Systems biology.

The remainder of the paper is organized as follows. A short description of the ABC and ABC–SMC algorithms is provided in Section 2, followed by a description of the Dirichlet process Gaussian mixture model. We then present how DPGMM can be used as a transition kernel in the ABC–SMC framework. In Section 3 some biological background on the Wnt signalling pathway, the biological process we are interested in, is given accompanied by a mathematical model describing the main chemical reactions in the pathway. The methodology is applied both to a toy example and the Wnt pathway model in Section 4. Finally, some concluding remarks and future extensions are discussed in Section 5.

2 Methods

2.1 Approximate Bayesian Computation

Approximate Bayesian Computation (ABC) methods (Tavaré *et al.*, 1997; Pritchard *et al.*, 1999; Beaumont *et al.*, 2002) replace the calculation of the likelihood with direct sampling from the posterior $p(\boldsymbol{\theta}|D) \propto p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})$ through simulation. Given the prior distribution $p(\boldsymbol{\theta})$ on the parameter space, a candidate parameter vector $\boldsymbol{\theta}^*$ is sampled from $p(\boldsymbol{\theta})$ and a simulated data set D^s is generated from the likelihood function $p(D|\boldsymbol{\theta}^*)$. The parameter vector is accepted if the simulated data set is close to the observed data set D^o according to a distance function d $(d(D^o, D^s) \leq \epsilon)$. The distance function detrermines the discrepancy between the two datasets and ϵ is a tolerance value. If the data are too intricate or complicated it is common to replace a comparison of the observed and simulated data by a comparison of suitable summary statistics. An ABC rejection algorithm was used in Pritchard *et al.* (1999) for a model of the variation in human Y chromosome and its basic scheme follows Algorithm 1.

Input: Prior distribution $p(\boldsymbol{\theta})$, observed data set D^o, likelihood function $p(D|\boldsymbol{\theta})$, number of accepted particles N, tolerance ϵ

for $i = 1 \rightarrow N$ **do**

    Sample $\boldsymbol{\theta}^*$ from $p(\boldsymbol{\theta})$.

    Simulate a data set $D^s \sim p(D|\boldsymbol{\theta})$.

    while $d(D^o, D^s) > \epsilon$ **do**

        Sample $\boldsymbol{\theta}^*$ from $p(\boldsymbol{\theta})$.

        Simulate a data set $D^s \sim p(D|\boldsymbol{\theta})$.

    end

    Set $\boldsymbol{\theta}_i = \boldsymbol{\theta}^*$

end

Algorithm 1: ABC rejection algorithm

The output of the ABC rejection algorithm is a sample of parameter vectors from the distribution $p(\boldsymbol{\theta}|d(D^o, D^s) \leq \epsilon)$. In the case that ϵ is sufficiently small, the distribution $p(\boldsymbol{\theta}|d(D^o, D^s) \leq \epsilon)$ is considered a good approximation of the posterior distribution $p(\boldsymbol{\theta}|D^o)$. The acceptance rate of the ABC rejection algorithm depends on the proposal distribution $p(\boldsymbol{\theta})$ and it can become very low for a diffuse prior. Marjoram *et al.* (2003) proposed an ABC method based on Markov Chain Monte Carlo (ABC–MCMC) to overcome the low acceptance rate problem. Although the algorithm is guaranteed to converge to the approximate posterior distribution $p(\boldsymbol{\theta}|d(D^o, D^s) \leq \epsilon)$ while substantially improving the acceptance rate over the ABC rejection algorithm (Marjoram *et al.*, 2003), the samples in the generated Markov Chain are highly correlated and the chain may stuck in areas of low probability for long periods of time (Sisson *et al.*, 2007).

2.2 Sequential Monte Carlo for ABC

Algorithms combining ABC and Sequential Monte Carlo sampling have been proposed (Sisson *et al.*, 2007; Toni *et al.*, 2009) to avoid some of the disadvantages of the ABC–MCMC methods. To improve acceptance rates due to the mismatch between the proposal and the target distribution, sequential sampling methods (Del Moral *et al.*, 2006) define a sequence of intermediate distributions $p_1, \ldots, p_T$ that evolve gradually from a proposal distribution to the target distribution. The intermediate distributions are approximated by a set of random samples, the so–called particles. Starting with a distribution p_1 that is easy to approximate using the prior p, a number of particles, $\{\boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_N}\}$, are sampled from the prior and targeting p_1. The particles are then moved using a Markov kernel. It corresponds to sampling from an importance sampling distribution and the new particles are thus weighted with importance sampling weights (Toni *et al.*, 2009), taking into account how well a particle conforms to the next target distribution and the importance sampling distribution used to generate the samples. In the ABC framework the intermediate distributions are constructed by considering different tolerance values, *i.e.* $p_i = p(\boldsymbol{\theta}|d(D^o, D^s) \leq \epsilon^i)$ where $\epsilon^1 > \epsilon^2 > \ldots > \epsilon^T \geq 0$. In this study an ABC–SMC algorithm (Toni *et al.*, 2009) based on sequential importance sampling (Del Moral *et al.*, 2006) is used and its scheme for a deterministic system is shown in Algorithm 2.

Apart from the increased efficiency in comparison to the ABC rejection algorithm, ABC–SMC does not stuck in low probability regions and the particles are not correlated, which is the case with ABC–MCMC (Sisson *et al.*, 2007). Moreover, given that a population of particles is used, the method is more efficient when complex posteriors are considered (Sisson *et al.*, 2007). However, the performance of the algorithm clearly depends on the selected tolerance sequence. Fast decreasing tolerances can result in low acceptance rates and poor performance, while if they decrease slowly the algorithm becomes computationally intensive (Del Moral *et al.*, 2012). For the optimal selection of the tolerance sequence, Del Moral *et al.* (2012) proposed an adaptive ABC–SMC algorithm. The kernels K_j that are used to draw new particles starting from the previous ones also affect ABC–SMC efficiency. Being easy to implement, local random–walk moves (Gaussian or uniform) are usually employed (Sisson *et al.*, 2007; Toni *et al.*, 2009). The parameters of the kernels are determined by some statistics on the previous sample population. Such approaches can be problematic especially in the case of complex multimodal posteriors (Filippi *et al.*, 2013). Filippi *et al.* (2013) have derived optimality criteria based on the Kullback–Leibler divergence between a target distribution and its estimate. They have also showed that for complicated posterior distributions, locally adapted kernels that are designed taking into account the K–nearest neighbors for each particle tend to show the best performance. The main drawback of this approach is that the number K has to be fixed *a*

Input: Prior distribution $p(\boldsymbol{\theta})$, an observed data set D^o, a likelihood function $p(D|\boldsymbol{\theta})$, number of accepted particles N, a tolerance schedule $\epsilon^1 > \ldots > \epsilon^T \geq 0$

for $j = 1 \rightarrow T$ **do**

  if $j == 1$ **then**

    for $i = 1 \rightarrow N$ **do**

      Sample $\boldsymbol{\theta}^*$ from $p(\boldsymbol{\theta})$.

      Simulate a data set $D^s \sim p(D|\boldsymbol{\theta^*})$.

      while $d(D^o, D^s) > \epsilon^j$ **do**

        Sample $\boldsymbol{\theta}^*$ from $p(\boldsymbol{\theta})$.

        Simulate a data set $D^s \sim p(D|\boldsymbol{\theta^*})$.

      end

      Set $\boldsymbol{\theta}_i^j = \boldsymbol{\theta}^*$;

      $w_i^j = 1$

    end

    Normalize the weights w_j^i

  else

    for $i = 1 \rightarrow N$ **do**

      Sample $\boldsymbol{\theta^*}$ from $\sum_{i=1}^{N} w_i^{j-1} K(\boldsymbol{\theta}|\boldsymbol{\theta}_i^{j-1})$.;

      Simulate a data set $D^s \sim p(D|\boldsymbol{\theta^*})$.

      while $d(D^o, D^s) > \epsilon^j$ **do**

        Sample $\boldsymbol{\theta^*}$ from $\sum_{i=1}^{N} w_i^{j-1} K(\boldsymbol{\theta}|\boldsymbol{\theta}_i^{j-1})$.;

        Simulate a data set $D^s \sim p(D|\boldsymbol{\theta^*})$.

      end

      Set $\boldsymbol{\theta}_i^j = \boldsymbol{\theta}^*$;

      $w_i^j \propto \frac{p(\boldsymbol{\theta}_i^j)}{\sum_{i=1}^{N} w_i^{j-1} K(\boldsymbol{\theta}_i^j|\boldsymbol{\theta}_i^{j-1})}$.

    end

    Normalize the weights w_i^j.

  end

end

Algorithm 2: ABC–SMC algorithm

priori and its value affects the exploration of the parameter space (Filippi *et al.*, 2013). For this reason, we propose to use a Dirichlet Process Gaussian Mixture Model to generate the proposal distributions at each iteration of the SMC-ABC scheme.

2.3 Dirichlet Process Gaussian Mixture Model

Gaussian mixtures (GM) are a class of parametric mixture models that are used for density estimation (McLachlan and Peel, 2004). Cappé *et al.* (2008), Wraith *et al.* (2009) and Cornuet *et al.* (2012) have proposed the use of GM's as importance densities in Adaptive Importance Sampling approaches. As with the K–nearest neighbors approach in (Filippi *et al.*, 2013), the number of components has to be pre–selected and in most approaches model selection is used to select the optimal number of components. A workaround to this problem is the use of a non–parametric Infinite Mixture model that assumes that the data come from a mixture of an infinite number of distributions. In the following we consider the Dirichlet Process Mixture (DPM) model (Antoniak, 1974) and more precisely the Dirichlet Process Gaussian Mixture model (DPGMM) which has been extensively used for density estimation and data clustering (Müller *et al.*, 1996; Rasmussen, 1999; Escobar, 1994; MacEachern, 1994; Escobar and West, 1995). Starting with a finite GM, we can obtain a DPGMM by taking the limit as the number of components goes to infinity.

A GM with K components can be written as:

$$p(\boldsymbol{x} | (\pi_1, \boldsymbol{\mu}_1, \Sigma_1), \ldots, (\pi_K, \boldsymbol{\mu}_K, \Sigma_K)) = \sum_{c=1}^{K} \pi_c \mathcal{N}(\boldsymbol{\mu_c}, \Sigma_c) \tag{1}$$

where π_c, $\boldsymbol{\mu}_c$ and Σ_c are respectively the mixing proportion, the mean vector and the covariance matrix for component c respectively. If we assume a Dirichlet prior on the mixing proportions and a prior distribution H on the mean and the covariance, the model can be written:

$$\begin{aligned}
\boldsymbol{x}_i | z_i, (\mu_1, \Sigma_1), \ldots (\mu_K, \Sigma_K) &\sim& \mathcal{N}(\boldsymbol{\mu}_{z_i}, \Sigma_{z_i}) \\
z_i | \pi_1, \ldots, \pi_K &\sim& \mathrm{Cat}(\pi_1, \ldots, \pi_K) \\
\pi_1, \ldots, \pi_K | \alpha &\sim& \mathrm{Dir}(\alpha/K, \ldots, \alpha/K) \\
(\boldsymbol{\mu}_i, \Sigma_i) &\sim& H
\end{aligned} \tag{2}$$

where z_i is the indicator variable and corresponds to the component to which the observation x_i belongs. Given a set of observations $\{x_1, x_2, \ldots, x_N\}$ the probability of an observation to belong to one of the K components is (Görür and Rasmussen, 2010):

$$p(z_l = j | \boldsymbol{z}_{\setminus l}, \alpha) = \frac{C_{\setminus l, j} + \alpha/K}{N - 1 + \alpha} \tag{3}$$

where $C_{\setminus l, k} = \sum_{i=1, \ldots, N i \neq l} \delta_k(z_i)$ is the number of observations except l that belong to component j. Taking the limit as K goes to infinity the probability that an observation belongs to a non–empty component (*i.e* components with $C_{\setminus l, j} > 0$) is:

$$p(z_l = j | \boldsymbol{z}_{\setminus l}, \alpha) = \frac{C_{\setminus l, j}}{N - 1 + \alpha} \tag{4}$$

and the probability to belong to one of the infinitely many empty components is:

$$p(z_l \neq z_k \ \forall \ k \neq l | \boldsymbol{z}_{\setminus l}, \alpha) = \frac{\alpha}{N - 1 + \alpha}. \tag{5}$$

Probabilities (4-5) are the same with the probabilities in a Chinese restaurant process that is used to define a Dirichlet process. Hence, the infinite limit of a GMM (2) is a DPGMM (Görür and Rasmussen, 2010) which can be compactly written:

$$\begin{aligned}
\boldsymbol{x}_i | (\boldsymbol{\mu}_i, \Sigma_i) &\sim& \mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i) \\
(\boldsymbol{\mu}_i, \Sigma_i) &\sim& G \\
G &\sim& \mathrm{DP}(H, \alpha).
\end{aligned} \tag{6}$$

where $\mathrm{DP}(H, \alpha)$ is the Dirichlet process with concentration α and base distribution H.

The base distribution H is the mean of the DP, which means that the DP draws distributions "around" H as a normal distribution draws real numbers around its mean. A strong property of DP is that the distributions drawn are almost surely discrete even if the base distribution is continuous. The concentration parameter α specifies how strong the discretization is. For an α close to 0, the distributions are concentrated on a single value, while as α goes to ∞ the realizations are close to H. The component

parameters are then sampled from the distribution drawn from the DP. Both the choice of the base distribution and the value of the concentration parameter are important for the performance of the model. While the concentration parameter can sometimes be inferred from the data, it is harder to decide on the base distribution. In most cases the base distribution is chosen so as to facilitate the inference of the DPGMM model.

2.4 DPGMM estimation

DPM models are usually inferred using a MCMC algorithm (Neal, 2000). Most methods are based on Gibbs sampling, where variables (parameters, hyperparameters and indicator variables) are updated in turn by sampling from their posteriors conditional on the other variables. In the case that the base distribution is the conjugate prior of the probability distribution of the observations generated from a component, Gaussian in our case, a collapsed Gibbs sampler (MacEachern, 1994; Sudderth, 2006) is more efficient than explicitly sampling parameters. In this case, the state of the Markov chain constructed from this sampler consists of the indicator variables only that are updated by iteratively reassigning each observation to a new component based on the assignments of the remaining observations. Consider a set of observations $\boldsymbol{x} = \{x_1, x_2, \ldots, x_N\}$ from a GM (1) with unknown number of components K. The probability of an observation l to belong to a component j is:

$$p(z_l = j | \boldsymbol{z}_{\backslash l}, \boldsymbol{x}, \alpha, \kappa, \theta, \nu, \Delta) \quad = \quad p(z_l = j | \boldsymbol{z}_{\backslash l}, \alpha) \tag{7}$$
$$p(x_l | z_l = j, \boldsymbol{z}_{\backslash l}, \boldsymbol{x}_{\backslash l}, \kappa, \theta, \nu, \Delta)$$

where $p(z_l = j | \boldsymbol{z}_{\backslash l}, \alpha)$ is given by (4) or (5) and $p(x_l | z_l = j, \boldsymbol{z}_{\backslash l}, \boldsymbol{x}_{\backslash l}, \kappa, \theta, \nu, \Delta)$ can be seen as the likelihood of x_l to have been generated from the same Gaussian with the observations that already belong to the component j. The latter can be approximated by a Gaussian with mean and covariance computed by the observations belonging to the component j (Sudderth, 2006). The Gibbs sampler starts by assigning all observations to a single component, and iteratively reassigns each observation to a new component based on the assignments of the remaining observations according to (7).

In the case of DPGMM a natural selection for H is the Normal–Inverse–Wishart distribution, which is the conjugate prior of normal distribution. The covariance matrix Σ is drawn from an inverse–Wishart distribution:

$$\mathcal{W}^{-1}(\Sigma | \nu, \Delta) \propto |\Sigma|^{\left(-\frac{\nu+d+1}{2}\right)} \exp\left\{-\frac{1}{2} tr(\nu \Delta \Sigma^{-1})\right\} \tag{8}$$

where Δ is the covariance parameter, ν the degrees of freedom and d the dimension. Conditioned on Σ the mean $\boldsymbol{\mu}$ is drawn from a Normal distribution

$$\mathcal{N}(\theta, \Sigma/\kappa). \tag{9}$$

A detailed description of the collapsed Gibbs sampler that was used for this study can be found in Section 2.5.3, Algorithm 2.3 in Sudderth (2006). For the concentration parameter α an inverse gamma prior was used as proposed in (Rasmussen, 1999).

2.5 ABC–SMC using DPGMM kernels

In the ABC–SMC framework, DPGMM can be used to fit a mixture on the samples of each population and new samples can be drawn from the identified Gaussian components. More precisely, assume that $\boldsymbol{\Theta^{j-1}} = \{\boldsymbol{\theta_k^{j-1}}\}_{k=1}^N$ are the samples accepted during iteration $j-1$ of the ABC–SMC algorithm, $\boldsymbol{W^{j-1}} = \{w_k^{j-1}\}_{k=1}^N$ the corresponding weights, and $\boldsymbol{z^{j-1}} = \{z_k^{j-1}\}_{k=1}^N$ the indicator variables estimated using the previously described Gibbs sampler. Samples for the next iteration (lines 16 and 19 in Algorithm 2) can be directly drawn from:

$$\sum_{c=1}^{K^{j-1}} \hat{\pi}_c \mathcal{N}(\boldsymbol{\mu_c^{j-1}}, \Sigma_c^{j-1}) \tag{10}$$

where K^{j-1} is the number of non–empty components of the DPGMM and

$$\hat{\pi}_c = \sum_{k \in C_c^{j-1}} w_k^{j-1} \tag{11}$$

the sum of the importance weights of the particles $C_c^{j-1} = \{k = 1, \ldots, N | z_k^{j-1} = c\}$ that belong to the component.

It is important to note that the sampling weights are also considered in the calculation of the parameters of each component, *i.e.* the mean and the covariance. The mean of a component c is given by the weighted mean:

$$\boldsymbol{\mu_c^{j-1}} = \sum_{k \in C_c^{j-1}} w_k^{j-1} \boldsymbol{\theta_k^{t-1}} \tag{12}$$

and the covariance matrix by the weighted covariance matrix:

$$\Sigma_c^{j-1} = \frac{V_c^{j-1}}{\left(V_c^{j-1}\right)^2 - \mathcal{V}_c^{j-1}} \sum_{k \in C_c^{j-1}} \left[w_k^{j-1} \left(\boldsymbol{\theta_k^{j-1}} - \boldsymbol{\mu_c^{j-1}}\right)^T \left(\boldsymbol{\theta_k^{j-1}} - \boldsymbol{\mu_c^{j-1}}\right) \right] \tag{13}$$

where V_c^{j-1} and $\mathcal{V}_c^{j-1}$ are the sum of the weights and the sum of the squares of the weights.

$$V_c^{j-1} = \sum_{k \in C_c^{j-1}} w_k^{j-1} \tag{14}$$

$$\mathcal{V}_c^{j-1} = \sum_{k \in C_c^{j-1}} \left(w_k^{j-1}\right)^2 \tag{15}$$

3 The Wnt pathway

Wnt pathways are a family of *intracellular signalling transduction* pathways known for playing a key role in cell's development (e.g. cell proliferation, stem cell maintenance, differentiation, cardiac development). They are the subject of intensive studies both in relation to cancer development as well as to embryonic development pathologies (Fleiss *et al.*, 2015). Two families of Wnt pathways have been, so far, characterized according to their dependency on the multifunctional β-catenin protein (Kühl and Rao, 2010). *Canonical* Wnt pathways (Wnt/β-catenin) regulate gene transcription via β-catenin mediated transduction of Wnt signal, whereas *non-canonical* Wnt pathways, like Wnt/Ca^{2+} (De, 2011), are involved in the regulation of a cell's physical characteristics, *i.e.* the cell shape, and its functioning. All Wnt pathways share the same activation mechanism: the signal is triggered through the binding of a Wnt protein to a receptor (of the Frizzled family) on the cell's membrane, and propagated inside the cell through the cytoplasmic *dishevelled* (Dsh) proteins. In this paper we consider a model of the canonical Wnt pathway.

Wnt/β-catenin signalling (basics). Signal transduction along the Wnt/β-catenin pathway is initiated when extracellular Wnt proteins bind to receptors of the Frizzled family and to coreceptors such as the lipoprotein receptor–related protein (LRP5/6). It triggers a cascade of events that involve a number of *actors* both in the cytoplasm and in the nucleus of the cell. The core species known to be involved in the pathway include (Table 1): the Dishvelled (Dsh) protein, the β-catenin protein, the Axin protein, the adenomatous polyposis coli (APC), the glycogen synthase kinase 3 (GSK-3β), the multimeric protein complex APC/GSK3β/Axin (given by complexation of APC, Axin and GSK-3β) also called the *destruction complex* (DC), and at nucleic level, the T–Cell factor (TCF) and the lymphocyte enhancing factor (LEF) which yield gene expression.

In the absence of extracellular Wnt molecules, β–catenin is destructed by the degradation complex, which forms around the activated form of Axin (AxinP), in the cytosol. When Wnt binds to LRP at the membrane, the assembly of the destruction complex is inhibited, leading to an accumulation of β–catenin in the cytosol. Consecutively, β–catenin is relocated to the nucleus where it activates the transcription of genes including the gene encoding for Axin protein.

Models of canonincal Wnt pathway. Many computational models of the Wnt/β-catenin pathway have been proposed (Mazemondet *et al.*, 2012; MacLean *et al.*, 2016; Jensen *et al.*, 2010; Schmitz *et al.*, 2013; Lee *et al.*, 2003; Tan *et al.*, 2014). They may be distinguished with respect to the biological features they capture, for example whether they consider the compartmentalization of the cell, or the granularity with which molecular activities are represented (MacLean *et al.*, 2016). Some of the models (Tan *et al.*, 2014; Schmitz *et al.*, 2013; Mazemondet *et al.*, 2012) account for the cytoplasm/nucleus relocation of some of the species. In Tan *et al.* (2014) the dynamics of β-catenin compartmental diffusion have been investigated through a double-compartment ODEs model of Wnt-pathway which has been validated through time series measurements on Humans Epithelial Kidney (HEK293T) cells. In Jensen *et al.* (2010) and Mazemondet *et al.* (2012) the negative feedback loop formed through the regulation of Axin expression by nuclear β–catenin is modeled explicitly. In Jensen *et al.* (2010) the Wnt induced oscillatory

Species			
	Acronym	var. ID	description
Monomers	Dsh$_a$	x_1	active dishevelled
	Dsh$_i$	x_2	inactive dishevelled
	GSK3	x_5	glycocen synthase kinase
	APC	x_7	adenomatous polyposis coli
	β-cat*	x_{10}	phosporylated beta catenin
	β-cat	x_{11}	beta catenin
	Axin	x_{12}	axin protein
	TCF	x_{13}	T-Cell factor
Polymers	APC*/Axin*/GSK3	x_3	complex
	APC/Axin/GSK3	x_4	complex
	APC/Axin	x_6	complex
	APC*/Axin*/GSK3/β-cat	x_8	complex
	APC*/Axin*/GSK3/β-cat*	x_9	complex
	TCF/β-cat	x_{14}	complex
	APC/β-cat	x_{15}	complex

Table 1: Core species of Lee Wnt/β-catenin pathway model

dynamics of a β-catenin/Axin negative feedback loop have been unraveled in ODEs models of somatogenesis. Most of the existing Wnt pathway models are based on the seminal work by Lee *et al.* (2003), who proposed an ODE, mono–compartment, *core* model of canonical Wnt pathway, validated through *in-vitro* measurement on the eggs of the *Xenopus* frog. Because of the amount of wet-lab experimental data, to date, the Lee model represents the most accurate/complete effort in Wnt pathway modelling. In the context of this paper, we also focus on the Lee model and present some numerical experiments (Section 4) performed on it. We proceed below with a detailed description of the Lee model.

3.1 The *Lee* model.

The model consists of 6 core species (Dsh, Axin, β-catenin, TCF, APC and GSK3) and their basic interactions, (i.e. synthesis, degradation, phosphorylation, reversible complexations). The overall dynamics of the model are described by 15 *mass action kinetics* ordinary differential equations (16–29). In the Lee *et al.* (2003) model, the existence of intracellular compartments (i.e. cytosol, nucleus) is abstracted away, hence all species are assumed to be contained in a single cellular compartment. Furthermore, the Wnt signal is not explicitly represented, rather it is assimilated to the state of the dishevelled (Dsh) protein (active Dsh corresponding to the presence of Wnt signal on the membrane receptors, inactive Dsh corresponding to its absence).

In absence of Wnt signal (Dsh$_i$), the DC (Axin*/APC*/GSK3) is free to form through a multistep reaction, consisting of: first the reversible dimerization of Axin with APC (APC/Axin, binding at rate proportional to α_8, unbinding at rate proportional to α_9), then the reversible formation of the Axin/APC/GSK3 trimer (binding at rate α_6, unbinding at rate proportional to α_7, if Dsh is inactive, or to α_5 if Dsh is active), followed by the reversible phosphorylation of Axin and APC parts of the Axin/APC/GSK3 trimer to obtain the actual DC (Axin*/APC*/GSK3). The DC can then reversibly sequester free β-catenin molecules (sequestration at rate proportional to α_{10}, liberation at rate proportional to α_{11}). β-catenin sequestration by the DC eventually leads to β-catenin degradation (at rate proportional to α_{14}) through an intermediate, irreversible, phosphorylation step (phosporylation at rate proportional to α_{12}). Notice that the decomplexation of the unphosphohorylated form of DC (i.e. the Axin/APC/GSK3 complex) is regulated by the active form of Dsh (i.e. Dsh$_a$): in presence of Dsh$_a$ the decomplexation of Axin/APC/GSK3 is enhanced by a rate which is proportional to the level of Wnt signal (i.e. the amount of Dsh$_a$) as well as to the constant rate α_5. The model also accounts for constant production and degradation of both β-catenin (produced at constant rate α_{15}, degraded at rate proportional to α_{16}) and Axin (produced at constant rate α_{17}, degraded at rate proportional to α_{18}). Activation/inactivation of Dsh (corresponding to the binding/unbinding of Wnt molecules to the cell's membrane) are assumed to occur at rate proportional to α_1, respectively α_2.

$$R_1 \quad : \quad \text{Dsh}_i \xrightleftharpoons[\alpha_2]{\alpha_1} \text{Dsh}_a$$

$$R_2 \quad : \quad \text{APC*/Axin*/GSK3} \xrightleftharpoons[\alpha_3]{\alpha_4} \text{APC/Axin/GSK3}$$

$$R_3 \quad : \quad \text{APC/Axin} + \text{GSK3} \xrightleftharpoons[\alpha_5[\text{Dsh}_a]+\alpha_7]{\alpha_6} \text{APC/Axin/GSK3}$$

$$R_4 \quad : \quad \text{APC} + \text{Axin} \xrightleftharpoons[\alpha_9]{\alpha_8} \text{APC/Axin}$$

$$R_5 \quad : \quad \text{APC*/Axin*/GSK3} + \beta\text{-cat} \xrightleftharpoons[\alpha_{11}]{\alpha_{12}} \text{APC*/Axin*/GSK3}/\beta\text{-cat}$$

$$R_6 \quad : \quad \text{APC*/Axin*/GSK3}/\beta\text{-cat} \xrightarrow{\alpha_{12}} \text{APC*/Axin*/GSK3}/\beta\text{-cat*}$$

$$R_7 \quad : \quad \text{APC*/Axin*/GSK3}/\beta\text{-cat*} \xrightarrow{\alpha_{13}} \text{APC/Axin/GSK3}/ + \beta\text{-cat*}$$

$$R_8 \quad : \quad \beta\text{-cat*} \xrightarrow{\alpha_{14}} \emptyset$$

$$R_9 \quad : \quad \emptyset \xrightarrow{\alpha_{15}} \beta\text{-cat}$$

$$R_{10} \quad : \quad \beta\text{-cat} \xrightarrow{\alpha_{16}} \emptyset$$

$$R_{11} \quad : \quad \emptyset \xrightarrow{\alpha_{17}} \text{Axin}$$

$$R_{12} \quad : \quad \text{Axin} \xrightarrow{\alpha_{18}} \emptyset$$

$$R_{13} \quad : \quad \beta\text{-cat} + TCF \xrightleftharpoons[\alpha_{20}]{\alpha_{19}} \beta\text{-cat}/TCF$$
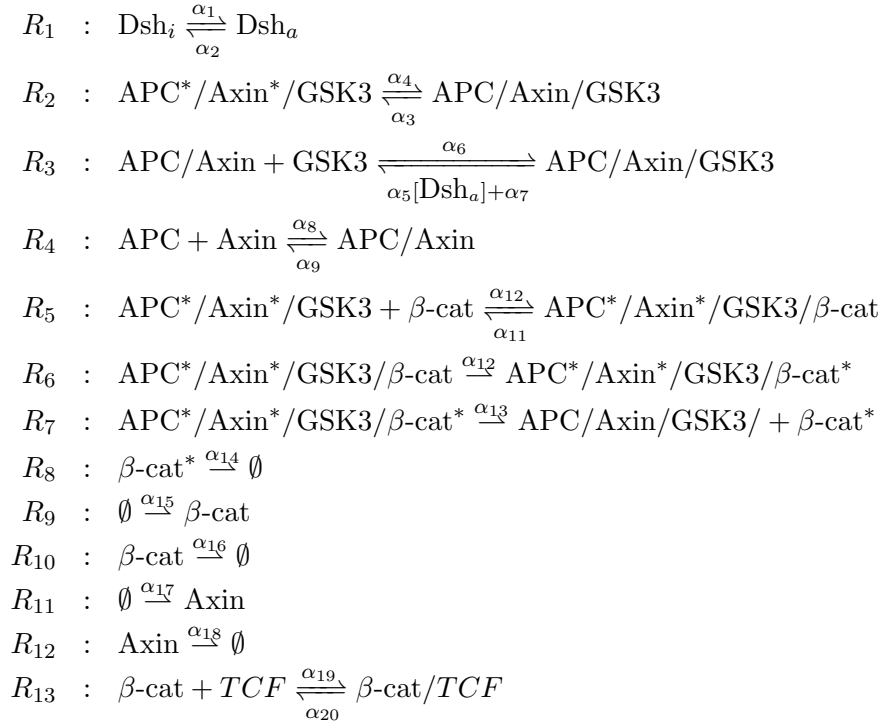
Figure 1: Chemical reactions in the Lee *et al.* (2003) model.

3.2 A compartmentalized *Lee* model

An important feature of the Lee *et al.* (2003) model to study the role of the Wnt pathway in diseases is that it focuses on the assembly of the DC from its constituent parts. While available experimental data are not adequate to fully characterize the dynamics of DC assembly, modelling these reactions enables the prediction of system's behavior under mutations. On the other hand, the most important lack of the model is that it does not distinguish between the nucleus and cytoplasm. Mazemondet *et al.* (2012) developed a model based on the Lee *et al.* (2003) model that distinguishes the two compartments and also includes the β–catenin/Axin feedback loop. However, the specific model simplifies the DC assembly procedure. Here, we combine the two models.

The new model consists of all the chemical reactions in the Lee *et al.* (2003) model except for R_{13} (Fig. 1) which is replaced by the reaction

$$R'_{13} \quad : \quad \beta\text{-cat}_n + TCF \xrightleftharpoons[\alpha_{20}]{\alpha_{19}} \beta\text{-cat}_n/TCF$$

where β-cat$_n$ corresponds to the β–cateninin molecules localized in the nucleus. The model is completed with one reaction for the β–cateninin transportation

$$R'_{14} \quad : \quad \beta\text{-cat} \xrightleftharpoons[\alpha_{24}]{\alpha_{23}} \beta\text{-cat}_n$$
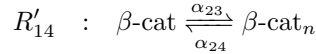
and one for the regulation of the expression of Axin by the β-cat$_n$/TCF complex

$$R'_{15} \quad : \quad \beta\text{-cat}_n/TCF \xrightarrow{\alpha_{25}} \beta\text{-cat}_n/TCF + Axin.$$
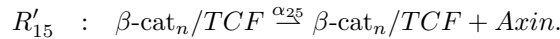
The differential equations for Axin and β–catenin become

$$\frac{dx_{11}}{dt} = -\alpha_{10}x_{11}x_3 + \alpha_{11}x_8 + \alpha_{15} - \alpha_{16}x_{11} - \alpha_{21}x_{11}x_7$$
$$+\alpha_{22}x_{15} - \alpha_{23}x_11 + \alpha_{24}x_{16} \tag{30}$$

$$\frac{dx_{12}}{dt} = -\alpha_8 x_7 x_{12} + \alpha_9 x_6 + \alpha_{25}x_{16} - \alpha_{18}x_{12} \tag{31}$$

where x_{16} denotes the concentration of β–catenin in the nucleus governed by the following differential equation:

$$\frac{dx_{16}}{dt} = -\alpha_{19} * x_{16} * x_{13} + \alpha_{20} * x_{14} + \alpha_{23}x_{11}$$
$$-\alpha_{24}x_{16} \tag{32}$$

9

$$\frac{dx_1}{dt} = -\alpha_1 x_1 + \alpha_2 x_2 \tag{16}$$

$$\frac{dx_2}{dt} = \alpha_1 x_1 - \alpha_2 x_2 \tag{17}$$

$$\frac{dx_3}{dt} = \alpha_3 x_4 - \alpha_4 x_3 - \alpha_{10} x_{11} x_3 + \alpha_{11} x_8 + \alpha_{13} x_9 \tag{18}$$

$$\frac{dx_4}{dt} = -\alpha_5 x_2 x_4 - \alpha_3 x_4 + \alpha_4 x_3 + \alpha_6 x_5 x_6 - \alpha_7 x_4 \tag{19}$$

$$\frac{dx_5}{dt} = \alpha_5 x_2 x_4 - \alpha_6 x_5 x_6 + \alpha_7 x_4 \tag{20}$$

$$\frac{dx_6}{dt} = \alpha_5 x_2 x_4 - \alpha_6 x_5 x_6 + \alpha_7 x_4 + \alpha_8 x_7 x_{12} - \alpha_9 x_6 \tag{21}$$

$$\frac{dx_8}{dt} = \alpha_{10} x_{11} x_3 - \alpha_{11} x_8 - \alpha_{12} x_8 \tag{22}$$

$$\frac{dx_9}{dt} = \alpha_{12} x_8 - \alpha_{13} x_9 \tag{23}$$

$$\frac{dx_{10}}{dt} = \alpha_{13} x_9 - \alpha_{14} x_{10} \tag{24}$$

$$\frac{dx_{11}}{dt} = -\alpha_{10} x_{11} x_3 + \alpha_{11} x_8 + \alpha_{15} - \alpha_{16} x_{11} - \alpha_{19} x_{11} x_{13} + \alpha_{20} x_{14} - \alpha_{21} x_{11} x_7 + \alpha_{22} x_{15} \tag{25}$$

$$\frac{dx_{12}}{dt} = -\alpha_8 x_7 x_{12} + \alpha_9 x_6 + \alpha_{17} - \alpha_{18} x_{12} \tag{26}$$

$$\frac{dx_{13}}{dt} = -\alpha_{19} x_{11} x_{13} + \alpha_{20} x_{14} - \alpha_{21} x_{11} x_7 + \alpha_{22} x_{15} \tag{27}$$

$$\frac{dx_{14}}{dt} = \alpha_{19} x_{11} x_{13} - \alpha_{20} x_{14} \tag{28}$$

$$\frac{dx_{15}}{dt} = \alpha_{21} x_{11} x_7 - \alpha_{22} x_{15} \tag{29}$$

Figure 2: Differential equations in the Lee *et al.* (2003) model.

4 Results

An implementation of the ABC–SMC method (Algorithm 2) is available in the Python package ABC–SysBio (Liepe *et al.*, 2010). Several kernels are already available in the ABC–SysBio. We used the component–wise Gaussian (cwG), the multivariate Gaussian (mvG) and the K–neighbors multivariate Gaussian (KmvG) for comparison. Assume a set of parameter vectors $\boldsymbol{\Theta} = (\boldsymbol{\theta^1}, \ldots, \boldsymbol{\theta^N})$ with weights $\boldsymbol{w} = (w^1, \ldots, w^N)$, where each parameter vector has d components $\boldsymbol{\theta^i} = (\theta^i_1, \ldots, \theta^i_d)$.

In the cwG kernel each component θ_j of the parameter vector is perturbed independently according to a Gaussian distribution with mean θ_j and variance:

$$\sigma_j^2 = \frac{\sum_{i=1}^N w^i (\theta_j^i - \mu_j)^2}{V - (\mathcal{V}/V)} \tag{33}$$

where $V = \sum_{i=1}^N w^i$ and $\mathcal{V} = \sum_{i=1}^N {w^i}^2$ are the sum of the weights and the sum of the squares of the weights respectively and $\mu_j = \sum_{i=1}^N w^{(i)} \theta_j^i$ is the weighted mean for the specific component.

In the mvG kernel, the correlations between the components of the parameter vector are taken into account. A particle $\boldsymbol{\theta_j}$ is perturbed using a multivariate normal distribution with mean $\boldsymbol{\theta_j}$ and covariance matrix:

$$\Sigma = \frac{V}{V^2 - \mathcal{V}} = \sum_{i=1}^N w^i (\boldsymbol{\theta^i} - \boldsymbol{\mu})^T (\boldsymbol{\theta^i} - \boldsymbol{\mu}) \tag{34}$$

where $\boldsymbol{\mu} = \sum_{i=1}^N w^i \boldsymbol{\theta^i}$ is the weighted mean.

The KmvG kernel, starts by selecting for each particle $\boldsymbol{\theta^j}$ the K–nearest neighbors. Then a multivariate normal distribution with mean $\boldsymbol{\theta_j}$ and covariance matrix computed by (34) but taking into account the K–nearest neighbors only, is used to draw a new particle. In ABC-SysBio, K is set by default to a quarter of the number of particles, and this value was used in the following studies.

Further to the above described packages, featured by the ABC-SysBio framework, we have developed a Python package for the DPGMM and the Gibbs sampler described in Section 2. Regarding the DPGMM inference several iterations may be needed for the sampler to converge. In order to increase the overall computational efficiency, a reduced number of iterations (50) was pre–selected. At each iteration, the Bayesian information criterion (BIC) was used to evaluate how well the data fit to the proposed model. The cluster assignment with the lowest BIC was used to define the transition kernels in the ABC–SMC. For the ABC–SMC, we used 1000 particles and an Euclidean distance measure to compare simulated and experimental data. While different tolerance sequences may be optimal for each kernel, we used the same tolerance sequence for all of them as we wanted to compare their efficiency under the same conditions.

4.1   Toy example

In order to highlight the advantages of the DPGMM kernel, we initially apply ABC–SMC to a toy example with known posterior. We assume that the data are produced by the likelihood function:

$$p(D|\boldsymbol{\theta}) = \mathcal{N}\left([\theta_1^2\ \theta_2^2], \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}\right) \tag{35}$$

and that $D^o = [1\ 1]$ is observed. The posterior density is then

$$p(\theta_1, \theta_2|D^s) = \phi\left([1\ 1] \middle| [\theta_1^2\ \theta_2^2], \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}\right) \tag{36}$$

where $\phi(\cdot|\boldsymbol{\mu}, \Sigma)$ is the two dimensional normal density with mean μ and covariance matrix Σ. As it is shown in Figure 3A, there are 4 modes in the posterior distribution around the points $[1\ 1]$, $[1\ -1]$, $[-1\ 1]$ and $[-1\ -1]$.

Using ABC–SMC and a uniform prior distribution on the space $[-10\ 10] \times [-10\ 10]$ we tried to estimate the posterior distribution. As it is shown in Figures 3C–F, the posteriors predicted by all 4 kernels are close to the true posterior. However, there is a significant difference in the efficiency of each kernel. When the cwG and mvG kernels are used many more particles have to be generated in order to accept 1000 particles from the next intermediate distribution in comparison to the locally concentrated KmvG and DPGMM kernels. The acceptance rates of the latter are in some cases over two times higher than the former ones. This is expected as the parameters of cwG and mvG kernels are derived from all the particles in the previous population and thus more and more of the generated particles fall into the area surrounded by the 4 modes. Moreover, the performance of the two kernels is almost identical, which is normal given that there is no correlation between the two parameters, due to the diagonal covariance matrix. From the other two kernels, DPGMM consistently provides better acceptance rates than KmvG. Especially in the last 5 populations, in which the acceptance rate is low for all kernels, $50 - 70\%$ more particles have to be generated with the KmvG than with the DPGMM kernel.

4.2   Parameter estimation for the Wnt pathway

The Lee–model (Section 3.1) consists of 22 parameters. While several of them have been measured through experiments on the *Xenopus* frog (Lee *et al.*, 2003), we expect their values to be different for measurements performed on Human cells (Tan *et al.*, 2014). Nevertheless, given the scarce experimental data available, we fixed all parameters to constant values that have been used in previous studies (MacLean *et al.*, 2016) except for three: the binding rate between β–catenin and the destruction complex (α_{11}), the β–catenin degradation rate (α_{16}), and the binding rate of β–catenin and TCF (α_{19}). Uniform priors with ranges $[0\ 100]$, $[0\ 1]$, $[0\ 1.5]$ were used for each parameter respectively. We also aimed at estimating the initial concentration of β–catenin. For the remaining species we used the initial concentrations reported in (Tan *et al.*, 2012). For β–catenin, x_{11}, we used a uniform prior with range $[0.5\ 200]$. A similar approach was used for the extended Lee model (Section 3.2). For the new parameters introduced due to the new reactions, the parameter values reported in (Mazemondet *et al.*, 2012) were used. As with the Lee model all parameters were fixed except for: the binding rate between β–catenin and the destruction complex (α_{11}), the β–catenin degradation rate (α_{16}), and the β–catenin trasportation rate α_{23}. For the first two the same uniforms as before were used, while for α_{23} we used a uniform prior with range $[0\ 10]$. Given that two pools of β–catenin are considered in the extended model, we estimated the initial concentrations of both cytosolic and nuclear β–catenin, x_{11} and x_{16}. The simulated data were compared to the experimental data from (Tan *et al.*, 2014). In Tan *et al.* (2014) the β–catenin concentration in the cytosol and the nucleus was measured at four time points (0, 60, 120, and 240 minutes) with and without Wnt. Here we consider the total β–catenin concentration after pathway activation. For the Lee model simulated data correspond to the concentration x_{11}, while for the extended model we used the total β–catenin concentration, *i.e.* $x_{11} + x_{16}$.
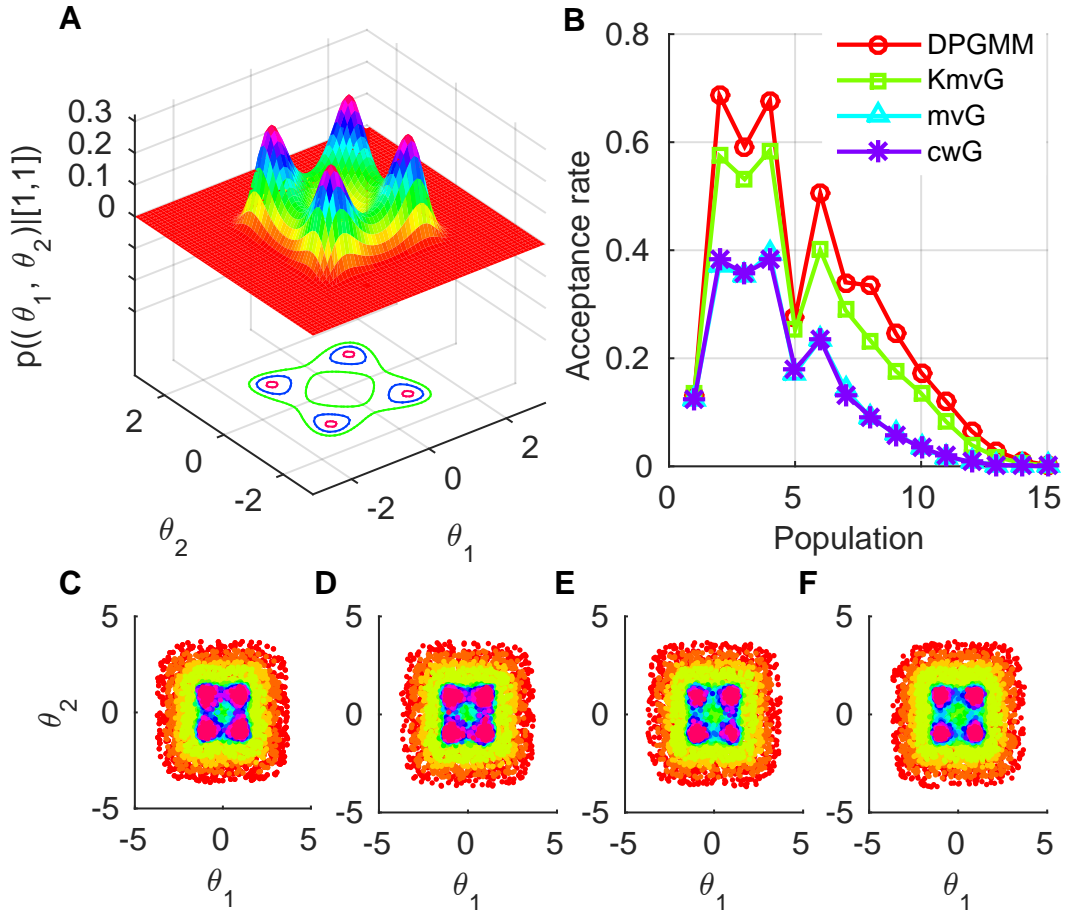
11

Figure 3: Parameter estimation results for the toy example. (**A**) True posterior distribution for the likelihood function in (35). (**B**) Acceptance rates (fraction of generated particles that are accepted) for the cwG purple), the mvG (cyan), the KmvG (green) and the DPGMM (red) kernels for each population. (**C–F**) Intermediate distributions for different tolerance values, starting with a high tolerance (red) to a low tolerance (purple) approximated using cwG (**C**), mvG (**D**), KmvG (**E**) and DPGMM (**F**) kernels.

Being mostly interested on the results than comparing the different kernels, we did not pre–define a tolerance sequence, but we used an adaptive approach to estimate the next tolerance ϵ based on the particles accepted at the current population. More precisely, after having accepted N particles for which the distance between simulated and experimental data is lower than the current tolerance ϵ_i, we select the next tolerance ϵ_{i+1} to be equal to the higher distance of the αN best particles, where $0 < \alpha < 1$. This means that the αN particles from the current population will also be in the next population. A value of α close to 1 guarantees that the intermediate target distributions do not differ significantly. We set the α equal to 0.75 and the lowest tolerance $\epsilon_T = 0.15$. Based on the results from the toy example we used the DPGMM and the K–nearest neighbors kernels to ensure that the results will be the same for the two kernels.

The posterior distributions, as approximated using the DPGMM kernel, are shown in Figures 4 and 5 for the model parameters and the concentrations respectively. Similar results were obtained using the K–nearest neighbors kernel. For the β–catenin–DC binding rate (α_{11}) the results are different for the two models. For the Lee *et al.* (2003) model it spans the prior range, while for the compartmentalized model it takes values over the higher half of its prior range. The β–catenin degradation rate (a_{16}) values are concentrated towards the lower half for both models but the range for the Lee model is much wider. Finally, the values of the third parameter, the binding rate of β–catenin and TCF (α_{19}) span the prior range, while the β–catenin trasportation rate α_{23} is concentrated around the upper limit of the prior range. It is clear from the results that some parameters affect more the results than others, for which their values do not significantly alter the results (*i.e.* the parameters whose values span the whole prior range). In summary for both models there exist sets of parameters for which model predictions match the experimental data well even if the remaining parameters have been fixed most probably to unrealistic values. This is not unreasonable if we consider the complexity of the models with respect to
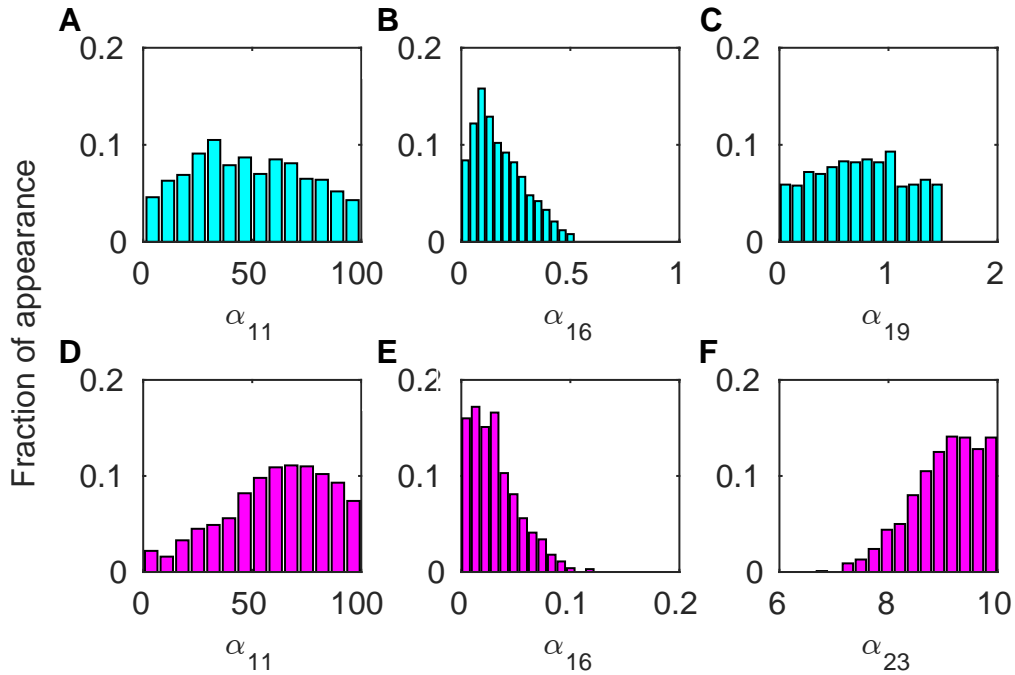
12

Figure 4: Marginal posterior distributions for the Lee *et al.* (2003) model parameters (**A–C**) and the compartmentalized model (**D–F**) approximated using ABC–SMC with DPGMM transition kernels.

the data (MacLean *et al.*, 2016). The results for the initial concentrations, on the other side, provide
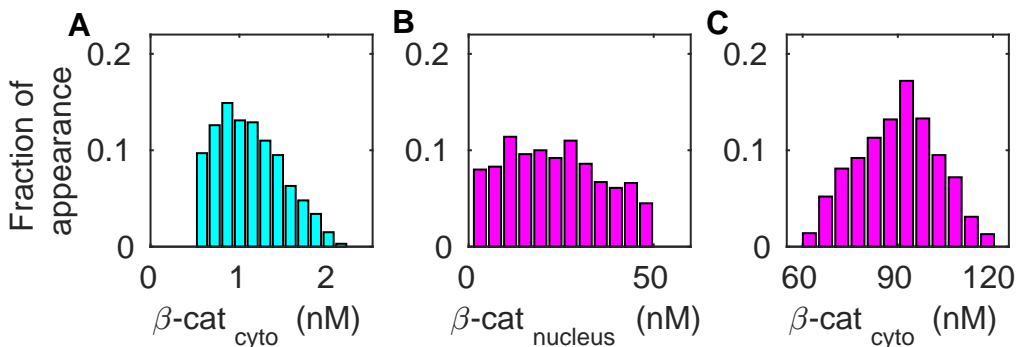

Figure 5: Marginal posterior distributions for the initial concentrations of (**A**) the cytosolic β–catenin (x_{11}) in the Lee *et al.* (2003) model, and (**B**) the cytosolic and (**C**) nuclear β–catenin in the compartmentalized model approximated using ABC–SMC with DPGMM transition kernels.

much more information than the parameter estimations. An interesting observation is that the initial β–catenin concentration for the Lee model has to be very low (the posterior is concentrated around 1, Fig. 5A) for the model predictions to be close to the experimental data. The posterior distributions for the compartmentalized model are significantly more spread. The concentration of the cytosolic β–catenin ranges from 0 to 50 nM, with a small peak around 20–25 nM, while the one for the β–catenin in the nucleus varies between 70 and 130 nM with a peak around 120 nM. These numbers are in accordance with the initial concentrations (25 nM in the cytosol and 147 nM in the nucleus) that have been exerimentally measured in HEK293T cells (Tan *et al.*, 2012).

Regarding the performance of the two kernels, the proposed DPGMM kernel provides either better or similar acceptance rates with the KmvG kernel (Fig 6A). However, the running times of the ABC–SMC algorithm (Fig. 6B) with the KmvG kernel are better than DPGMM. The reason is that the implemented Gibbs sampler is computationally expensive. However, it is important to note that the computational cost of simulating the data was extremely low for the model we consider here. One simulation of the ODE model lasted around 0.042 seconds. Stiffer or stochatic models usually require considerably more time. Projections of the running times of the ABC–SMC algorithm with acceptance rates equal to the ones observed using DPGMM and KmvG kernels and assuming simulations lasting 1 sec and 5 secs are
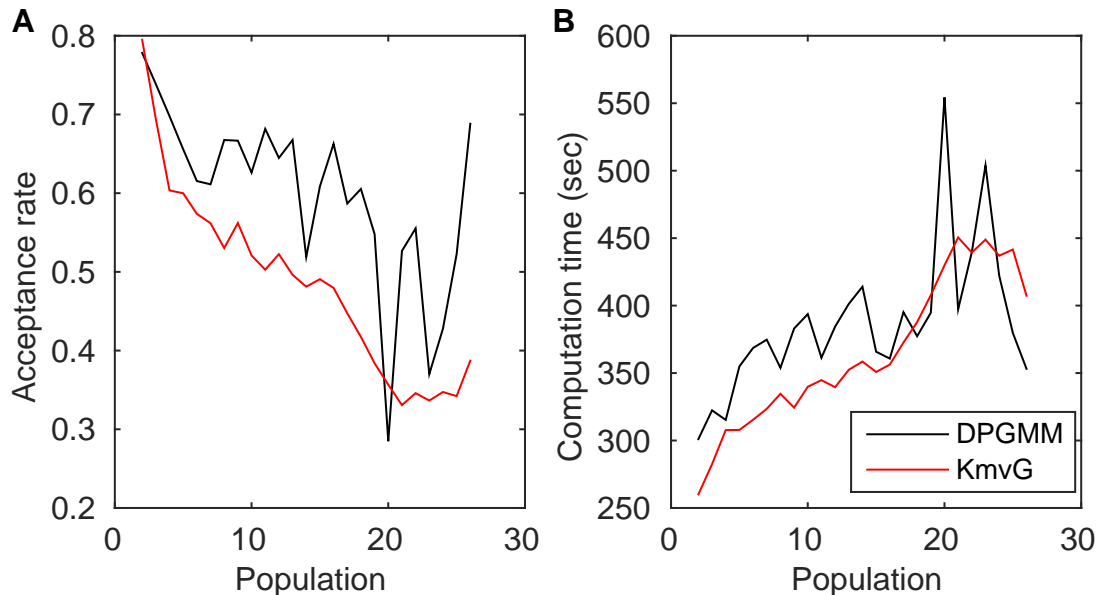
Figure 6: Performance analysis using experimental data. (**A**) Acceptance rates (fraction of generated particles that are accepted) using the KmvG (red) and the DPGMM (black) kernels (**B**) Computational times for the two kernels (solid lines). Projection of the computational times assuming a simulation time of 1 sec (dashed lines) and 5 secs (dotted lines).

shown in Figure 6B (dashed and dotted lines respectively).

5   Conclusions

We presented a likelihood-free methodology to learn the parameters of mathematical models of complex biological systems. The proposed approach is based on Dirichlet process mixtures (DPMs) that have been previously used in density estimation. We used the DPMs in a ABC–SMC framework to efficiently explore the parameter space. It was shown that the proposed approach can bring potential benefits compared to commonly used alternatives improving the efficiency of SMC mathods in targeting complex, multimodal distributions. Apart from the improved performance, another advantage of DPMs is that they are user friendly as, due to their non–parametric nature, they do not require a predefined number of Gaussian components. The advantages of the method were highlighted using both a toy example and a real biological system, the Wnt signalling pathway.

While the proposed methodology was used with ABC–SMC its applicability is not limited to the specific framework. It can be easily extended to several bayesian inference algorithms, like MCMC, to design proposal distributions. Similarly, it could be used for other problems like model selection. In the ABC–SysBio package an ABC–SMC algorithm for model selection is available. Thus, the use of DPMs for model selection is straight forward. An interesting extension of the present work is the development of an ABC–SMC algorithm to estimate model parameters using single–cell experimental data. We envisage that the increased efficiency of DPMs in estimating multimodal distributions will be beneficial in studying cell–to–cell variability and infering multidimensional parameter distributions describing a population of cells that may be grouped according to their environmental conditions.

A drawback with DPM kernels is that they may become computationally expensive. The implemented Gibbs sampler requires a significant amount of time to infer the DPGMM. However, it is important to note that the computational cost of simulating biological systems may also be high. Thus, the increase in the acceptance rate may be beneficial when the computational cost to simulate complex data is not negligible. Apart from Gibbs sampling, several alternative inference approaches for DPMs are available (Blei *et al.*, 2006; Wang and Dunson, 2011) and it would worth investigating their efficiency.

References

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The annals of statistics*, pages 1152–1174.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, **162**(4), 2025–2035.

Blei, D. M., Jordan, M. I., *et al.* (2006). Variational inference for Dirichlet process mixtures. *Bayesian analysis*, **1**(1), 121–143.

Cappé, O., Douc, R., Guillin, A., Marin, J.-M., and Robert, C. P. (2008). Adaptive importance sampling in general mixture classes. *Statistics and Computing*, **18**(4), 447–459.

Cornuet, J., Marin, J.-M., Mira, A., and Robert, C. P. (2012). Adaptive multiple importance sampling. *Scandinavian Journal of Statistics*, **39**(4), 798–812.

De, A. (2011). Wnt/ca2+ signaling pathway: a brief overview. *Acta Biochim Biophys Sin*, **43**, 745–756.

Del Moral, P., Doucet, A., and Jasra, A. (2006). Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68**(3), 411–436.

Del Moral, P., Doucet, A., and Jasra, A. (2012). An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statistics and Computing*, **22**(5), 1009–1020.

Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, **89**(425), 268–277.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, **90**(430), 577–588.

Filippi, S., Barnes, C. P., Cornebise, J., and Stumpf, M. P. (2013). On optimality of kernels for approximate bayesian computation using sequential monte carlo. *Statistical applications in genetics and molecular biology*, **12**(1), 87–107.

Fleiss, B., Van Steenwinckel, J., Schang, A., Sigaut, S., Krishnan, M., Montame, A., Degos, V., Hennebert, O., Lebon, S., Schwendimann, L., *et al.* (2015). Microglial wnt signaling inhibition promotes microglia activation and oligodendrocyte maturation blockade. In *GLIA*, volume 63. WILEY-BLACKWELL 111 RIVER ST, HOBOKEN 07030-5774, NJ USA.

Givens, G. H. and Raftery, A. E. (1996). Local adaptive importance sampling for multivariate densities with strong nonlinear relationships. *Journal of the American Statistical Association*, **91**(433), 132–141.

Görür, D. and Rasmussen, C. E. (2010). Dirichlet process gaussian mixture models: Choice of the base distribution. *Journal of Computer Science and Technology*, **25**(4), 653–664.

Jensen, P. B., Pedersen, L., Krishna, S., and Jensen, M. H. (2010). A wnt oscillator model for somitogenesis. *Biophysical Journal*, **98**, 943950.

Kühl, M. and Rao, T. P. (2010). An updated overview on wnt signaling pathways: A prelude for more. *Circulation Research*, **106**, 1798–1806.

Lee, E., Salic, A., Krüger, R., Heinrich, R., and Kirschner, M. W. (2003). The roles of apc and axin derived from experimental and theoretical analysis of the wnt pathway. *PLoS Biology*, **1**(5), 745–756.

Liepe, J., Barnes, C., Cule, E., Erguler, K., Kirk, P., Toni, T., and Stumpf, M. P. (2010). ABC–SysBio–approximate Bayesian computation in Python with GPU support. *Bioinformatics*, **26**(14), 1797–1799.

MacEachern, S. N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics-Simulation and Computation*, **23**(3), 727–741.

MacLean, A., Harrington, H., Stumpf, M., and Byrne, H. (2016). Mathematical and statistical techniques for systems medicine: The wnt signaling pathway as a case study. *Methods Mol Biol*, **1386**, 405–439.

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, **100**(26), 15324–15328.

Mazemondet, O., John, M., Leye, S., Rolfs, A., and Uhrmacher, A. (2012). Elucidating the Sources of β–Catenin Dynamics in Human Neural Progenitor Cells. *PLoS ONE*, **7**.

McLachlan, G. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

Milias-Argeitis, A., Summers, S., Stewart-Ornstein, J., Zuleta, I., Pincus, D., El-Samad, H., Khammash, M., and Lygeros, J. (2011). In silico feedback for in vivo regulation of a gene expression circuit. *Nature biotechnology*, **29**(12), 1114–1116.

Müller, P., Erkanli, A., and West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, **83**(1), 67–79.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, **9**(2), 249–265.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular biology and evolution*, **16**(12), 1791–1798.

Rasmussen, C. E. (1999). The Infinite Gaussian Mixture Model. In *NIPS*, volume 12, pages 554–560.

Schmitz, Y., Rateitschak, K., and Wolkenhauer, O. (2013). Analysing the impact of nucleo-cytoplasmic shuttling of β–catenin and its antagonists APC, Axin and GSK3 on Wnt/β-catenin signalling. *Cellular Signalling*, **25**, 22102221.

Sisson, S. A., Fan, Y., and Tanaka, M. M. (2007). Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, **104**(6), 1760–1765.

Sudderth, E. B. (2006). *Graphical models for visual object recognition and tracking*. Ph.D. thesis, Massachusetts Institute of Technology.

Tan, C., Gardiner, B., Hirokawa, Y., Smith, D., and Burgess, A. (2014). Analysis of wnt signaling β–catenin spatial dynamics in hek293t cells. *BMS Systems Biology*, **44**, 1–18.

Tan, C. W., Gardiner, B. S., Hirokawa, Y., Layton, M. J., Smith, D. W., and Burgess, A. W. (2012). Wnt signalling pathway parameters for mammalian cells. *PLoS One*, **7**(2), e31882.

Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics*, **145**(2), 505–518.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, **6**(31), 187–202.

Uhlendorf, J., Miermont, A., Delaveau, T., Charvin, G., Fages, F., Bottani, S., Batt, G., and Hersen, P. (2012). Long-term model predictive control of gene expression at the population and single-cell levels. *Proceedings of the National Academy of Sciences*, **109**(35), 14271–14276.

Wang, L. and Dunson, D. B. (2011). Fast Bayesian inference in Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, **20**(1), 196–216.

Wraith, D., Kilbinger, M., Benabed, K., Cappé, O., Cardoso, J.-F., Fort, G., Prunet, S., and Robert, C. P. (2009). Estimation of cosmological parameters using adaptive importance sampling. *Physical Review D*, **80**(2), 023507.