Eric Kernfeld

Report on Darren Wilkinson's "Parameter inference for stochastic kinetic models of bacterial gene regulation," also referred to as "the paper" or "W10." It is a chapter in the proceedings of the ninth Valencia meeting on Bayesian statistics [1].

Abstract

In this paper, Wilkinson attempts to infer reaction rates for biochemical networks in a setting with discrete observations, missing data, and measurement error. He uses vague priors and likelihood-free MCMC methods within a Bayesian model. He runs four main simulations. The first three iterate through successively more difficult and realistic measurement models, and they show the approach can accurately infer three key reaction rates with a useful precision. The fourth studies a naive model, showing it leads to overconfident, incorrect inferences. I implement the method in Julia and attempt to reproduce the experiments. I review some alternative methods in detail and discuss the relative merits of W10's approach.

Contents

1 Introduction: Parameter Inference for Biological Models

Modern biology has progressed to the point of creating *in silico* models of entire cells. The potential benefit is enormous, because unlike real cells, which must be observed via microscopy or high-throughput methods, a simulated cell can reveal its internal state. The obvious drawback is that simulations do not necessarily correspond to reality either in terms of their mechanisms or in terms of their results. One key limiting factor: even though the structure of biochemical networks is often well known, precise information about how quickly interactions play out has not kept pace. If Protein A promotes transcription of Gene B, and they mix for five minutes, we still need to know whether to expect 10, 100 or 1000 new copies of B's messenger RNA at the end. Wilkinson's paper (W10) confronts a subproblem in this domain.

Chemical reactions are often modeled using ordinary differential equations (ODE's), but Wilkinson's exact subproblem has an extra complication that rules ODE's out: natural stochasticity. W10's model organism[1] , the bacterium *Bacillus subtilis*, varies its behavior so that even if two bacteria begin in similar initial conditions, one may become mobile and the other may not. Here is one explanation, which the underlying physics support and which Wilkinson's paper subscribes to. Interactions among molecules are themselves random, driven by Brownian motion. If there are only tens or hundreds of molecules, the randomness persists in the system dynamics rather than canceling out. This phenomenon is common, and some natural systems even amplify it in order to produce usefully random behavior [2]. Wilkinson expresses this in terms of a particular stochastic model, and the main contribution of his paper is a method for inference assuming that model.

1.1 Poisson process modeling of chemical systems

For a well-stirred chemical system that maintains thermal equilibrium, Gillespie [3] derived a model from first principles. The ultimate result is a collection of competing Poisson processes, one for each chemical reaction to be modeled. Suppose there are $x_i(t)$ molecules of type i at time t, $i \in \{1...u\}$. These molecules interact via a set of reactions $\mathcal{R}_j$, $j \in \{1...v\}$, with $\mathcal{R}_j$ consuming p_{ij} particles of type i and producing q_{ij} molecules of type i. Let $R_j(t)$ denote the number of reactions of type j that happened in $[0, t]$. I'll refer to all these variables collectively using the time-varying vectors $x(t)$ and $R(t)$ along with static matrices P and Q. Defining the matrix S to be $Q - P$, the equation $x(t) - x(0) = SR(t)$ shows how the reactions affect the molecule counts. Since the particle counts are integers, the system is not constantly in flux: there will be windows, albeit short ones, where no reactions occur. The exact form of the model is this: over a reaction-free window, $R_j(t)$ is a Poisson process with intensity $\theta_j \int_0^t \prod_{i=1}^u \binom{x_i(t)}{p_{ij}}$.

The θ_i's are sometimes unknown, and they are the quantities needed for biological models.

1.2 Forward simulation and inference with complete data

Given that every reaction is observed without error, inference is still needed because the system is stochastic. Many inference methods center around some form of forward simulation, and in fact forward simulations can be done exactly and easily.

The key insight is that the intensities are constant on the interval until the next reaction. So, as with any homogeneous Poisson process, draw an exponential random variable with rate equal to the intensity of the process; that gives the correctly distributed waiting time until the next event, so long as nothing happens in between. In particular, the minimum wait time is always correctly distributed. For a collection of independent Poisson processes, use the sum of their intensities as the rate, because in distribution this equals the minimum of the individual wait times (i.e. the first reaction to occur). Choose the type of reaction from a categorical distribution with cell probabilities equal to the normalized intensities. This scheme applies to this setting because even though the process is not homogeneous, it is temporarily homogenous in between reactions. The rates can be recomputed after each new reaction. Popularized by [4], this is known as the Gillespie algorithm. Pseudocode appears in Algorithm 1.

[1] Held up against the usual meaning in statistics, this use of the word "model" is different. Similar to the way that statisticians use models to derive insight about natural phenomena, biologists use particular species to derive broader biological insights. These species are known as model organisms.

Algorithm 1: The Gillespie algorithm

```
Given:
  A simulation duration $T$
  $x(0)$, the initial particle counts
  $S$, a "stoichiometry matrix" whose $i, j$ entry says how many molecules of type $i$ appear or vanish in
  a reaction of type $j$ (net change)
  $P$, a matrix whose $i, j$ entry says how many molecules of type $i$ enter a rxn of type $j$ (not a net
  change)
  $c$, a vector of reaction rates
  -
Do this:
  Initialize $x$ to $x(0)$ and $t$ to 0.
  While true:
    Calculate $\alpha_j = \theta_j \prod_i \binom{x_i}{P_{ij}}$
    Increment $t$ by Exponential($\sum_j \alpha_j$) ($\sum_j \alpha_j$ is the rate parameter)
    If $t > T$, quit and return $x$.
    Otherwise, choose an integer $j$ with probability $\frac{\alpha_j}{\sum_j \alpha_j}$.
    Increment $x$ by adding column $j$ of $S$.
```

Returning to the likelihood, suppose ν is a vector so that the ith reaction has type ν_i and t is a vector so that the ith reaction happens at time t_i. The likelihood has a term corresponding to the exponential with the sum of the Poisson intensities. Conditioned on the wait times, the next term consists of the categorical probability given to the actual reaction type observed. That happens for every observed reaction. The normalizing constants of the factors cancel, and at the end of the day, the likelihood is

$$L(\theta|\nu, t) = \prod_{i=1}^{n} \theta_{\nu_i} \prod_{j=1}^{u} \binom{x_j(t_{i-1})}{p_{\nu_i j}} \exp\left(-\theta_{\nu_i}(t_i - t_{i-1}) \binom{x_j(t_i)}{p_{\nu_i j}}\right). \tag{1}$$

In terms of maximum likelihood inference, this function and its derivatives can be evaluated. In terms of Bayesian inference, it has a conjugate prior (independent Gamma distributions). Computational scale is not an issue: Wilkinson's examples have only 13 molecule types and 18 reactions. The problem arises because of discretely observed data.

2 Approaching inference without complete data

In Wilkinson's data, not every reaction is observed. Rather, measurements of molecule counts are taken every five minutes. Many possible sequences of reactions could have resulted in the same observations (infinitely many, counting production-decay or binding-unbinding loops). For each sequence, uncountably many variations exist: consider stretching or shrinking the wait times. Let ν range over the set of eligible reaction sequences, meaning it matches the observed data. Let Ω_ν be the set of possible times for reactions in ν. The likelihood becomes

$$L(\theta|x(0), x(t_1), ...x(T)) = \sum_{\nu} \int_{t \in \Omega_\nu} \prod_{i=1}^{n} \theta_{\nu_i} \prod_{j=1}^{u} \binom{x_j(t_{i-1})}{p_{\nu_i j}} \exp\left(-\theta_{\nu_i}(t_i - t_{i-1}) \binom{x_j(t_i)}{p_{\nu_i j}}\right). \tag{2}$$

The main obstacle here is the summation over reaction sequences. I have not made a serious attempt to evaluate it, and some papers, including W10, claim it is intractable. To further complicate the problem, most reaction sequences are not included in the sum, if we define "most" via the measure induced by simulation. In other words, rejection sampling is not a promising approach: most forward simulations, even with correct initial conditions, will not match the data.

Instead, expectation-maximization is a common recourse: alternate between calculating expected complete-data log likelihoods and updating parameters based on equation 1. Since calculating expected latent reaction times and types is not necessarily possible, as it would be in the case of a time-homogenous process, Horváth and Manini [5] instead draw samples using the Gillespie method. This produces only trajectories at odds with the observations, which an exact method with infinite computing power could afford to discard. After multiple simulations, they discard all but the trajectory closest to fitting the data, and then they improve the fit of this trajectory by altering some reactions. This makes their method inexact, though they argue the distortion of the likelihood is minimal unless molecule counts are close to zero. Daigle et al. [6] take a very similar tactic, using EM with a sampling-based E-step done by rejecting incorrect trajectories. To reduce wasted simulation time, Daigle et al. use a tool they call multi-level splitting, and it bears an eerie resemblence to the sequential sampling method that Wilkinson contributes. Another EM-based method contributed by Bayer et al. [7] confronts the same problem of how to obtain trajectories consistent with data. Their scheme is to simulate forward from one observed time-point and backwards from the next, hoping the two samples meet in the middle. A common theme in these search-based methods is that careful initialization helps reduce the fraction of useless samples. This idea, too, will be mirrored in the Bayesian literature on the topic.

In Wilkinson's work, some molecule counts are zero or close to zero, which renders the approximation in [5] dubious. Though missing data has not yet been mentioned, at time t_i, Wilkinson observes only one molecule out of a system with 13, or in other words one coordinate of the 13-dimensional vector $X(t_i)$. This disqualifies [7] by Bayer et al.'s own admission. The method by Daigle et al., [6], can accomodate partial observations, so it remains competitive, and furthermore it claims to perform better than a different paper co-authored by Wilkinson [8]. However, [6] was published two years after W10. Among these papers, W10 had no direct competitors when it was published.

Wilkinson's papers tend to use Bayesian methods for modeling and MCMC for inference. A 2003 paper by Marjoram et al. [9] opened the way for Metropolis-Hastings samplers that do not require likelihood evaluations. Wilkinson introduces one of these, claims it will not work, and makes, as his paper's main contribution, an adaptation of likelihood-free MCMC (LF-MCMC) for hidden discrete or continuous-time Markov models.

2.1 Methods similar to Wilkinson's

Wilkinson's adaptation of LF-MCMC is subtle and strange. Because it contains ingredients of several Monte Carlo schemes, the discussion leading up to Wilkinson's algorithm will briefly detour through sequential Monte Carlo and LF-MCMC. Readers familiar with these schemes may wish to skip to the pseudocode in Algorithm 2 and the attending description.

On its own, LF-MCMC is just a form of Metropolis-Hastings. To produce samples from a *target distribution* π over a hidden state x and parameters θ, recall that Metropolis-Hastings says to propose a new sample x^*, θ^* with distribution $q(x^*, \theta^*|x, \theta)$, then accept the proposal with probability $\min\{1, A\}$ if $A = \frac{q(x,\theta|x^*,\theta^*)}{q(x^*,\theta^*|x,\theta)} \times \frac{\pi(x^*,\theta^*)}{\pi(x,\theta)}$. It will help to keep in mind this "recipe" for Metropolis-Hastings, but to motivate and describe LF-MCMC, I move back to the stochastic chemical model.

Suppose we have a set of observations at times t_i, $i \in 1...I$, and for each i we label the observations $\mathcal{D}_{t_i}$. These might be real-valued scalars measuring brightness from a glowing "reporter molecule." The observations depend on the current hidden state, which is the vector of counts for each molecule type. They may also depend on θ, which aggregates parameters of the measurement error model along with any unknown reaction rates. At time t_i, Wilkinson observes only one coordinate of the 13-dimensional vector $x(t_i)$, and it is a noisy observation. On the bright side, the revealed coordinate is fixed and known, and $P(\mathcal{D}_{t_i}|x(t_i), \theta)$ is assumed to be known and tractable. W10 imposes a prior $P(\theta)$, also tractable. The target distribution for the MCMC is chosen to be the posterior $P(x, \theta|\mathcal{D})$, meaning the term $\frac{P(x^*,\theta^*|\mathcal{D})}{P(x,\theta|\mathcal{D})}$ needs to be computed at each iteration. This term can be rewritten $\frac{P(\mathcal{D}|x^*,\theta^*)P(x^*|\theta^*)P(\theta^*)}{P(\mathcal{D}|x,\theta)P(x|\theta)P(\theta)}$. The only intractable calculation is $P(x^*|\theta^*)$, which is a special case of Equation 2. To avoid it, the LF-MCMC comes in.

Marjoram et al.'s LF-MCMC dodges this type of problem not by approximation, but by making the intractable term appear in a second location and cancel the first. In Metropolis-Hastings, each iteration needs the quantity $A = \frac{q(x,\theta|x^*,\theta^*)}{q(x^*,\theta^*|x,\theta)} \times \frac{\pi(x^*,\theta^*)}{\pi(x,\theta)}$. Noticing that the asterisks appear in the numerator for q and the denominator for π, the inspiration is to build $P(x|\theta)$ into the *proposal.* By proposing only θ, then simulating x from the correctly-formed model (here, via the Gillespie algorithm), becomes $A =$

$\frac{q(\theta|\theta^*)}{q(\theta^*|\theta)} \frac{P(x|\theta)}{P(x^*|\theta^*)} \times \frac{P(D|x^*,\theta^*)P(x^*|\theta^*)P(\theta^*)}{P(D|x,\theta)P(x|\theta)P(\theta)}$. The intractable terms cancel.

Wilkinson points out that this scheme suffers from the same problem that besets the sampling-based EM: most forward simulations are incongruent with the data. Once the chain gets to taste a single well-placed sample sequence of reactions, it is unlikely to explore parameters and initial conditions other than what produced the winning simulation. The rejection rate is high, and the mixing time is bad. Wilkinson reasonably claims that this problem gets worse with less measurement error or with higher-dimensional observations, and that Marjoram et al.'s method will not work "out of the box." He offers no theory, citations, or experiments to support this claim, except for a clause containing the phrase "in practice," which implies experiments not reported in W10. Wilkinson's work centers around building a sampler with tolerable rejection rates.

The idea is similar to sequential Monte Carlo (SMC), which uses a sequence of distributions rather than moving directly from proposal to posterior. Most SMC methods resemble importance sampling: propose a collection of independent, identically distributed samples from the wrong distribution. To render the result unbiased, calculate functionals only after appropriately up- or down-weighting each of the samples.

As is done in this branch of MCMC research, I will begin to refer to each individual sample as a particle. SMC solves a problem in importance sampling where most particle weights are tiny, and just a few particles contribute most of the information. The SMC solution is to introduce several intermediate stages of resampling in which successful particles can multiply and low-weight particles can vanish. Some strategies use perturbations so that the duplicate particles disperse during resampling. If the weights at intermediate stages come from distributions that progress gradually from proposal to target, the effect is a particle cloud slowly migrating into the correct formation. Simple importance sampling would produce a particle cloud whose body all but vanishes and whose wispy edges determine the final estimate.

Because Wilkinson's scheme involves elements of both SMC and MCMC, it is useful to consider both side by side. In particular, both methods render dependent samples, but for different reasons. In SMC, samples are dependent because many of them have walked randomly away from a common ancestor. MCMC includes the other extreme. If I use an independence proposal, meaning my $q(\theta^*|\theta)$ does not depend on θ, the issue of slow random walks does not arise. Rather, dependence arises because of rejections, because each time, θ must be repeated in order to preserve the correct target density.

2.2 Inference via Wilkinson's method

Wilkinson's method indeed lies at this extreme of MCMC: his proposal distribution does not depend on the previous sample. As mentioned in Section 2.1, dependence arises because of rejects (i.e. repeated samples), and the particle cloud is coaxed into the correct shape because some proposals are rejected more often than others.

The exact method follows in Algorithm 2. As it involves nested loops, pseudocode may be an easier format to read it in, but the essence is this. Draw a large sample from the prior of choice $P(\theta, x(0))$ on the initial state and the parameters. Its empirical distribution will be called the "current distribution." Use the current distribution (or a smoothed version) as a proposal distribution. The first step will convert it to a sample from $P(\theta, x(t_1)|\mathcal{D}_1)$, which indicates two changes. First, the sample comes from a distribution conditioned on one observation, not from the prior, and second, it describes the state at the time of that observation, rather than the initial state. This pattern continues through the rest of the stages: each time, samples condition on one more datum, and states are sampled at the time when this datum was recorded. In more precise terms, the ith step is a Metropolis-Hastings procedure, and it results in samples from $P(\theta, x(t_i)|\mathcal{D}_1, ...\mathcal{D}_i)$. As far as the procedure, it proposes $(x(t_{i-1})^*, \theta^*)$ from the current distribution, then produces $x(t_i)^*$ via forward simulation using $x(t_{i-1})^*$ as the initial condition and parameters θ^*. It accepts the proposal with probability the lesser of 1 and $A = \min(1, \frac{P(\mathcal{D}_{t_i}|x(t_i)^*,\theta^*)}{P(\mathcal{D}_{t_i}|x(t_i),\theta)})$. After throwing away the first few states as a burn-in and more as desired to thin out the chain, begin to regard the results as the "current sample" and move to the next iteration.

To be clear, instead of running MCMC just once, this method runs another five million steps through the sampler for every time point in the dataset. It is not as outrageous as it sounds; many competitors scale linearly in both the forward simulation cost and the length of the Markov chain. For example, the approximate likelihood used in every MCMC step in [10] is linear in the time-series' length, and the differential equations giving the approximate likelihood inside the MCMC loop in [11] and [12] are analogous to forward simulations.

Algorithm 2: Wilkinson's sequence of MCMC Samplers

```
Given a hidden continuous-time Markov process $\{x_t\}_{t=0}^{T}$ with:
    Unknown parameters $\theta$
    Known initial state $x_0$
    Data points $\mathcal{D}_{t_i}$ at times $t_i$, $i \in \{1, ...I\}$
    A simple, tractable error model $P(\mathcal{D}_{t_i}|x(t_i), \theta)$
    A simulator for paths of $x$ given $\theta$ the process
    A big array $B_0$ (Wilkinson uses length 1,000,000) of samples from a prior on $\theta, x_0$
    Empty arrays $B_i$ of the same length

For each time point (for $i \in \{1, ...I\}$), fill $B_i$ with samples from this Metropolis-Hastings scheme:
    Initialize $(\theta, x(t_i))$
    Until $B_i$ is full:
        Draw $(\theta^*, x(t_{i-1})^*)$ from $B_{i-1}$ or a kernel density estimate (KDE) from its contents (note
        $t_0 \equiv 0$)
        Using $(\theta^*, x(t_{i-1})^*)$, simulate up to $x(t_i)^*$, the state at time $t_i$
        Set $A = \min(1, \frac{P(\mathcal{D}_{t_i}|x(t_i)^*, \theta^*)}{P(\mathcal{D}_{t_i}|x(t_i), \theta)})$
        With probability $A$, overwrite $(\theta, x(t_i))$ with $(\theta^*, x(t_i)^*)$
        If the number of times through this loop exceeds 1000 (for the burn-in) and equals one
        modulo five (for the thinning), add $(\theta, x(t_i))$ to $B_i$
```

Why is the target as claimed? To be precise, this would be done by mathematical induction, but I show only a proof that stage i is correct given stage $i-1$ is correct. From the Metropolis-Hastings recipe, the acceptance probability is either 1 or the proposal ratio times the posterior ratio. For step i, the proposal ratio × posterior ratio can be expanded as follows. The posterior ratio is written as a ratio of filtering probabilities $P(\mathcal{D}_{t_i}, x(t_{1:i}), \theta|\mathcal{D}_{t_{1:i-1}})$.

$$\frac{\overbrace{P(x(t_{1:i-1})^*, \theta^*|\mathcal{D}_{t_{1:i-1}})}^{\text{from the induction hypothesis}}}{P(x(t_{1:i-1}), \theta|\mathcal{D}_{t_{1:i-1}})} \frac{P(x(t_i)^*|x(t_{1:i-1})^*, \theta^*{\color{gray}, \mathcal{D}_{t_{1:i-1}}})}{\underbrace{P(x(t_i)|x(t_{1:i-1}), \theta{\color{gray}, \mathcal{D}_{t_{1:i-1}}})}_{\text{from the forward simulation}}} \times$$

$$\frac{\overbrace{P(x(t_{1:i-1}), \theta|\mathcal{D}_{t_{1:i-1}})}^{\text{For the filtering probability, start here...}}}{P(x(t_{1:i-1})^*, \theta^*|\mathcal{D}_{t_{1:i-1}})} \quad \frac{P(x(t_i)|x(t_{1:i-1}), \theta{\color{gray}, \mathcal{D}_{t_{1:i-1}}})}{\underbrace{P(x(t_i)^*|x(t_{1:i-1})^*, \theta^*{\color{gray}, \mathcal{D}_{t_{1:i-1}}})}_{\text{...then look here...}}} \frac{P(\mathcal{D}_{t_i}|x(t_i), \theta{\color{gray}, \mathcal{D}_{t_{1:i-1}}})}{\underbrace{P(\mathcal{D}_{t_i}|x(t_i)^*, \theta^*{\color{gray}, \mathcal{D}_{t_{1:i-1}}})}_{\text{...and finally, look here.}}}.$$

Most of this cancels, which shows how the brief expression in Algorithm 2 arises. Notice the gray terms: the hidden Markov model structure lets us include or omit them at will. This is key: otherwise the proposal, which is not conditioned on the data directly, would not cancel the intractable likelihood term, which conditions on the data. As a side note, using a KDE affects this ratio, because the densities $\frac{P(x(t_{i-1})^*, \theta^*|\mathcal{D}_{t_{i-1}})}{P(x(t_{i-1}), \theta|\mathcal{D}_{t_{i-1}})}$ in the proposal get convolved with the kernel but their counterparts in the target do not.

3 Experiments and discussion of the method

3.1 Reproducing Wilkinson's experiments

Wilkinson runs four main experiments, all of them close to identical. In each of them, he uses a 13-molecule, 18-reaction system. He creates simulated data from the Markov process model, generating 24 observations at 300-second intervals and assuming Gaussian noise with a standard deviation of 10 molecules. He treats the noise distribution as known, along with all but three reaction rates. His experiments are biologically motivated: a different molecule is observed each in each experiment, and the results show which molecules contain enough information to pin down the parameters. In the fourth experiment, the form of the model is mis-specified, and he shows that this leads to strong and incorrect posteriors.

Figure 1 is representative of the results of the experiments done with properly specified models. It shows a contour plot of the posterior density over two of three unknown log parameters. The correlation in the posterior appears because the parameters represent binding and unbinding of the same pair of molecules. Chemical systems exist in dynamic equilibrium; in other words, near-constant amounts of a substance reflect equal rates of gain and loss rather than underlying stasis. Given a dynamic equilibrium with some molecules bound together and others free, uncertainty remains about how dynamic the equilibrium is.

Because these experiments were similar to one another and biologically motivated, I planned at first to reproduce only one of them. However, all of my replication attempts failed. Below, I detail a few of my results and the settings I used.

W10 specifies most of the details about the experiments, including priors, the chemical system, the initial amount of each molecule, which molecule was observed, the burn-in time, the number of particles, and the thinning of the Markov chain. The simulated data are not given, and it is unclear whether noise was added to the already-stochastic simulated data or whether the model assumed the noise without it actually being present. W10 also does not discuss the choice of KDE in the resampling step, but via email, Wilkinson suggests a Normal kernel using the bandwidth from Silverman's rule of thumb [13] divided by ten. My experiments used a Normal kernel with standard deviations 0.05, 0.5, and Silverman over ten; analogues to W10's results (Figure 1) appear in Figure 2.

Though the 0.5-bandwidth trial in the middle pane of Figure 2 assigns substantial density near the true value, there is still cause for doubt. W10 gives higher rates than the truth for both reactions, which makes sense because lower-than-expected binding could mask lower-than-expected unbinding. This replication

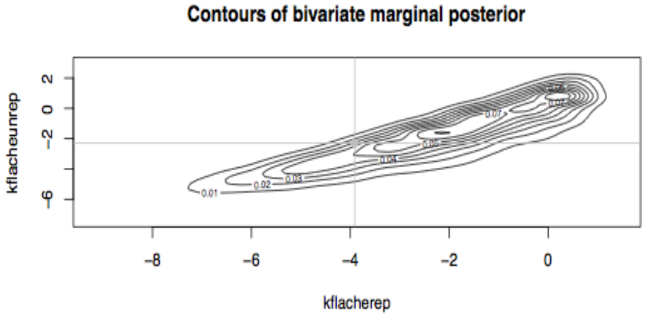

Figure 1: Bivariate marginal of the posterior from Wilkinson's figure 3. Axes show log reaction rates for binding and unbinding rates of two molecules. True rates appear as intersecting gray lines.

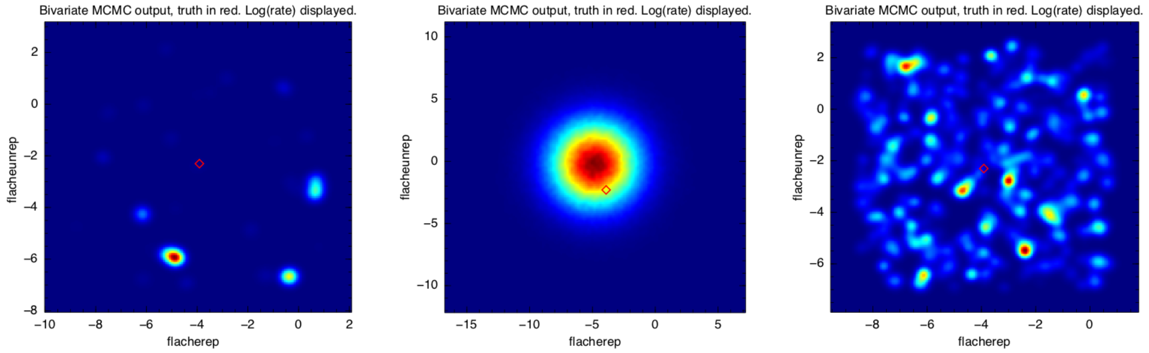

Figure 2: Bivariate marginal of the posterior from several attempts at replicating W10's figure 3. Axes show log reaction rates for binding and unbinding rates of two molecules. True rates appear as a red diamond. From left, attempts use bandwidth 0.5 and 0.05 (absolute) and then one tenth of the Silverman bandwidth.

makes mistakes in opposite directions for the two coordinates. Furthermore, W10 shows a thin, diagonal posterior: the method consistently opted for matched binding and unbinding rates. Even though W10's results are based on unpublished, simulated data and the replication is based on different data, the results should share a basic property like posterior correlation of rates for opposing processes.

3.1.1 A more basic experiment

The estimates above show a discouraging lack of sensible spatial patterns: individual modes show none of the posterior correlation of W10's results, and the modes do not congregate towards the ground truth above other regions of prior support. At this point, it made sense to ask whether the method could generate sensible posterior estimates in a simple system. I encoded a system with only one molecule, using a constant, unknown immigration rate and a constant, unknown decay rate. Figure 3 shows the prior sample, the particle cloud conditioned on one data point, and the particle cloud conditioned on two data points. The signal is weak, but the algorithm discards regions with too much production and too much decay.

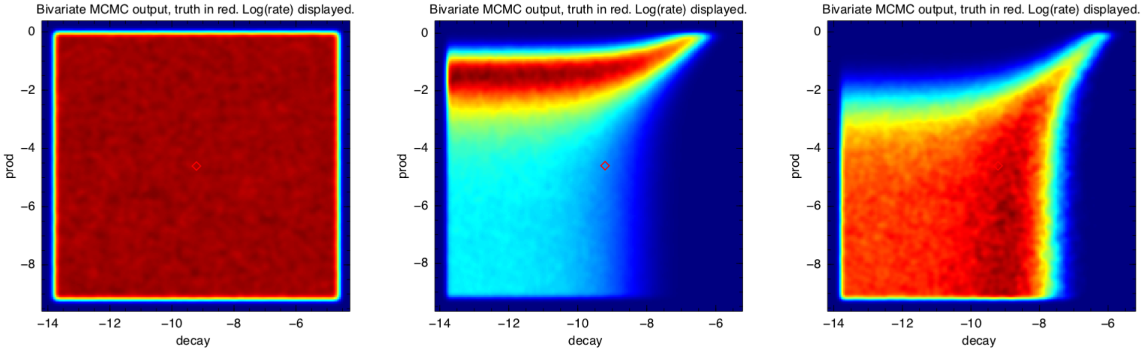

Figure 3: Bivariate marginals of the distribution of production (vertical) and decay (horizontal) log rates in a simple system. From left to right, distributions condition on zero data points, one, and two.

3.2 Software documentation and testing

I built the infrastructure for this project in three Julia modules: one for simulation of chemical systems, another for particle MCMC, and a third to manage the complexity involved in the various experiments. The project is freely available on GitHub, and the README file included with it contains further explanation of the code structure and example calls to functions that automate experiments. As with any software project, new features constantly forced me to modify the interfaces I had designed, so I apologize if any of the documentation has escaped my efforts to keep it up to date.

Testing and debugging mathematical code is difficult, and within that arena, stochastic algorithms are particularly puzzling because tests cannot expect any particular result given a particular input. Of these, MCMC is still more problematic because it is used to gain access to distributions not available by any other method. Likelihood-free methods, designed to circumvent an extra unknown distribution, are by nature the worst of the bunch. Given the failure to reproduce W10's results, it is only fair to ask whether the code is correct.

I believe that it is. Here is a summary of references on MCMC debugging, along with the evidence that convinces me my implementation is correct.

- For debugging MCMC, Gu [14] outlines a "5-possibility headache." Elsner [15] gives similar advice. Here is a list of possibilities that can account for MCMC problems, along with remarks on whether each applies.
  - Unreasonable prior: this cannot be the problem, because the ground truth is known and the prior is centered correctly.
  - Incorrect code: I address this below.
  - Incorrect math: W10 has taken care of the mathematics, and I verify them in section 2.2.
  - Bad mixing: I believe this is the problem. I discuss it in section 3.3.
  - Bad model: this Markov jump process model has strong grounding in physics. Besides, the data here are synthetic.
- As far as the code, Geyer [16] suggests printing all of the internals, including random numbers generated, desired acceptance probabilities, acceptance or rejection events, the chain state before the event, and the state after. Given a random draw, the algorithm's response is deterministic. As long as it is safe to assume elementary random numbers are being drawn correctly, this removes the stochastic testing problem. I have implemented this scheme as a "verbose" option for both the MCMC routine and the Gillespie routine. Scrutinizing the results, I found no unexpected behavior.
- The Geweke test [17] alternates between sampling emissions given parameters/latent variables and vice versa. Both operations preserve the joint distribution over $\theta, X, \mathcal{D}$, so the results can be compared

against a direct sample from the generative model being used. Furthermore, this alternation tends to amplify errors in the code: incorrect unobservables produce incorrect data, which leads to further divergence in the unobservables, and so on. I implemented a simple Geweke test where the forward model for $X(1)$ consists of adding θ to $X(0)$, then rounding to the nearest integer. I used a unif(0, 10) prior for theta, an initial state of 1, and a standard normal noise model. When re-sampling $X(1)$ and θ, I used 1,000 particles and returned the 10th sample. For each method (iid generative vs. MCMC), I generated 10,000 samples. Figure 4 shows my results in the form of Q-Q plots. The two distributions match.

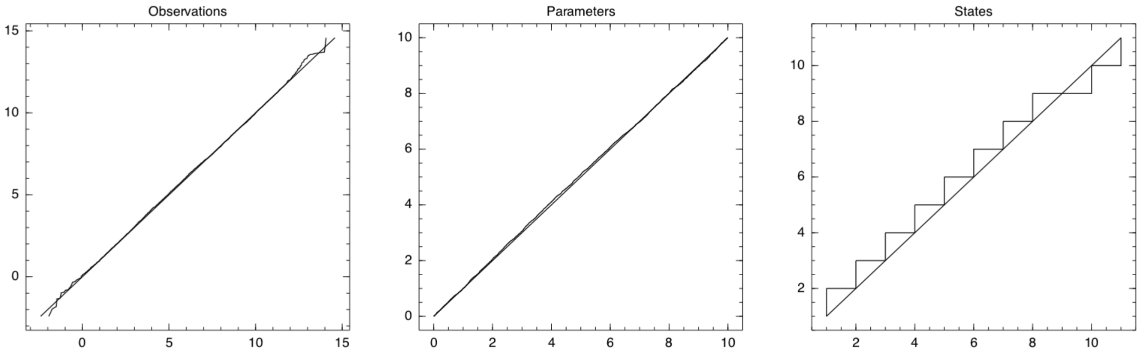

Figure 4: Results from a Geweke test of the pMCMC_julia.jl module. Each pane gives a Q-Q plot comparing observations (left), parameters (center), or hidden states at time 1 (right) from (iid generative vs. MCMC). Each plot also shows $y = x$ for reference.

- Unlike MCMC, any implementation of the Gillespie algorithm can be compared against analytical results for simple systems. For example, a single-model, single-reaction system with n_0 molecules decaying at rate k can be repeated many times, and after t seconds, the average number of molecules left will be $n_0 \exp(-kt)$. Figure illustrates such a comparison (left pane), along with a typical simulation from the *B. subtilis* motility regulation model from my experiments (center) and W10's (right).

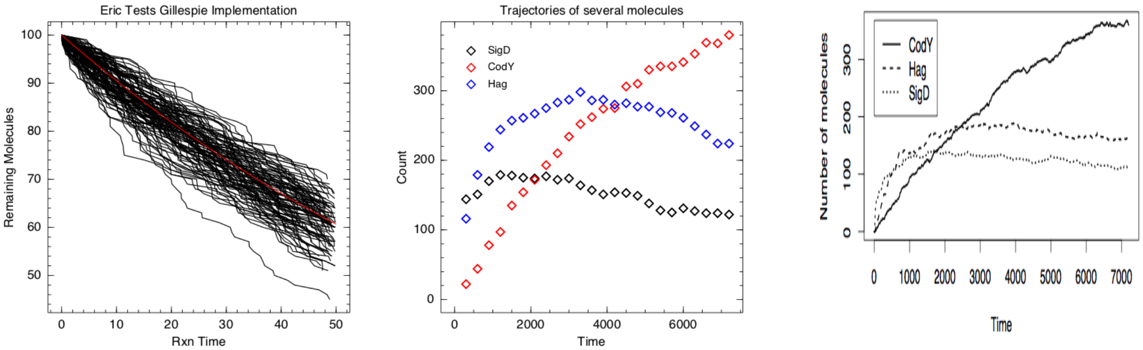

Figure 5: Left: comparison of the stochastic mean $n_0 \exp(-kt)$ (red) with multiple simulations of the corresponding Markov process. Remainder of figure: representative simulations from my work (center) and W10 (right).

3.3 Sample impoverishment

Sequential Monte Carlo suffers from a problem termed sample impoverishment, which means that eventually most samples are duplicates of a common ancestor and few unique values remain. Plots of results even with 1,000,000 samples seem to display this issue. To give an example, Figure 2 shows the posterior distribution over two unknown rates from a replicate of one of W10's experiments. The acceptance rate in the first round was 2 percent, [2] but it ranged between 25 and 95 percent thereafter. Though the samples are scattered around the true value, the plot indicates very few unique particles persisting to the final round.

This warrants a brief analysis. Wilkinson records the first of every five samples for use in the next stage's proposal. Ten rejections in a row thus guarantees a repeat. Recording the empirical rejection rates from the trial in the figure and taking each to the tenth power gives convervative estimates of the rate of 10-rejections-in-a-row, as if the trials were independent. This seems charitable: rejections could be correlated, since any small string of rejections signals a desirable starting point and thus more rejections to come. Converting these to rates of at-least-one-acceptance-after-10-rounds, we get proportions p_i such that N unique particles before stage i will result in about Np_i unique particles after stage i. Multiplying the stages together, the result is about 0.0003. So, being charitable, I would expect the final sample to have about 300 unique values for every 1,000,000 in the prior sample.

In fact, this treats the method too kindly for another reason: I would expect some more repeats to arise because the proposal distribution will begin to contain repeats. When proposed, these might be more likely to be accepted, as they were already selected to be probable based on earlier data.

Wilkinson, with lead author Andrew Golightly, criticizes W10's sample impoverishment problem in a later paper [10].

3.4 Comments on the method

3.4.1 Computing time

Wilkinson's example lies at a scale tailored to custom experiments with far fewer molecules and reactions than the whole-cell efforts cited earlier. The entire system describes only 13 molecules and 18 reactions. The experiments use a synthetic time-series with only 24 measurements. Attempting to replicate these results using Wilkinson's priors, my Julia-language implementation of the method takes about 38 hours to run. As opposed to Matlab or R, Julia offers fast performance, including loops, to mimic that of a statically typed language, and my implementation has been optimized using the Julia `@profile` macro. It takes advantage of sparsity in the stoichiometry matrix, and almost all of the time is spent on random number generation and floating-point operations. Wilkinson also notes computation time as an area for improvement, particularly for future work when it will be necessary to process data from entire batches of cells undergoing the same experiment.

A notable weakness of this algorithm is that it cannot be parallelized at any stage. Looking at the outside loop, each time a data point is added, the whole chain must run before adding another data point, because the proposals need results from the previous stage. Within a single chain, each calculation depends on the state of the chain, which depends on the previous calculation. Within each iteration, the forward simulation consumes the bulk of the power, but this too is a Markov process requiring serial computation. One sensible modification would be to run many chains at each stage, burning each of them in separately. Wilkinson does not report running multiple chains.

3.4.2 Modeling error

All of Wilkinson's experiments use synthetic data, generated from the correct model. Having neglected all but the most important chemicals, Wilkinson bears the burden of showing that modeling error in this reduced model is not too great. The method can perform no better than the model it is given, and this is doubly true because even if better models exist, the method is not computationally viable for bigger systems. Wilkinson stresses that much more information about his example network is available than is used, but

[2] Interestingly, the highest rejection rates are concentrated in the first couple of rounds, when moving between $P(\theta)$, $P(\theta|\mathcal{D}_{t_1})$ and $P(\theta|\mathcal{D}_{t_1}, \mathcal{D}_{t_2})$. This suggests a simple adaptation: spend more effort in the first round, perhaps by thinning more aggressively.

offers no citations to show that his reduction works well or that, in general, models can isolate a dozen molecules out of a complicated biochemical soup and still perform well.

If model error is possible, that raises another question: how much extra effort towards exact inference is warranted? W10 is meant to study systems for which a diffusion approximation does not serve, but would some faster approximation be worthwhile? For example, a technique known as tau-leaping uses a piecewise-constant-intensity approximation, and it preserves both stochasticity and discreteness; a 2009 work allows rates to change quadratically over leaps [18].

3.4.3 Look-ahead steps and preliminary smoothing

In many experiments, the particle cloud survived several stages, then underwent a sudden collapse down to a single mode. This seemed to match up with spots in the simulated data where consecutive error terms went in opposite directions. This suggests simple ways to improve mixing: precede the pMCMC inference with some Nadaraya-Watson smoothing, or artificially inflate the assumed error variance in the style of Approximate Bayesian Computation [19], or implement some type of look-ahead step.

3.4.4 Mixing time

Wilkinson discards 1000 samples as a burn-in at every stage of the sampler. Is this sufficient? He offers experiments showing scenarios where the method does or doesn't give good results, but he never addresses this concern about whether it has converged, not even via citation or theoretical mixing time analysis. Meanwhile, other authors using MCMC describe discarding far more of their samples as a burn-in, between 10 and 50 percent [20, 21].

3.4.5 Prior Selection (cons)

In experiments, Wilkinson's priors cover four orders of magnitude, but they are centered (on a log scale) exactly over the true parameters. Are the results robust to choices of prior that merely cover the true values?

3.4.6 Prior Selection (pros)

Unlike some competitors, W10's priors need not be dictated by convenience.

3.4.7 Flexibility

Wilkinson's approach is wonderfully adaptable. On the surface, Marjoram et al.'s LF-MCMC makes no HMM assumption and so is more flexible, but Wilkinson claims convincingly that LF-MCMC does not mix properly in this setting. The technique of constructing intermediate distributions by adding one datum at a time would generalize to any hidden Markov model, and coupled with the likelihood-free approach, this allows tremendous flexibility for indirect observation of stochastic and/or nonlinear systems. Adaptation for different types of observation (for example, a fluorescent protein instead of the protein of interest) is simple. Wilkinson points out that posteriors could be fed straight back into the algorithm as priors to analyze a second dataset, and this approach allows for analysis of different models that share parameters. The main aspect limiting flexibility was already mentioned: computing cost.

4 Discussion and alternative methods

This area seems to have progressed rapidly starting around 2006: Reinker et al write in [2] that they are not aware of methods that can cope with measurement error in systems with few molecules. In 2008, Boys *et al.* write "...we believe this is the first systematic attempt to conduct rigorous 'exact' inference for partially and discretely observed stochastic kinetic models" [8]. Now, in mid-2015, many strategies exist. In fact, there is too much to describe it all unless I am to turn this project into a review paper. I attempt to digest and organize a sort of a snowball sample of this literature, centering around the citations within W10. Some papers, based on EM, are described in Section 2. Others include methods of moments, variational inference, and myriad adaptations of MCMC. My main concerns in assessing these models are tolerance for missing data and scaling.

4.1 Moment closure

Some schemes, such as [11], [22], and [12], match moments to choose parameters. The core of these methods is an analytically-derived differential equation system that changes with the parameters. The system is typically infinite in size: only an infinite number of moments can completely encode the distributions that arise from the model. To facilitate a solution, higher moments are set to zero; this gives rise to the term "moment closure" that describes the methods. For each set of candidate parameters, the system is solved numerically. In [11], the end result is a setup where quickly-solvable ODE's for mean and covariance give a Gaussian to approximate the true density, and [11] embeds this inside of a Bayesian random-walk MCMC scheme. To distinguish this MCMC from LF-MCMC, it does evaluate an approximation to the likelihood: that is exactly what the moments are used for. MCMC also appears in [12]. There, MCMC functions as a search algorithm to maximize an approximate posterior, which is evaluated as the prior times a moment-based normal approximation to the likelihood. Thus, these schemes require solving multidimensional ODE systems inside of MCMC samplers. The project [22] uses a penalty function involving higher moments, rather than a moment-matched approximate density. Despite being less interpretable in terms of probability theory, the results in a fast procedure that allows for partial observations, as well as some assessment of uncertainty through repeated runs. For more information on different uses of moment closure, [23] cites many applied projects using it. W10 precedes all of this work.

4.2 Variational inference

Mean-field variational inference approximates a posterior distribution by finding a nearby joint distribution whose coordinates are independent. For Markov jump processes, one effort at mean field variational inference [24] scales well for large systems. This work precedes W10 by a year and has some intriguing properties. It is one of few inference methods in this subfield that requires no sampling or stochastic search, and it allows for fairly flexible priors on hidden states. W10's hidden state priors arise implicitly from his parameter priors and initial state priors, meaning that in the hidden state, they model only uncertainty due to intrinsic randomness (aleatory uncertainty), but in the parameters they model uncertainty due to incomplete knowledge (epistemic uncertainty). Variational inference in [24] offers a probabilistic way to model epistemic uncertainty in the hidden states. This could be useful, for example in inferring initial conditions: W10's experiments assume the initial state is known, though Wilkinson acknowledges that the assumption is unrealistic and done for the sake of simplicity.

To estimate parameters, [24] would need to be extended, perhaps as part of a variational EM algorithm. As written, the variational approximation is for the latent states, not parameters. Even variational EM does not give estimates of parameter uncertainty, so this work is not a viable alternative to W10 for scientific parameter estimation. For even more evidence that [24] cannot handle parameter estimation properly, the method gives intervals with close-to-nominal coverage except when attempting projection, and authors attribute this failure to untreated parameter uncertainty.

4.3 Maximum *a posteriori* inference

James Rawlings' group has developed sampling algorithms that produce a semianalytical posterior distribution or likelihood over θ, which they then optimize using derivative-free methods similar to simulated annealing [20, 25]. This work came after W10 by five years. They use an ingenious importance sampling scheme: if q is the importance distribution, the idea starts as

$$\begin{aligned} P(\theta|D) &= \int_x P(\theta|x)P(x|D)dx \\ &\approx \sum_x P(\theta|x)\frac{P(x|D)}{q(x)}dx \\ &\approx \frac{1}{P(D)}\sum_x P(\theta|x)\frac{P(D|x)P(x)}{q(x)}dx. \end{aligned}$$

$P(D|x)$ is tractable. So is $P(\theta|x)$ if a conjugate gamma prior is used. They deal with $\frac{1}{P(D)}$ by normalizing the importance weights, and $P(x)$ they evaluate in closed form as $\frac{P(x|\theta)P(\theta)}{P(\theta|x)}$, which is a ratio of gamma distribution normalizing constants.

This requires use of gamma priors, which is a restriction: the posterior is a mixture of many gammas, so it cannot be recycled as the prior when processing a new dataset. Also, domain experts often use log-normal or log-uniform priors, as does W10, and both sources regard prior choice as important. However, the scheme is exact, and well-designed q can make it more efficient than MCMC-based schemes. If the prior choice issue could be resolved, this alternative might be well suited to Wilkinson's problem.

4.4 Bayesian inference via MCMC

As noted, there is too much to literature on parameter estimation to cover all of it in this project. Paradoxically, focusing on "only" the ones using MCMC seems to make the problem worse. Bearing in mind that this review is incomplete, a good jumping-off point might be [10], because (like W10) it uses multiple common techniques. One numerical trick in particular bears mention: sometimes, likelihood terms appearing in proposal ratios can be replaced with noisy but unbiased estimates, [3] and the target distribution will still be correct (details appear in [26]). This lets [10] embed an SMC estimate of the intractable discrete-data likelihood inside of a random-walk MCMC scheme, retaining exactness. The price: noisy likelihood approximations at each step make for a slower mixing chain overall; the noisier, the slower. Furthermore, the SMC inside the main loop requires forward simulations from the Gillespie algorithm. Though SMC can be run in parallel, communication between processors at every new proposal makes this impractical. Displeased with the computational cost of SMC incurred at each iteration, [10] employs another trick: they approximate the process using a continuous-state diffusion process satisfying a certain SDE. The SDE itself is intractable, so they use an Euler approximation to draw samples.

The paper [27] is similar to [10], with both using the acceptance-ratio approximation from [26]. In [27], though, multiple chains are run in parallel. To reduce burn-in costs, which scale with the number of processors used, [27] initializes each chain from an inexact importance-sampling-like procedure known as Approximate Bayesian Computation (ABC) that can be run in parallel. Other permutations and combinations of MCMC, SMC, SDE/ODE approximation, and ABC appear in the literature. For more on ABC, the paper [28] compares it against exact methods. Other papers using SDE approximations or related techniques include [29], [8] and [30].

For most methods, endpoint-conditioned simulation is the primary issue in this problem; it affects LF-MCMC, sampling-based EM schemes, and (indirectly, via the SMC estimate variance) the schemes in [10, 27]. The paper [31] would have been another good jumping-off point, as it explores several proposal-generating options and offers a rich set of references. The paper [32] is also noteworthy for detailing a tidy, though approximate, solution. The set of eligible reaction totals given the change in chemical counts is phrased as a lattice, a vector-space-like set containing only integers. A matrix spanning it is obtained, and it suffices to multiply a random vector of integers into that matrix. This work in [32] is sophisticated enough to deal with partial measurements at a given time point, and it anticipates and deals with the potential criticism (e.g. [10]) that separate block updates to parameters and latent state can lead to low acceptance rates.

The paper [21] is similar to W10, using MCMC updates inside a sequential Monte Carlo scheme. This work is a step up from the rest in terms of model complexity, treating batches of cells with some shared

[3] Surprisingly, this works even with biased estimates, as long as the bias does not depend on the parameters being sampled.

parameters and some parameters that vary by cell. The mathematical machinery involves extra tools to marginalize over between-cell variability and block updates in places where Wilkinson doesn't need them.

4.5 Conclusions

For a statistician working in 2009, W10 was a big step forward. Statistically, it is a feat: exact inference using discrete data, measured with error, to infer parameters of an arbitrary stochastic nonlinear system, with reasonable quantification of uncertainty, and all without a single likelihood evaluation. Unfortunately, the restrictions imposed to reach this combination of properties leave little room to adapt the method for parallel processing or in ways that would improve mixing. For a mathematical biologist shopping for parameter inference algorithms, I recommend newer algorithms like [27]or [21]. These scale better than W10, taking advantage of the flexibility offered by the method of moments or the particle marginal Metropolis-Hastings algorithm from [26].

References

[1] Bernardo, J.M., Bayarri, M.J., Berger, J.O., Dawid, A.P., Heckerman, D., Smith, A.F.M., West, M.: Ninth Valencia international meeting on Bayesian statistics, Benidorm, Spain, 03-08.06.2010. Oxford U.P., Oxford (2012)

[2] Reinker, S., Altman, R., Timmer, J.: Parameter estimation in stochastic biochemical reactions. IEE Proceedings-Systems Biology **153**(4) (2006) 168–178

[3] Gillespie, D.T.: A rigorous derivation of the chemical master equation. Physica A: Statistical Mechanics and its Applications **188**(1) (1992) 404–425

[4] Gillespie, D.T.: Exact stochastic simulation of coupled chemical reactions. The Journal of Physical Chemistry **81**(25) (1977) 2340–2361

[5] Horváth, A., Manini, D.: Parameter estimation of kinetic rates in stochastic reaction networks by the em method. In: BioMedical Engineering and Informatics, 2008. BMEI 2008. International Conference on. Volume 1., IEEE (2008) 713–717

[6] Daigle, B.J., Roh, M.K., Petzold, L.R., Niemi, J.: Accelerated maximum likelihood parameter estimation for stochastic biochemical systems. BMC bioinformatics **13**(1) (2012) 68

[7] Bayer, C., Moraes, A., Tempone, R., Vilanova, P.: An efficient forward-reverse expectation-maximization algorithm for statistical inference in stochastic reaction networks. arXiv preprint arXiv:1504.04155 (2015)

[8] Boys, R.J., Wilkinson, D.J., Kirkwood, T.B.: Bayesian inference for a discretely observed stochastic kinetic model. Statistics and Computing **18**(2) (June 2008) 125–135

[9] Marjoram, P., Molitor, J., Plagnol, V., Tavaré, S.: Markov chain monte carlo without likelihoods. Proceedings of the National Academy of Sciences **100**(26) (2003) 15324–15328

[10] Golightly, A., Wilkinson, D.J.: Bayesian parameter inference for stochastic biochemical network models using particle markov chain monte carlo. Interface focus (2011) rsfs20110047

[11] Milner, P., Gillespie, C.S., Wilkinson, D.J.: Moment closure based parameter inference of stochastic kinetic models. Statistics and Computing **23**(2) (2013) 287–295

[12] Zechner, C., Ruess, J., Krenn, P., Pelet, S., Peter, M., Lygeros, J., Koeppl, H.: Moment-based inference predicts bimodality in transient gene expression. Proceedings of the National Academy of Sciences **109**(21) (2012) 8340–8345

[13] Silverman, B.W.: Density estimation for statistics and data analysis. Volume 26. CRC press (1986)

[14] Gu, K.: Debugging mcmc

[15] Elsner, M.: Debugging samplers: Making mcmc work in practice

[16] Geyer, C.: Mcmc: Does it work? how can we tell?

[17] Geweke, J.: Getting it right: Joint distribution tests of posterior simulators. Journal of the American Statistical Association **99**(467) (2004) 799–804

[18] Sehl, M., Alekseyenko, A.V., Lange, K.L.: Accurate stochastic simulation via the step anticipation τ-leaping (sal) algorithm. Journal of Computational Biology **16**(9) (2009) 1195–1208

[19] Beaumont, M.A., Zhang, W., Balding, D.J.: Approximate bayesian computation in population genetics. Genetics **162**(4) (2002) 2025–2035

[20] Gupta, A., Rawlings, J.B.: Comparison of parameter estimation methods in stochastic chemical kinetic models: Examples in systems biology. AIChE Journal **60**(4) (2014) 1253–1268

[21] Zechner, C., Unger, M., Pelet, S., Peter, M., Koeppl, H.: Scalable inference of heterogeneous reaction kinetics from pooled single-cell recordings. Nature methods **11**(2) (2014) 197–202

[22] Kügler, P.: Moment fitting for parameter inference in repeatedly and partially observed stochastic biological models. PloS one **7**(8) (2012) e43001

[23] Milner, P., Gillespie, C.S., Wilkinson, D.J.: Moment closure approximations for stochastic kinetic models with rational rate laws. Mathematical Biosciences **231**(2) (2011) 99 – 104

[24] Opper, M., Sanguinetti, G.: Variational inference for markov jump processes. In: Advances in Neural Information Processing Systems. (2008) 1105–1112

[25] Srivastava, R., Rawlings, J.B.: Parameter estimation in stochastic chemical kinetic models using derivative free optimization and bootstrapping. Computers & chemical engineering **63** (2014) 152–158

[26] Andrieu, C., Doucet, A., Holenstein, R.: Particle markov chain monte carlo methods. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **72**(3) (2010) 269–342

[27] Owen, J., Wilkinson, D.J., Gillespie, C.S.: Scalable inference for markov processes with intractable likelihoods. Statistics and Computing (2014) 1–12

[28] Owen, J., Wilkinson, D.J., Gillespie, C.S.: Likelihood free inference for markov processes: a comparison. arXiv preprint arXiv:1410.0524 (2014)

[29] Golightly, A., Wilkinson, D.J.: Bayesian inference for stochastic kinetic models using a diffusion approximation. Biometrics **61**(3) (2005) 781–788

[30] Fearnhead, P., Giagos, V., Sherlock, C.: Inference for reaction networks using the linear noise approximation. Biometrics **70**(2) (2014) 457–466

[31] Golightly, A., Wilkinson, D.J.: Bayesian inference for markov jump processes with informative observations. Statistical Applications in Genetics and Molecular Biology (2014)

[32] Amrein, M., Künsch, H.R.: Rate estimation in partially observed markov jump processes with measurement errors. Statistics and Computing **22**(2) (2012) 513–526