Eric Kernfeld

Report on Darren Wilkinson's "Parameter inference for stochastic kinetic models of bacterial gene regulation," also referred to as "the paper" or "W09." 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 review some alternative methods in detail and discuss the relative merits of W09's approach. I implement the method in Julia and attempt to reproduce the experiments EMK: more on this later!. I assess the method using additional experiments designed to study mixing time, scaling with system size, and prior sensitivity EMK: more on this later!.

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 regularly dump its entire internal state to a file. 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 is this: 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 we let them 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 (W09) 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. W09's model organism, 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. 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 [?] 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} particles 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, perhaps 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 $c_j \int_0^t \prod_{i=1}^u \binom{X_i(t)}{p_{ij}}$. The c_i's are sometimes unknown, and they are the quantities needed for biological models.

1.2 Inference with complete data and simulation

Given that every reaction is observed without error, inference is still needed because the system is stochastic. To rapidly understand the likelihood, think of how to simulate the system.

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. For a collection of independent Poisson processes, use the sum of their intensities as the rate. 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 [?], this is known as the Gillespie algorithm. Pseudocode appears in Algorithm 1.

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 the final formula is

$$L(c|\nu, t) = \prod_{i=1}^{n} c_{\nu_i} \prod_{j=1}^{u} \binom{X_j(t_{i-1})}{p_{\nu_i j}} \exp\left(-c_{\nu_i} \sum_{i=1}^{n} (t_i - t_{i-1}) \binom{X_j(t_i)}{p_{\nu_i j}} \right).$$

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 $\alpha_j = c_j \prod_i \binom{X_i}{P_{ij}}$
    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 $S$.
```

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

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

The domain of integration Ω_ν has dimension n, and n changes depending on ν. Most reaction sequences are ineligible, if we define "most" via a the measure induced by simulation. In other words, most forward simulations, even with correct initial conditions, will not match the data.

To my knowledge, this likelihood has not been evaluated. Instead, expectation-maximization is a common recourse: alternate between updating parameters and calculating expected latent reaction times/types. 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 [3] instead draw samples using the Gillespie method. This produces mostly trajectories at odds with the observations, so after a few simulations, they discard all but the trajectory closest to fitting the data, and then they tweak 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. [4] take a very similar tactic, also 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 MCMC method that Wilkinson contributes. Another EM-based method contributed by Bayer et al. [5] 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 this domain 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 [3] 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 [5] by Bayer et al.'s own admission. The method by Daigle et al., [4], can accomodate partial observations, so it remains competitive, and furthermore it claims to beat a different paper co-authored by Wilkinson [6]. However, [4] was not published when Wilkinson came out with the method reproduced here.

Wilkinson's papers tend to use Bayesian methods, and they tend to use MCMC for inference. A 2003 paper by Marjoram et al. [7] 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, we 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 D_{t_i}. These might be real-valued scalars. 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 $P(D_{t_i}|x_{t_i}, \theta)$ is assumed to be tractable. So is the prior $P(\theta)$. Since the target distribution for the MCMC is the posterior $P(x, \theta|D)$, the term $\frac{P(x^*,\theta^*|D)}{P(x,\theta|D)}$ can be rewritten $\frac{P(D|x^*,\theta^*)P(x^*|\theta^*)P(\theta^*)}{P(D|x,\theta)P(x|\theta)P(\theta)}$. The only intractable calculation is $P(x^*|\theta^*)$, a special case of Equation 1. To avoid it, the LF-MCMC comes in.

Marjoram et al. dodge this type of problem not by approximation, but by making the intractable term appear in a second location. 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), the entire thing 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 this paper. 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 where most particle weights become small, 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, where 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 share a common ancestor. In MCMC, one case lies at the other extreme: using an independence proposal, meaning $q(\theta^*|\theta)$ does not depend on θ, the probability of proposing two samples with a common ancestor is zero unless q is discrete. Rather, dependence arises because of rejections, in which case 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 cloud is coaxed into the correct shape because some proposals are rejected more often than others.

The exact method follows in Algorithm 2. 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 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 $P(\theta, x(t_1)|D_1)$, which indicates two changes: first, that the sample comes from a distribution conditioned on one observation, and second, that it describes the state at the time of that observation, rather than the initial state. The ith step is Metropolis-Hastings, and it results in samples from $P(\theta, x_{t_i}|D_1, ...D_i)$. 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 initial the 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. Repeat that until no data remains.

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 isnt as bad as it sounds: for a sampler that runs only once, the chain will not mix without a sophisticated proposal, and constructing a usable proposal is often linear in the size of the data anyway.

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? 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:

$$\frac{P(x^*_{t_i}|x^*_{t_{i-1}}, \theta^*{\color{gray}, \mathcal{D}_{t_{i-1}}})}{P(x_{t_i}|x_{t_{i-1}}, \theta{\color{gray}, \mathcal{D}_{t_{i-1}}})} \frac{P(x^*_{t_{i-1}}, \theta^*|\mathcal{D}_{t_{i-1}})}{P(x_{t_{i-1}}, \theta|\mathcal{D}_{t_{i-1}})} \times \frac{P(x_{t_i}|x_{t_{i-1}}, \theta{\color{gray}, \mathcal{D}_{t_{i-1}}})}{P(x^*_{t_i}|x^*_{t_{i-1}}, \theta^*{\color{gray}, \mathcal{D}_{t_{i-1}}})} \frac{P(x_{t_{i-1}}, \theta|\mathcal{D}_{t_{i-1}})}{P(x^*_{t_{i-1}}, \theta^*|\mathcal{D}_{t_{i-1}})} \frac{P(\mathcal{D}_{t_i}|x_{t_i}, \theta)}{P(\mathcal{D}_{t_i}|x^*_{t_i}, \theta^*)}.$$

This shows how the compact form in Algorithm 2 arises. Notice the gray terms: the hidden Markov model would allow us to include or omit them at will. This is key: otherwise the proposal, which is not conditioned on the data due to the difficulty of endpoint-conditioned simulation, would not cancel the intractable likelihood term, which conditions on the data.

3 Alternatives, Potential Flaws, and Experiments

3.1 Replication and problems with sample impoverishment

3.1.1 Sample impoverishment

Sequential MCMC 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 EMK: Add a figure here?. While writing about other work, I found that Wilkinson himself, with lead author Andrew Golightly, makes this criticism when describing W09 in a later paper [**?**].

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. Taking the empirical rejection rates from a trial run with 100,000 particles 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: they could be correlated, since any small string of rejections signals a desirable starting point and perhaps 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 3 unique values for every 10,000 in the prior sample. EMK: Re-do the calculations to describe the same million-sample test for which you have a figure. How can this be made transparent/reproducible? Providing code?

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 plausible values.

Solutions to sample impoverishment Interestingly, the highest rejection rates are concentrated in the first couple of rounds, when moving between $P(\theta)$ and $P(\theta|d_{t_1})$ and $P(\theta|d_{t_1}, d_{t_2})$. This suggests a simple adaptation: spend more effort in the first round, perhaps by thinning more aggressively.

Another solution to this problem is to sample from a KDE of the previous distribution, rather than just resampling. Wilkinson mentions using a KDE instead of straight resampling, but he does not explain how important it is, and he does not mention any of the literature (for example [**?**]) that helps explain why and how to carry out a KDE in this setting. Using a KDE renders the method no longer exact, but according to [**?**], it works better than no KDE. Unfortunately, using a KDE in log-space with a normal kernel, sd of 0.05, yields a small cloud, centered tightly in the wrong place. This is not a genuine improvement over a point-mass posterior estimate. EMK: Write more here once results of KDE tests are in.

3.2 Potential flaws of the method and extensions to experiments

3.2.1 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 one day to run. As opposed to Matlab or R, Julia offers fast performance, including loops, to mimic that of a statically typed language. The implementation has been optimized using the Julia `@profile` macro, though it could likely be done more quickly by taking advantage of sparsity in the stoichiometry matrix. Wilkinson also notes computation time as an area for improvement, particularly when processing data from entire batches of cells undergoing the same experiment.

3.2.2 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 pint-size (more accurately, cubic-nanometer-size) model is not too great. The model is an upper bound on the accuracy of the method, because even if better models exist, the method is not useful for systems that go up to EMK: figure this out with little runs. Wilkinson stresses that much more information about his example network is available than is used, but offers no citations to show EMK: Look for papers on this! that his reduction works well.

3.2.3 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 is working properly, not even via citation or theoretical mixing time analysis. Meanwhile, other authors describe discarding far more of their samples as a burn-in, between 10 and 50 percent [8, 9]. EMK: Design experiments to confront this. Simple conjugate posterior? Discordant results between multiple runs?

3.2.4 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.

3.2.5 Prior Selection

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? EMK: Experiment with that once you finish the replication.

3.3 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. Now, in May of 2015, many strategies exist. Some, 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: predictive performance, tolerance for missing data, scaling, and uncertainty assessment.

Some schemes, such as [10], [**?**], and [11], 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, and for each set of candidate parameters, the system is solved numerically. In [10], the end result is a setup where quickly-solvable ODE's for mean and covariance give a Gaussian to approximate the true density, and [10] embeds this inside of a random-walk MCMC scheme. To distinguish this from LF-MCMC, it does evaluate an approximation to the likelihood: that is exactly what the moments are used for. MCMC also appears in [11], despite it falling in the "method of moments" category. The 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 [**?**] uses a penalty function involving higher moments, rather than a moment-matched approximate density. Despite being less interpretable in terms of statistics, 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, [**?**] cites many applied projects using it. W09 precedes all of this work.

Mean-field variational inference for Markov jump processes [12] scales well for large systems. This work precedes W09 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. W09's hidden state priors come implicitly from his parameter priors and initial state priors, meaning that in the hidden state, they treat only uncertainty due to intrinsic randomness, but in the parameters they treat uncertainty due to incomplete knowledge. Thus, ([12]) is a good option for inference when you want to fold in outside information about specific transitions.

To estimate parameters, [12] would need to be extended, perhaps as part of a variational EM algorithm. As written, the method produces only smoothed state estimates. Even variational EM does not give estimates of parameter uncertainty, so this work is not a viable alternative to W09 for scientific parameter estimation. For even more evidence that [12] cannot handle parameter estimation properly, the method performs well except for projection, and authors attribute this failure to untreated parameter uncertainty.

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 [8, 13]. This work came after W09 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 of a new dataset. Also, domain experts often use log-normal or log-uniform priors, as does W09, 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.

As far as MCMC, a good place to start might be [**?**], because it seems to use every trick in the book. One numerical trick in particular bears mention: sometimes, likelihood terms appearing in proposal ratios can be replaced with noisy but unbiased estimates [1] , and the target distribution will still be correct (details appear in [**?**]). This lets [**?**] embed an SMC estimate of the intractable discrete-data likelihood inside of a random-walk MCMC scheme, retaining exactness. The price: noisier approximations at each step make for a slower mixing chain overall. Furthermore, the SMC inside the main loop requires forward simulations from the Gillespie algorithm. Displeased with the computational cost of SMC incurred at each iteration, [**?**] 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. Other papers using SDE approximations or related techniques include [14] and [16].

Other non-exact methods include Approximate Bayesian Computation (ABC), which is compared with LF-MCMC in [17].

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 scheme in [**?**]. Thus, the paper [**?**] is 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 matrix spanning it is obtained; it suffices to multiply a random vector of integers into that matrix. This work in [**?**] is sophisticated enough to deal with partial measurements at a given time point, and it anticipates and deals with the criticism (e.g. [**?**]) that separate block updates to parameters and latent state can lead to low acceptance rates.

Some MCMC methods are exact, with a general theme of using approximations to generate good proposals. For examples, [18] uses a parallelized ABC algorithm and [19] explores several proposal-generating options, giving a particularly nice set of references in the process. The paper [9] is similar to Wilkinson's, using MCMC updates inside a sequential Monte Carlo scheme. They use a more complex model that treats batches of cells with some shared 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.

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

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] 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

[4] 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

[5] 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)

[6] 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

[7] 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

[8] 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

[9] 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

[10] 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

[11] 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

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

[13] 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

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

[15] Golightly, A., Wilkinson, D.J.a.: Bayesian parameter inference for stochastic biochemical network models using particle markov chain monte carlo. Interface Focus **1**(6) (10 2011) 807–820

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

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

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

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