Eric Kernfeld

Summary of Wilkinson's "Parameter inference for stochastic kinetic models of bacterial gene regulation", a book chapter in [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 EMK: Do I have to spell out "Markov Chain Monte Carlo"? 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 shows that a naive model, where the fluorescent reporter protein is proportional to the protein of interest, does not work. All tests are conducted using synthetic data on only one vector of true parameters.

I plan to implement the method in Julia and reproduce the experiments. To honestly test the method, I need to do more simulations, too: what happens when the log-space mean of the prior is not near the true values, or when the true values are somewhere other than Wilkinson's choice? What happens when all the parameters are unknown, rather than just the three we are interested in?

1 Introduction

1.1 Summary

This paper develops tools to study bacterial behavior, which is sometimes random: three bacilli in similar environments will act differently. It investigates a possible mechanism for this randomness: fluctuations in biochemical systems that regulate cell metabolism. In this paper and others, these systems are modeled using continuous-time Markov jump processes. This gives a rigorous treatment of stochastic dynamics. One remaining issue, the subject of this paper, is parameter inference for these jump processes using data that are noisy, discrete-time, and partial. The paper addresses this problem via Bayesian inference and MCMC.

1.2 Bare-Bones Biology

As a motivating case, Wilkinson uses the "decision" of *Bacillus subtilis* whether to become mobile. The paper centers around a gene encoding *flagellin*, which is a protein component of organelles that allow motility. Because biological networks can be disorienting, I'll relay some of the biological relationships in short sentences in the next paragraph.

The protein *flagellin* helps bacteria move. The protein σ^D promotes *flagellin*. The *fla / che* operon [1] contains many motility-related genes, including the one for σ^D. The protein σ^A and the protein $CodY$ both suppress the *fla / che* operon. Thus, they suppress σ^D, and they indirectly suppress *flagellin*. In fact, $CodY$ also downregulates *flagellin* directly. This is easiest to digest as a figure.

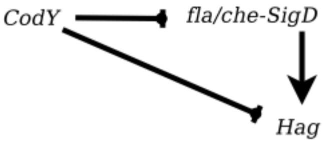

Figure 1: Regulatory relationships. Hag is the gene for $flagellin$, while $SigD$ encodes σ^D.

[1] Proteins are complex molecules that take myriad forms and roles within a cell. Operons are the basic transistor-like elements of the genome. In response to an outside stimulus, such as high levels of the sugar lactose, a normally active operon may become inactive or vice versa. By "active", I mean that the DNA encoded by the operon can be transcribed into RNA; this is the first step in the production of proteins.

2 The Model

Since we're simulating a biochemical system, suppose there are $X_j(t)$ particles of type j at time t, $j \in \{1...u\}$. These particles interact via a set of reactions $\mathcal{R}_i$, $i \in \{1...v\}$, with $\mathcal{R}_i$ consuming p_{ij} particles of type j and producing q_{ij} particles of type j. Let $R_i(t)$ denote the number of reactions of type i in $[0, t]$. I'll refer to these using the arrays $X(t)$, P, Q and once it $R_i(t)$ is defined, $R(t)$. Defining the matrix S to be $Q^T - P^T$, the model says that $X(t) - X(0) = SR(t)$. Under some assumptions EMK: maybe I should look into these, the different reaction channels evolve independently, and $R_i(t)$ is a Poisson process with intensity $c_i \int_0^t \prod_{j=1}^u \binom{X_j(t)}{p_{ij}}$. The c_i's are unknown.

Suppose n reactions occur between time 0 and time T. Suppose the ith one occurs at time t_i and suppose it has type ν_i. Define $t_0 \equiv 0$ and $t_{n+1} \equiv T$. When it comes to inference, the likelihood follows a competing-hazards model from continuous Markov chain theory[2].

$$P(\nu, t|c) = \prod_{i=1}^{n} c_{\nu_i} \prod_{j=1}^{u} \binom{X_j(t_{i-1})}{p_{\nu_i j}} \exp\left(-c_{\nu_i} \int_0^T \binom{X_j(t)}{p_{\nu_i j}} dt \right)$$

Wilkinson points out that $\int_0^T \binom{X_j(t_{i-1})}{p_{\nu_i j}} dt$ is tractable; in fact, I derived it as $\sum_{i=0}^n (t_{i+1} - t_i) \binom{X_j(t_i)}{p_{\nu_i j}}$.

If the number of reactions of type k is r_k, then the likelihood becomes

$$P(\nu, t|c) = \prod_{k=1}^{u} \left\{ c_k^{r_k} \left\{ \prod_{j=1}^{u} \binom{X_j(t_{i-1})}{p_{kj}}^{r_k} \right\} \exp\left(-c_k \int_0^T \sum_j \left(\binom{X_j(t)}{p_{kj}} \right) \right) dt \right\}.$$

Still calling it the likelihood, Wilkinson omits the bracketed term $\prod_{j=1}^u \binom{X_j(t_{i-1})}{p_{kj}}^{r_k}$, probably because it does not involve c. Regardless, the gamma distribution is conjugate for c_k, and setting independent priors $c_k \sim \Gamma(a_k, b_k)$, the posteriors are independent with $c_k \sim \Gamma(a_k + r_k, b_k + \int_0^T \sum_j \binom{X_j(t)}{p_{kj}} dt)$. This is an ideal scenario; missing data make inference more challenging.

3 Inference

In Wilkinson's data, not every reaction is recorded. Measurements are intermittent, and only some of the particle counts are measured. There is error in the measurements, too. Conjugate priors cannot be used because even computing the likelihood is infeasible. Wilkinson says it better than I can: "Consider first the best-case scenario–perfect observation of the system at discrete times. Conditional on discrete-time observations, the Markov process breaks up into a collection of independent bridge processes that appear not to be analytically tractable." He goes on to mention alternative strategies, which include the MCMC scheme of [2] and the approximation [3] paired with inference via EMK: cite more Wilkinson papers here.

The crux of the paper is a competitor to all of these. It builds upon the fact that exact simulations from this model are possible, and also on the fact that measurement error helps soften the requirements on where the bridge process should begin and end. Let θ include c, controlling the reaction rates, and τ, controlling the scale of measurement error. Let x denote the true state of the chain, but measured only at discrete times. The likelihood $P(x|\theta)$ is thought to be intractable. Let $\mathcal{D}$ be x measured with error and possibly with missing data for some particle types.

If we want to construct a Metropolis-Hastings scheme to sample from $P(x, \theta|\mathcal{D}) \propto P(\theta)P(x|\theta)P(\mathcal{D}|x, \theta)$ using a proposal $f(\theta^*, x^*|\theta, x)$, it works out that the acceptance ratio must be the min of 1 and

$$\frac{f(\theta^*, x^*|\theta, x)}{f(\theta, x|\theta^*, x^*)} \times \frac{P(\theta)}{P(\theta^*)} \times \frac{P(x|\theta)}{P(x^*|\theta^*)} \times \frac{P(\mathcal{D}|x, \theta)}{P(\mathcal{D}|x^*, \theta^*)}.$$

In the model Wilkinson considers, $P(\mathcal{D}|x, \theta)$ is simple, but $P(x|\theta)$ is not feasible to compute. The insight is that one can cancel the term $\frac{P(x|\theta)}{P(x^*|\theta^*)}$ by constructing a proposal that contains $\frac{P(x^*|\theta^*)}{P(x|\theta)}$ as a factor. This

[2] At times, I choose to use equations that stand free of sentences. These I do not punctuate.

spills straight out if instead of drawing both x and θ from out-of-the-box proposals, we draw $\theta^* \sim f(\theta^*|\theta)$ and compute x^* via simulation with parameters θ^*. The end result is that the ratio of interest simplifies:

$$\begin{aligned}
&\frac{f(\theta^*, x^*|\theta, x)}{f(\theta, x|\theta^*, x^*)} \times \frac{P(x|\theta)}{P(x^*|\theta^*)} \times \frac{P(\theta)}{P(\theta^*)} \times \frac{P(\mathcal{D}|x, \theta)}{P(\mathcal{D}|x^*, \theta^*)} \\
&= \frac{f(\theta^*|\theta)}{f(\theta|\theta^*)} \times \frac{P(x^*|\theta^*)}{P(x|\theta)} \times \frac{P(x|\theta)}{P(x^*|\theta^*)} \times \frac{P(\theta)}{P(\theta^*)} \times \frac{P(\mathcal{D}|x, \theta)}{P(\mathcal{D}|x^*, \theta^*)} \\
&= \frac{f(\theta^*|\theta)}{f(\theta|\theta^*)} \times \frac{P(\theta)}{P(\theta^*)} \times \frac{P(\mathcal{D}|x, \theta)}{P(\mathcal{D}|x^*, \theta^*)}.
\end{aligned}$$

To simplify further, use an independence sampler with the prior as a proposal and your calculation reduces to $\frac{P(\mathcal{D}|x,\theta)}{P(\mathcal{D}|x^*,\theta^*)}$.

If the measurement error is small, or $\mathcal{D}$ is high-dimensional, or both, this scheme leads to very high rejection rates, so it is not usable. Instead, you can break down X and $\mathcal{D}$, adding only one data point at a time. The exact approach is best understood by referring to Algorithm 1. In broad strokes, the procedure alternates between generating large samples from $P(\theta, x_{t_{1:i}}|\mathcal{D}_{t_{1:i}})$, done via likelihood-free MCMC, and folding in new data (incrementing i). To be clear, instead of running MCMC just once, in this method runs another five million steps through the sampler *for every time point in the dataset.* It isn't 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 1: 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 KDE of 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 $B_i$
```

In any M-H scheme, 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:

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

this shows how the compact form in Algorithm 1 arises. The Markov model would allow us to include or omit red terms at will. Note that using a KDE does not affect the ratio, because usually the kernel $K(\cdot|\cdot)$ is symmetric, so the extra term $\frac{K(\theta^*|\theta)}{K(\theta|\theta^*)}$ is just 1.

4 Experiments, Extensions and Further Checks

4.1 Experiments

Wilkinson runs four main simulations. All tests are conducted using synthetic data on only one vector of true parameters. The first three iterate through successively more difficult and realistic measurement models:

- First experiment: observe σ^D directly
- Second experiment: observe Hag rather than σ^D
- Third experiment: observe only GFP

The fourth shows that a naive model, assuming the fluorescent reporter protein is proportional to the protein of interest, leads to strong and incorrect claims in the posterior probabilities.

4.1.1 Details needed to reproduce experiments

Wilkinson describes a twelve-reaction regulatory model for these chemicals, spelling it out in table 1, and he lists three scientifically important reaction rates that, for tests of the inference method, will be treated as a "ground truth." The prior distributions cover 4 orders of magnitude, and they are uniform on a log scale. The experiment assumes D_t is the number of molecules observed with Gaussian error of standard deviation 10 molecules. EMK: Gaussian noise for integer data? Could do Bin$(p = 0.5, n = 40)$. What is the physical mechanism behind the measurement error? The initial state of the cell is assumed known EMK: but does he specify it?, and observations occur every 5 minutes (300 seconds) for 2 hours (7200 seconds).

4.2 Further checks

Some authors [4] claim that reaction rates range over seven orders of magnitude. Wilkinson's prior covers only four orders of magnitude. What happens when the log-space mean of the prior is not near the true values? In general, where does the prior information come from?

Wilkinson also treats parameters as known except for the three parameters of interest. How well does this method work when the rest of the reaction rates must be inferred, or when the measurement error must be inferred?

4.3 Extensions

Wilkinson mentions many areas for future work.

- His likelihood-free MCMC scheme applies to any hidden, discretely observed continuous Markov process. It also can be used for multiple time-series $\mathcal{D}^1, ..., \mathcal{D}^p$ by using posteriors from one as priors for the next.
- This method can fuse data from two bacterial strains, one missing a certain gene, by using posteriors from one experiment as priors for the next. It can also accomodate multiple fluorescent "reporters" for different molecules.
- Unsolved problems include scaling the method up to process data on batches of cells integrating data from microarrays or RNA-seq.

I also have some ideas for scaling.

- Could an EM algorithm handle the missing data more quickly?
- Do you really need such a large sample for every intermediate step?
- Could you build proposal distributions using variational methods?
- Could I build BEMC around this?

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] 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

[3] Gillespie, D.T.: The chemical langevin equation. The Journal of Chemical Physics **113**(1) (2000) 297–306

[4] Schlosshauer, M., Baker, D.: Realistic protein–protein association rates from a simple diffusional model neglecting long-range interactions, free energy barriers, and landscape ruggedness. Protein Science : A Publication of the Protein Society **13**(6) (06 2004) 1660–1669