Variational inference for Markov jump processes

Guido Sanguinetti

Joint work with Manfred Opper (T.U. Berlin)
Talk plan

- Markov jump processes in the life sciences
- Basics of jump processes
- Variational mean field
- Results
- Future extensions?
Markov Jump Processes

• Describe the dynamics of populations of interacting species
• Changes in populations are discrete (birth/death) and happen at random times
• Rates of births and deaths for a species depend on population levels for all species
• Many examples: chemical kinetics, predator-prey models, telecoms, etc.
Relevance to life sciences

- Environmental systems
- Single-cell dynamics
- Low concentration pathogens (see F. Bois talk)
- Usually dynamics is simulated using (variants of) Gillespie’s algorithm (SSA)
- SSA does not include observations and requires full specification of system
- Inference is hard: one MCMC approach (Boys et al; “A bloody mess”, A. Golightly, PESB2007)
Mathematical notation

- Mathematically, a Markov Jump Process (MJP) is a family of discrete random variables indexed by time $x(t)$.
- Markov property:
  \[ p(x(t_N) | x(t_{N-1}), \ldots, x(t_0)) = p(x(t_N) | x(t_{N-1})) \]
- Process rates $f(x'|x)$ defined by
  \[ \lim_{\delta t \to 0} p(x'(t+\delta t) | x(t)) = \delta_{x'x} + f(x'|x) \]
- Process rates satisfy $\sum f(x'|x) = 0$ (normalisation)
Master equation

- The process rates and the time-dependent probabilities $p(x)$ are linked by the Master equation

$$\frac{dp_t(x)}{dt} = \sum_{x' \neq x} \left[ -p_t(x) f(x'|x) + p_t(x') f(x'|x) \right].$$

- Notice that the Master equation is in fact a system of $S^D$ ODEs, with $S$ the number of states accessible to each species, and $D$ the number of species.
Variational Inference

• Deterministic approximate inference technique to approximate a probability distribution $p$ with a simpler one $q$.

• The measure of proximity between distributions is the Kullback-Leibler (KL) divergence

$$KL(q || p) = -\int dq \log \frac{p}{q}$$

• Notice that you do not need explicit knowledge of $p$, just of the expectation of $\log(p)$ under $q$
KL divergence between MJPs

• Consider two MJPs $p$ and $q$, with rates $f$ and $g$.
• A MJP can be viewed as a probability distribution over trajectories of the system; denote a trajectory as $x_{0:K}=(x(t_0),...,x(t_0+K\delta t))$.
• The KL divergence between these processes is

$$KL(q||p) = \sum_{x_{0:K}} q(x_{0:K}) \ln \frac{q(x_{0:K})}{p(x_{0:K})} =$$

$$= \sum_{k=0}^{K-1} \sum_{x_k} q(x_k) \sum_{x_{k+1}} q(x_{k+1}|x_k) \ln \frac{q(x_{k+1}|x_k)}{p(x_{k+1}|x_k)} + K_0$$
Continuous limit

• By taking the limit $\delta t \to 0$, we can rewrite the KL divergence in terms of process rates.

• Setting the initial KL $K_0 = 0$, we get

\[
KL (q \| p) = \int_0^T dt \sum_x q_t(x) \times \sum_{x', x' \neq x} \left\{ g(x' | x) \ln \frac{g(x' | x)}{f(x' | x)} + f(x' | x) - g(x' | x) \right\}
\]
Posterior processes

• We assume now to have discrete-time, noise-corrupted observations $y_l$ of the process.
• The noise model is given by $\hat{p}(y_l|x(t_l))$.
• The posterior process is still Markovian and is
  $$p_{post}(x_{0:K}) = \frac{1}{Z} p_{prior}(x_{0:K}) \times \prod_{l=1}^{N} \hat{p}(y_l|x(t_l)) .$$
• The KL divergence with a process $q$ is
  $$KL(q||p_{post}) = \ln Z + KL(q||p_{prior}) - \sum_{l=1}^{N} E_q [\ln p(y_l|x(t_l))] .$$
Mean-field approximation

- The main assumption we will make is that the approximating process $q$ is factorised

$$q_t(x) = \prod_{i=1}^{D} q_{it}(x_i) \quad g_t(x'|x) = \sum_{i=1}^{D} \prod_{j \neq i} \delta_{x'_j,x_j} g_{it}(x'_i|x_i)$$

- The KL-divergence then becomes

$$KL(q\|p_{post}) = \ln Z - \sum_{l=1}^{N} E_q [\ln \hat{p}(y_l|x(t_l))] +$$

$$\int_0^T dt \sum_i \sum_x q_{it}(x) \sum_{x':x' \neq x} \left\{ g_{it}(x'|x) \ln \frac{g_{it}(x'|x)}{\hat{f}_i(x'|x)} + \hat{f}_i(x'|x) - g_{it}(x'|x) \right\}$$

so that it decomposes as a sum over species $i$. 
Constraints

• Each of the factors in the approximating distribution satisfy a 1-dimensional Master equation.
• The complexity is decreased from $S^D$ to DS.
• Using Lagrange multiplier functions $\lambda_i(x,t)$, we obtain the Lagrangian
  
  $$L = KL(p_{\text{post}}) + 
  \sum_{i} \int_{0}^{T} dt \sum_{x} \lambda_i(x,t) \left( \frac{\partial}{\partial t} q_{il}(x) - \sum_{x' \neq x} \left\{ g_{il}(x|x') q_{il}(x') - g_{il}(x'|x) q_{il}(x) \right\} \right),$$

  with the standard boundary condition $\lambda_i(x,T)=0$. 
Functional derivatives

- To optimise the Lagrangian, we compute its functional derivatives

\[
\frac{\delta L}{\delta q_{i}(x)} = \sum_{x' \neq x} \left[ g_{it}(x'|x) \ln \frac{g_{it}(x'|x)}{\tilde{f}_{i}(x'|x)} - g_{it}(x'|x) + \tilde{f}_{i}(x'|x) \right] + \partial_{t}\lambda_{i}(x, t) + \\
\sum_{x'} g_{it}(x'|x) \left\{ \lambda_{i}(x', t) - \lambda_{i}(x, t) \right\} - \sum_{l} \ln \tilde{p}(y_{l}|x(t)) \delta(t - t_{l}) = 0
\]

(1)

\[
\frac{\delta L}{\delta g_{it}(x'|x)} = q_{it}(x) \left( \ln \frac{g_{it}(x'|x)}{\tilde{f}_{i}(x'|x)} + \lambda_{i}(x', t) - \lambda_{i}(x, t) \right) = 0
\]

(2)
Backward equation

• Inserting equation (2) into (1) and defining $r_i(x, t) = e^{-\lambda_i(x, t)}$ we obtain

$$\frac{dr_i(x, t)}{dt} = \sum_{x' \neq x} \left( \tilde{f}_i(x' | x) r_i(x, t) - \hat{f}_i(x' | x) r_i(x', t) \right)$$

• For each species, this is a system of $S$ linear differential equations valid at all times except the observation times.
Including observations

• We assume for simplicity that observations for different species are independent
  \[ \hat{p}(y_i|x(t)) = \prod_i \hat{p}_i(y_{il}|x_i(t_l)) \quad \forall l \]
  
• At observations, the Lagrange multiplier has a discontinuity
  \[ \lim_{t \to t_l^-} r_i(x, t) = \hat{p}_i(y_{il}|x_i(t_l)) \lim_{t \to t_l^+} r_i(x, t). \]
  
• Numerically, this can become difficult to solve.
• Easier to solve the ODE for the ratios of r.
Algorithm: approximating the posterior (E-step)

- We can find an approximate posterior as follows:
  1. Choose a species i;
  2. From an initial guess of $q_i(x)$, compute the averaged rates;
  3. Solve the backward equation backward in time from $r_i(x,T)=1$;
  4. Solve the Master equation forward in time to update $q_i(x)$;
  5. Iterate 2-4 until an optimum is found;
  6. Choose another species and follow 2-5;
  7. Iterate until convergence.

- This procedure guarantees a decrease in KL
Parameter estimation (M-step)

- Prior parameters can also be estimated by minimising the KL divergence, once an approximating distribution is obtained.
- E-step and M-step can be iterated until convergence is reached.
- Local minima can be a problem, thought must be given to initialisation.
Application: Lotka-Volterra

• Transition rates are given by

\[ f_{\text{prey}}(p + 1|p) = \alpha p \quad f_{\text{prey}}(p - 1|p) = \beta P p \]
\[ f_{\text{predator}}(P + 1|P) = \delta P p \quad f_{\text{predator}}(P - 1|P) = \gamma P \]

• M-step equations are given by

\[ \alpha = \frac{\int_0^T \langle \text{prey}_t (x + 1|x) \rangle_{\text{prey}_t}}{\int_0^T dt \langle x \rangle_{\text{prey}_t}} \]
\[ \beta = \frac{\int_0^T \langle \text{prey}_t (x - 1|x) \rangle_{\text{prey}_t}}{\int_0^T dt \langle x \rangle_{\text{prey}_t} \langle y \rangle_{\text{predator}_t}} \]
\[ \gamma = \frac{\int_0^T \langle \text{predator}_t (y - 1|y) \rangle_{\text{predator}_t}}{\int_0^T dt \langle y \rangle_{\text{predator}_t}} \]
\[ \delta = \frac{\int_0^T \langle \text{predator}_t (y + 1|y) \rangle_{\text{predator}_t}}{\int_0^T dt \langle y \rangle_{\text{predator}_t} \langle x \rangle_{\text{prey}_t}} \]
Results: Lotka-Volterra

Posterior predator (left) and prey (right) distribution. Diamonds are data, dashed-dotted posterior mode, solid posterior mean. The noise model is asymmetric and given by

\[
\tilde{p}_i(y_{il} | x_i(t_l)) \propto \left[ \frac{1}{5|y_{il} - x_i(t_l)|} + 0.001 \right]
\]
Parameter estimates

- Estimates of the parameters were reasonable

\[
\begin{align*}
\alpha &= 5.14 \times 10^{-4} \quad (5 \times 10^{-4}) \\
\beta &= 6.95 \times 10^{-5} \quad (1 \times 10^{-4}) \\
\gamma &= 7.26 \times 10^{-4} \quad (5 \times 10^{-4}) \\
\delta &= 5.77 \times 10^{-5} \quad (1 \times 10^{-4})
\end{align*}
\]

Estimates appear to converge to the true value when more observations are available. Results show average of five runs per data-set size.
Application: gene auto-regulation

- Autoregulation is one of the fundamental blocks in gene regulatory networks.
- Protein represses transcription of its own coding gene.
- We use a logical approximation. Process rates are

\[
\begin{align*}
\mathbf{f_{RNA}}(x+1|x, y) &= \alpha (1 - 0.99 \times \Theta(y - y_c)) \\
\mathbf{f_{RNA}}(x-1|x, y) &= \beta x \\
\mathbf{f_p}(y + 1|x, y) &= \gamma x \\
\mathbf{f_p}(y - 1|x, y) &= \delta y
\end{align*}
\]

The critical parameter \( y_c \) is the integer threshold for protein count above which repression begins. \( \Theta \) is the Heaviside step function.
Parameter estimation

- Fixed point equations for the parameters are obtained

\[
\alpha = \frac{\int_0^T dt \langle g_{RNA} (x + 1 | x) \rangle_{q_{RNA}}}{T \left(1 - 0.99 \frac{1}{T} \int h(y_c) \, dt\right)} \quad \beta = \frac{\int_0^T dt \langle g_{RNA} (x - 1 | x) \rangle_{q_{RNA}}}{\int_0^T dt \langle x \rangle_{q_{RNA}}}
\]

\[
\gamma = \frac{\int_0^T dt \langle g_p (x + 1 | x) \rangle_{q_p}}{\int_0^T dt \langle x \rangle_{q_{RNA}}} \quad \delta = \frac{\int_0^T dt \langle g_p (x - 1 | x) \rangle_{q_p}}{\int_0^T dt \langle y \rangle_{q_p}}
\]

where \( h(y_c) = \sum_{y \geq y_c} q_p(y) \) is the posterior probability that the protein levels will be above the critical threshold (function of time).
Identifiability of critical parameter

- The critical parameter $y_c$ is found by optimising (by search) the free energy

$$\mathcal{L}_{y_c} = \text{const} + \left\{ 2 \int_0^T dt \tilde{g} h(y_c) + \log \left[ 1 - 0.99 \frac{1}{T} \int_0^T h(y_c) \, dt \right] \int_0^T dt \tilde{g} \right\}$$

Clearly, if the protein levels in the data never exceed the threshold, $y_c$ is not identifiable. If it does, we get free energies with a well defined minimum.
Autoregulatory network: results

\[
\alpha = 3.4 \times 10^{-3} \quad (2 \times 10^{-3}) \\
\gamma = 1.9 \times 10^{-4} \quad (5 \times 10^{-4}) \\
\beta = 4.4 \times 10^{-5} \quad (6 \times 10^{-5}) \\
\delta = 6.7 \times 10^{-5} \quad (7 \times 10^{-5})
\]
Conclusions

• Efficient framework for posterior inference and parameter estimation in MJPs.
• Oustrips MCMC (Boys et al.) by orders of magnitude.
• Numerics are tricky and could be improved.
• Issues swept under the carpet: regularisation, cut-off on number of states.
• Readily extends to missing data.
Future work

• More complex, realistic networks.
• Missing data.
• Hybrid systems: one species has large numbers, so approximate as deterministic or diffusion.
• Hybrid systems: particles also have a spatial dimension, and diffuse in the environment.