

Andreas Rutter  
Guido Sanguinetti  
Manfred Opper

*We discuss the problem of statistical inference for Markov jump processes modeling biochemical reactions. Using a variational formulation of exact inference we derive two different approximations. A weak noise approach within a diffusion approximation is relevant when the number of individuals of a given species is rather large. On the other hand a mean field approximation takes the discreteness of the number of individuals into account but neglects correlations.*

---

## 10.1 Introduction

Recent technological advances are resulting in an increasing amount of information on the dynamics of cellular processes at the single cell level [Cai et al., 2006; Masamizu et al., 2006]. It is widely accepted that the low molecular copy numbers in these conditions mean that stochasticity can play a significant role in cellular dynamics [McAdams and Arkin, 1999], so that deterministic modelling is potentially inaccurate. This means that the learning problem is further complicated by the necessity of marginalizing the (unobserved) stochastic process, leading to potentially difficult inference problems.

In this chapter, we consider the stochastic modelling of the joint dynamics of groups (*species*) of interacting individuals using a particular class of stochastic processes known as *Markov jump processes*. These play an important role in a large number of application domains within systems biology, ranging from low count biochemical reactions to population models of interacting bacteria and macrophages. Exact statistical inference for such systems from a set of noisy data, *i.e.* the estimation of the true unobserved dynamics and the unknown parameters of the model,

quickly becomes intractable when the number of species/individuals is large. Hence, the development of approximation techniques is important.

In the following, we will derive equations for exact inference in Markov jump processes using a variational formulation [see *e.g.* Beal, 2003]. We will then develop two different techniques for approximating the exact inference which we expect to be useful for different applications. In a so-called weak noise expansion, we will assume that the number of individuals of a given species is large and number fluctuations are small. This allows us to use a continuum approximation resulting in equations of diffusion type. While the resulting multivariate Gaussian approximation allows us to keep correlations between species, the discreteness of the states is sacrificed. In a second type of approximation, we aim at keeping the discreteness, but sacrifice correlations. Here we use a mean field (MF) approach where the posterior process is approximated by a tractable factorizing process. This approach is better suited than the weak noise approximation, if the number of individuals of a given species is small and fluctuations can lead to qualitatively different results.

---

## 10.2 Markov Jump Processes

We consider a  $d$ -dimensional *stochastic process* in continuous time. In our examples, the values taken by  $X(t)$  will be restricted to the non-negative integers  $\mathbb{N}^d$ . The dimensionality  $d$  represents the number of (molecular) species present in the system; the components of the vector  $X(t)$  then represent the number of individuals of each species present at time  $t$ . We will model the dynamics as a Markov jump process (MJP). MJPs are a particular class of discrete stochastic processes exhibiting the Markov property in continuous time, *i.e.* the conditional probability of the state of the system at a time  $t$  given any sequence of states at previous times depends only on the most recent state. A MJP gives a probability measure over the infinite dimensional space of all possible *paths* of the system  $p(X_{0:T})$ . A MJP is characterized by its *process rates*  $f(X'|X)$ . The quantity  $f(X'|X)\Delta t$  determines the probability of transition from states  $X$  to states  $X'$  in an infinitesimal time interval  $\Delta t$ , *i.e.*

$$p(X'|X) \simeq \delta_{X',X} + \Delta t f(X'|X) \quad (10.1)$$

where  $\delta_{X',X}$  is the Kronecker delta and the equation becomes exact in the limit  $\Delta t \rightarrow 0$ . By normalization we get  $f(X|X) = -\sum_{X' \neq X} f(X'|X)$ . Another important quantity is the marginal probability  $p(X, t)$  of finding the system in state  $X$  at time  $t$ . Obviously, the marginal probability and the process rates are not independent. The relationship between them is given by the *Master equation* for the marginal probabilities

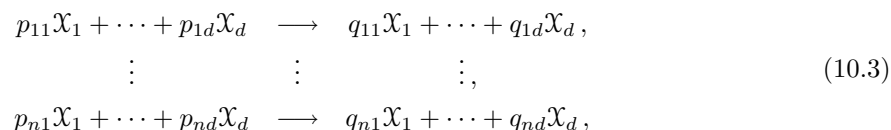
$$\frac{\partial}{\partial t} p(X, t) = \sum_{X' \neq X} \left( p(X', t) f(X|X') - p(X, t) f(X'|X) \right). \quad (10.2)$$

The Master equation is the discrete analog of the (forward) Fokker-Planck equation of diffusion systems and a special instance of the Chapman-Kolmogorov forward equation for continuous time Markovian processes. Its intuitive meaning is simple: at any time, the change in probability of finding the system in a specific state  $X$  is the probability of the system jumping in from another state minus the probability of the system jumping away from  $X$  to another state. In principle, inference could be based on the master equation directly if it could be solved. But usually it is intractable due to the size of the state space.

---

### 10.3 Reaction Systems

In a biochemical reaction system the model defining all possible jumps between states  $X$  is usually described as a set of chemical equations,



which states the species  $\mathcal{X}_j$  involved in the reactions as well as the stoichiometric factors  $p_{ij}$  and  $q_{ij}$  associated with reactants and products. Each time reaction  $i$  occurs, the number  $x_j$  of individuals of species  $\mathcal{X}_j$  is changed to

$$x'_j = x_j + q_{ij} - p_{ij}. \quad (10.4)$$

As a compact notation we use the  $d \times n$ -dimensional net effect reaction matrix  $S$  defined in Section 9.2, whose elements are given by  $S_{ji} = q_{ij} - p_{ij}$ . That way the jump caused by reaction  $i$  can be written in vector form,

$$X' = X + S\mathbf{e}_i, \quad (10.5)$$

where  $\mathbf{e}_i = (\delta_{i,1}, \delta_{i,2}, \dots, \delta_{i,n})^\top$  denotes the  $i$ -th base vector in the  $n$ -dimensional reaction space.

While  $S$  defines all possible reactions and their effects, we also need the rate vector  $\mathbf{h} = (h_1, h_2, \dots, h_n)^\top$  in order to describe the dynamics of the system. Each reaction rate  $h_i$  is a function of the current system state  $X$  and a rate constants  $c_i$ . For elementary chemical reactions we can usually assume mass action kinetics. In this case the rate law

$$h_i(X) = c_i \prod_{j=1}^d \prod_{k=0}^{p_{ij}-1} (x_j - k) \quad (10.6)$$

only depends on the numbers  $p_{ij}$  of reactants involved in a reaction, so that one can deduce it directly from the chemical equation. But as our methods do not take the special form of (10.6) into account, they also work with arbitrary rate laws,

such as those for fractal kinetics [Kopelman, 1988; Savageau, 1998].

Each reaction model defined by  $S$  and  $\mathbf{h}(X)$  corresponds to a MJP with process rates given by

$$f(X'|X) = \sum_{i=1}^n \delta_{X'-X, S\mathbf{e}_i} h_i(X) \quad (10.7)$$

for  $X' \neq X$ . Consequently, a reaction system is just a special case of a MJP. However, it is often easier to specify  $S$  and  $\mathbf{h}(X)$  instead of the process rate  $f(X'|X)$  for all possible jumps between different states.

## 10.4 A Variational Formulation of Inference

We assume that we have noisy observations  $Y_l$  ( $l = 1, \dots, N$ ) of the state of the system at discrete time points; the noise model is specified by a likelihood function  $\hat{p}(Y_l|X(t_l))$ . We can combine this likelihood with the prior process to obtain a posterior process. We will be working with probability measures over paths of the process. To simplify the treatment we will assume that time is discretized into small intervals of size  $\Delta t$  and represent such a trajectory as  $X_{0:K} = (X(t_0), \dots, X(t_0 + K\Delta t))$  where  $K$  is very large. Hence, we write the joint posterior probability as

$$p_{\text{post}}(X_{0:K}) = \frac{1}{Z} p_{\text{prior}}(X_{0:K}) \times \prod_{l=1}^N \hat{p}(Y_l|X(t_l)) \quad (10.8)$$

with

$$p_{\text{prior}}(X_{0:K}) = p(X_0) \prod_{k=0}^{K-1} p(X_{k+1}|X_k) \quad (10.9)$$

and  $Z = p(Y_1, \dots, Y_N)$ . Note that  $X(t_l) \in X_{0:K}$ .

Assuming the observation noise at different time points to be independent, it can be shown that the posterior process is still a MJP with a (time dependent) rate function. We will show that this rate function can be computed within a variational formulation. As is well known, the posterior minimizes the Kullback-Leibler (KL) divergence

$$\text{KL}[q \| p_{\text{post}}] = \log Z + \text{KL}[q \| p_{\text{prior}}] - \sum_{l=1}^N \left\langle \log \hat{p}(Y_l|X(t_l)) \right\rangle_q \quad (10.10)$$

with respect to possible measures over paths  $q$ .

The KL divergence between two MJPs defined by their path probabilities  $p(X_{0:K})$

and  $q(X_{0:K})$  is found to be

$$\begin{aligned} \text{KL}[q \parallel p] &= \sum_{X_{0:K}} q(X_{0:K}) \log \frac{q(X_{0:K})}{p(X_{0:K})} \\ &= \sum_{k=0}^{K-1} \sum_{X_k} q(X_k) \sum_{X_{k+1}} q(X_{k+1}|X_k) \log \frac{q(X_{k+1}|X_k)}{p(X_{k+1}|X_k)} + K_0 \end{aligned} \quad (10.11)$$

where we have used the Markov property of both processes. Assuming fixed initial marginal probabilities  $K_0 = \sum_{X_0} q(X_0)(\log q(X_0) - \log p(X_0))$  will be set to zero in the following. We use the short time scaling (10.1) for the conditional probabilities and let  $\Delta t \rightarrow 0$ ,  $K \rightarrow \infty$  with  $K \Delta t \rightarrow T$ , replace sums by integrals and obtain

$$\begin{aligned} \text{KL}[q \parallel p_{\text{prior}}] &= \int_0^T dt \sum_X q(X, t) \sum_{X': X' \neq X} \\ &\quad \left( g_t(X'|X) \log \frac{g_t(X'|X)}{f(X'|X)} + f(X'|X) - g_t(X'|X) \right) \end{aligned} \quad (10.12)$$

where  $f(X'|X)$  and  $g_t(X'|X)$  are the rates of the  $p_{\text{prior}}$  and  $q$  process respectively. After minimizing the KL divergence given in (10.10),  $g_t(X'|X)$  becomes the rate function of the posterior process  $p_{\text{post}}$ .

## 10.5 Exact Inference

Exact inference for MJPs is based on a generalization of the classic forward-backward algorithm for hidden Markov models (HMMs) to continuous time stochastic processes. We have already seen that the Master equation represents the forward equation of the system. We will now obtain a backward equation which allows us to compute the rates of the posterior process.

To find the rate  $g_t(X'|X)$  of the posterior process we minimize the KL divergence (10.12) jointly with respect to  $g_t(X'|X)$  and the marginal probabilities  $q(X, t)$ . To take their dependence, which is given by the Master equation (10.2), into account, we introduce a Lagrange multiplier function  $\lambda(X, t)$  and compute the stationary values of the Lagrangian

$$\begin{aligned} L &= \text{KL}[q \parallel p_{\text{post}}] - \int_0^T dt \sum_X \lambda(X, t) \\ &\quad \times \left[ \frac{\partial}{\partial t} q(X, t) - \sum_{X' \neq X} \left( g_t(X|X') q(X', t) - g_t(X'|X) q(X, t) \right) \right]. \end{aligned} \quad (10.13)$$

Taking functional derivatives of (10.13) with respect to  $q(X, t)$  and  $g_t(X'|X)$  yields

$$\begin{aligned} \frac{\delta L}{\delta q(X, t)} &= \sum_{X' \neq X} \left( g_t(X'|X) \log \frac{g_t(X'|X)}{f(X'|X)} - g_t(X'|X) + f(X'|X) \right) \\ &+ \frac{\partial}{\partial t} \lambda(X, t) + \sum_{X'} g_t(X'|X) (\lambda(X', t) - \lambda(X, t)) \\ &- \sum_l \log \hat{p}(Y_l|X(t)) \delta(t - t_l) = 0, \end{aligned} \quad (10.14)$$

$$\frac{\delta L}{\delta g_t(X'|X)} = q_t(X) \left( \log \frac{g_t(X'|X)}{f(X'|X)} + \lambda(X', t) - \lambda(X, t) \right) = 0. \quad (10.15)$$

Equation (10.15) then becomes

$$\frac{g_t(X'|X)}{f(X'|X)} = \frac{r(X', t)}{r(X, t)} \quad (10.16)$$

where we have defined  $r(X, t) = e^{-\lambda(X, t)}$ . Inserting equation (10.16) into (10.14), we obtain the linear (backward) differential equation

$$\frac{\partial}{\partial t} r(X, t) = \sum_{X' \neq X} f(X'|X) (r(X, t) - r(X', t)) \quad (10.17)$$

together with a set of jump conditions at the observations

$$\lim_{t \rightarrow t_l^-} r(X, t) = \hat{p}(Y_l|X(t_l)) \lim_{t \rightarrow t_l^+} r(X, t). \quad (10.18)$$

Using the equations of motion and an integration by parts we can show that

$$-\log p(Y_1, \dots, Y_N) = -\log Z = \sum_X p(X, t_1) \lambda(X, t_1). \quad (10.19)$$

This result can be used for a maximum likelihood estimation of system parameters. If we truncate the number of relevant states to  $R$ , (10.17) is a system of  $R^d$  coupled linear equations. Rather than attempting an exact solution of this system we will discuss two types of more tractable approximations in the following sections.

## 10.6 Weak Noise Approximation

We will next derive a weak noise approximation for inference in Markov jump processes which is in the spirit of van Kampen's expansion [van Kampen, 1981; Gardiner, 1996]. The method applies well to the case where the number of molecules of each type is large. In this limit, the number fluctuations, representing the internal "noise", are small and can be treated within a Gaussian diffusion approximation. We begin by approximating the backward equation (10.17). Writing  $r(X', t) = r(X + \epsilon(X' - X), t)$  we introduce a formal expansion parameter  $\epsilon$  which incorporates our assumption that changes of the state  $X' - X$  are much smaller than the typical

value for the state itself. Inserting this ansatz into (10.17), and expanding to second order in  $\epsilon$  we obtain

$$\left[ \frac{\partial}{\partial t} + \epsilon \mathbf{f}(X)^\top \nabla + \frac{1}{2} \epsilon^2 \text{tr} (D(X) \nabla \nabla^\top) \right] r(X, t) = 0 \quad (10.20)$$

where drift vector  $\mathbf{f}(X)$  and diffusion matrix  $D(X)$  are defined by the jump moments

$$\mathbf{f}(X) = \sum_{X' \neq X} f(X'|X)(X' - X), \quad (10.21)$$

$$D(X) = \sum_{X' \neq X} (X' - X) f(X'|X) (X' - X)^\top. \quad (10.22)$$

In the case of a reaction model we can use (10.7) in order to obtain  $\mathbf{f}(X)$  and  $D(X)$  in terms of  $S$  and  $\mathbf{h}(X)$ . This finally leads to

$$\mathbf{f}(X) = S \mathbf{h}(X) \quad D(X) = S \mathbf{h}(X) S^\top. \quad (10.23)$$

This drift vector and diffusion matrix correspond to the same diffusion approximation of the reaction process as the one presented in the chapter of Golightly and Wilkinson (Section 9.2.2). Following the discussions on the so-called system size expansion in van Kampen [1981]; Gardiner [1996], we can argue that the proper lowest order terms in the weak noise limit requires a further expansion. We use the fact that typical state vectors  $X$  are expected to be close to some dominant time dependent state  $\mathbf{b}(t)$ , *i.e.* we write  $X = \mathbf{b}(t) + \epsilon \mathbf{y}$  and set  $r(X, t) = \Psi(\mathbf{y}, t)$ . Again we expand up to terms of order  $\epsilon^2$ . If we require that

$$\frac{d\mathbf{b}}{dt} = \mathbf{f}(\mathbf{b}(t)) \quad (10.24)$$

then terms of order  $\epsilon$  cancel and we find (setting  $\epsilon = 1$  at the end)

$$\left[ \frac{\partial}{\partial t} + \mathbf{y}^\top A(X)^\top (\mathbf{b}(t)) \nabla + \frac{1}{2} \text{tr} (D(\mathbf{b}(t)) \nabla \nabla^\top) \right] \Psi(\mathbf{y}, t) = 0 \quad (10.25)$$

where the matrix  $A(X)$  is defined by

$$A_{ij}(X) = \frac{\partial f_i}{\partial x_j}. \quad (10.26)$$

This equation is easily solved by

$$r(X, t) \approx \eta(t) \exp \left( -\frac{1}{2} (X - \mathbf{b}(t))^\top B^{-1}(t) (X - \mathbf{b}(t)) \right) \quad (10.27)$$

where

$$\frac{dB}{dt} = A(\mathbf{b}(t))B(t) + B(t)A(\mathbf{b}(t))^\top - D(\mathbf{b}(t)) \quad (10.28)$$

and the normalization  $\eta(t)$  is given by

$$\frac{d\eta}{dt} = \eta(t) \operatorname{tr} (A(\mathbf{b}(t))). \quad (10.29)$$

We could use (10.27) in (10.16) to compute the posterior rate function  $g_t(X'|X)$ . However, we aim for a simple representation of the approximate posterior distribution  $q(X, t)$ . Thus it is more useful to apply a similar weak noise expansion for the posterior master equation and compute the corresponding drift vector

$$\mathbf{g}(X, t) = \sum_{X' \neq X} g_t(X'|X)(X' - X) = \sum_{X' \neq X} (X' - X) \frac{r(X', t)}{r(X, t)} f(X'|X). \quad (10.30)$$

By doing so we get

$$\mathbf{g}(X, t) \approx \mathbf{f}(X) + D(\mathbf{b}(t)) \nabla \log r(X, t) \quad (10.31)$$

$$\approx \mathbf{f}(X) - D(\mathbf{b}(t)) B^{-1}(t) (X - \mathbf{b}(t)), \quad (10.32)$$

where in the last step we have expanded  $r(X, t)$  to first order around  $X$ . The expansion proceeds in a similar way as the one for the backward equation. We end up with a Fokker-Planck equation for  $q(X, t)$  with a linearized drift term which is solved by the multivariate Gaussian

$$q(X, t) \approx (2\pi)^{-d/2} (\det C(t))^{-1/2} \times \exp \left( -\frac{1}{2} (X - \mathbf{m}(t))^\top C(t)^{-1} (X - \mathbf{m}(t)) \right), \quad (10.33)$$

where the mean state vector is a solution to

$$\frac{d\mathbf{m}}{dt} = \mathbf{g}(\mathbf{m}(t)) \quad (10.34)$$

and the covariance evolves according to

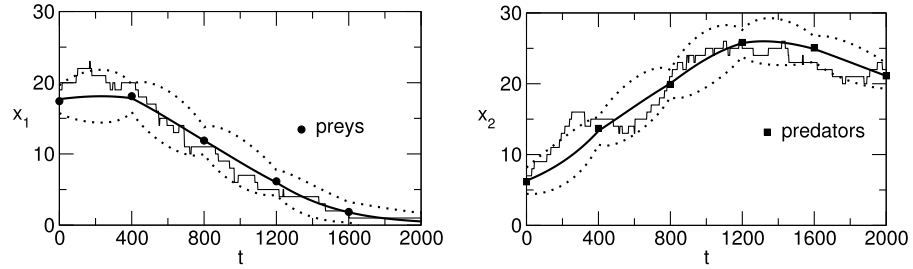
$$\frac{dC}{dt} = H(\mathbf{m}(t), t) C(t) + C(t) H(\mathbf{m}(t), t)^\top + D(\mathbf{m}(t)) \quad (10.35)$$

and where the matrix  $H$  is defined by

$$H_{ij}(X, t) = \frac{\partial g_i}{\partial x_j}. \quad (10.36)$$

In order to test our method we have applied it to data generated by the Lotka-Volterra model [Boys et al., 2008; Gilioli et al., 2008], which is a rather simple, but non-trivial, reaction system. It consists of two interacting species, traditionally named preys and predators. In this system four reactions are possible, as both preys  $\mathcal{X}_1$  and predators  $\mathcal{X}_2$  can be created or destroyed at any point in time. The rate laws are given by

$$\begin{aligned} h(\mathcal{X}_1 \rightarrow 2\mathcal{X}_1) &= \alpha x_1, & h(\mathcal{X}_1 \rightarrow \emptyset) &= \beta x_1 x_2, \\ h(\mathcal{X}_2 \rightarrow 2\mathcal{X}_2) &= \delta x_1 x_2, & h(\mathcal{X}_2 \rightarrow \emptyset) &= \gamma x_2, \end{aligned} \quad (10.37)$$



**Figure 10.1** Inference results for the Lotka-Volterra process with  $\alpha = 0.0005$ ,  $\beta = 0.0001$ ,  $\gamma = 0.0005$ ,  $\delta = 0.0001$ ,  $x_1(0) = 19$ ,  $x_2(0) = 7$ , and  $\sigma = 1$ . Symbols denote the noisy observations, thick lines show the posterior mean, and the 95% confidence region is surrounded by dotted lines. For comparison the original process is plotted as thin line.

where  $x_1$  denotes the number of prey and  $x_2$  the number of predators.

While the input of our algorithm was generated by a Lotka-Volterra process with fixed initial conditions  $x_1(0), x_2(0)$ , this knowledge was not used for inference purposes, as we assumed a flat prior. The observations were obtained at regular intervals and corrupted by Gaussian noise with standard deviation  $\sigma$ .

Figure 10.1 shows results of state inference for a single Lotka-Volterra process, which has been simulated using Gillespie’s algorithm [Gillespie, 1976]. It is clearly visible that the original process, which generated the observations, lies inside the inferred confidence interval most of the time. To estimate the quality of the inference algorithm quantitatively, we obtained the empirical distribution of the random variables

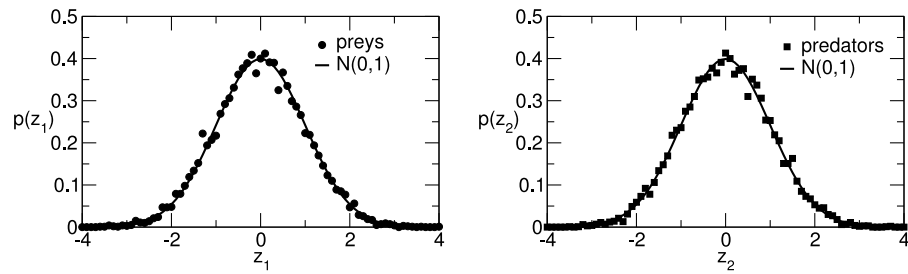
$$z_j(t) = \frac{x_j(t) - m_j(t)}{\sigma_j(t)} \tag{10.38}$$

by repeating this simulation with different sets of observations but unchanged parameters. As these quantities are the differences between the state  $X(t)$  of the original process and the maximum a posteriori prediction  $\mathbf{m}(t)$  rescaled by the calculated standard deviation  $\sigma_j(t) = \sqrt{C_{jj}(t)}$ , one expects that  $\mathbf{z}(t)$  has a normal Gaussian distribution. The results shown in Figure 10.2 confirm that the prediction is indeed well calibrated. Consequently, this method gives reliable information about the state of the reaction system, although the internal noise is rather large in this case due to the small number of prey and predators.

Parameter estimation is done by maximizing the log-likelihood of the data  $\log Z = \log p(Y_1, \dots, Y_N)$  from (10.19) using a standard gradient-based method. In the weak noise approximation this quantity is given by

$$\log Z = \log \int dX r(X, t_1) = \frac{d}{2} \log(2\pi) + \frac{1}{2} \log \det B(t_1) + \log \eta(t_1). \tag{10.39}$$

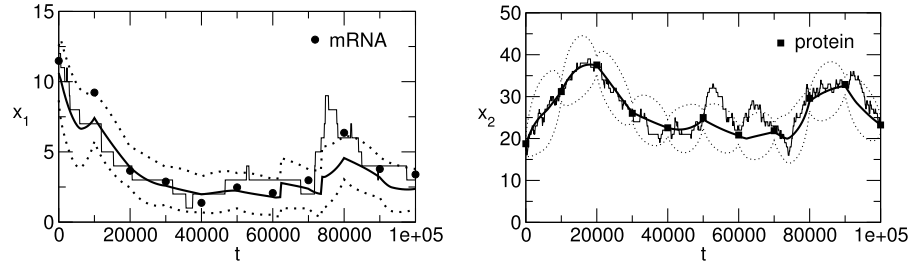
This works well, as shown in Table 10.1. The inferred values are reasonable close to the real values of the reaction constants. Additionally, this method is



**Figure 10.2** Distribution of the prediction error  $\mathbf{z}(t = 1000)$  for the Lotka-Volterra process with parameters given in Figure 10.1. Symbols denote the empirical results of the state inference algorithm applied to 1000 different data sets. For comparison the normal Gaussian distribution is shown as solid line.

**Table 10.1** Type II maximum likelihood estimates of reaction constants based on 11 observations at equidistant points in the time interval  $[0; 1000]$  with  $\sigma = 1$ . Mean values and standard deviations have been obtained by averaging over 600 samples from a Lotka-Volterra process.

parameter	inference result	real value
$\alpha$	$(6.5 \pm 5.4) \times 10^{-4}$	$5 \times 10^{-4}$
$\beta$	$(1.0 \pm 0.6) \times 10^{-4}$	$1 \times 10^{-4}$
$\gamma$	$(8.1 \pm 6.4) \times 10^{-4}$	$5 \times 10^{-4}$
$\delta$	$(0.9 \pm 0.4) \times 10^{-4}$	$1 \times 10^{-4}$



**Figure 10.3** Inference results for the genetic autoregulatory network with  $x_c = 20$ ,  $\alpha = 2 \times 10^{-3}$ ,  $\beta = 6 \times 10^{-5}$ ,  $\gamma = 5 \times 10^{-4}$ ,  $\delta = 7 \times 10^{-5}$ ,  $x_1(0) = 12$ ,  $x_2(0) = 17$ , and  $\sigma = 1$ . Symbols denote the noisy observations, thick lines show the posterior mean, and the 95% confidence region is surrounded by dotted lines. For comparison the original process is plotted as thin line.

quite fast, because only one backward integration is needed in each step of the minimization process. In fact, a single type II maximum likelihood estimate of the four reaction constants takes less than a minute, nearly independent of the number of observations.

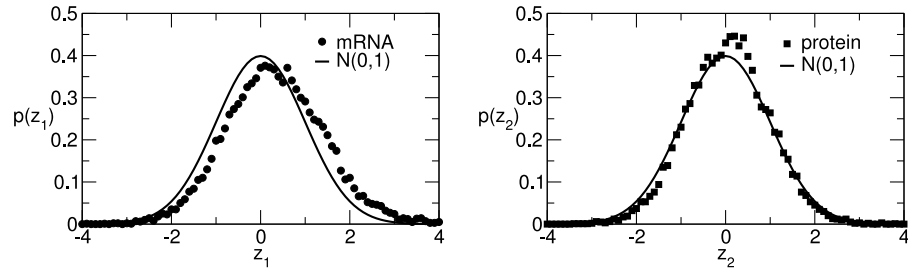
As a second example, we used a simple genetic autoregulatory network [Opper and Sanguinetti, 2007] as a test case for our inference method. Such a reaction system can be found as parts of the transcriptional regulatory network in biological cells. It again consists of two species, mRNA ( $\mathcal{X}_1$ ) and protein ( $\mathcal{X}_2$ ), and the rate laws

$$\begin{aligned} h(\emptyset \rightarrow \mathcal{X}_1) &= \alpha(1 - 0.99 \times \Theta(x_2 - x_c)), & h(\mathcal{X}_1 \rightarrow \emptyset) &= \beta x_1, \\ h(\emptyset \rightarrow \mathcal{X}_2) &= \gamma x_1, & h(\mathcal{X}_2 \rightarrow \emptyset) &= \delta x_2, \end{aligned} \quad (10.40)$$

where  $\Theta$  is the Heaviside step function. Both mRNA and proteins decay exponentially and proteins are produced by translation of mRNA which occurs proportional to the mRNA number  $x_1$ . In order to regulate the concentration of the protein, the production of mRNA is down-regulated significantly as soon as  $x_2$  increases beyond a critical parameter  $x_c$ .

Although this model is strongly non-linear because of the switching of the transcription process producing the mRNA, state inference works well as shown in Figure 10.3. However, it is clearly visible in Figure 10.4 that our inference method has a small bias for  $z_1$  due to the discontinuous rate law  $h(\emptyset \rightarrow \mathcal{X}_1)$  given in (10.40).

Estimates of the reaction constants are given in Table 10.2. As the discrete nature of the critical threshold  $x_c$  rules out gradient-based optimization methods, we had to use the Nelder-Mead simplex method [Nelder and Mead, 1965], which works without computing derivatives. Additionally, the weak noise approximation assumes linear behavior of the drift for small deviations. Although this assumption is not valid for  $h(\emptyset \rightarrow \mathcal{X}_1)$ , all inferred parameters are close to the real values. Consequently, the quality of inference is not affected much by the strong non-linear behavior of this genetic autoregulatory network.



**Figure 10.4** Distribution of the prediction error  $\mathbf{z}(t = 10^5)$  for the genetic autoregulatory network with parameters given in Figure 10.3. Symbols denote the empirical results of the state inference algorithm applied to 1000 different data sets. For comparison the normal Gaussian distribution is shown as solid line.

**Table 10.2** Type II maximum likelihood estimates of reaction constants based on 11 observations with  $\sigma = 1$ . Mean values and standard deviations have been obtained by averaging over 1000 samples from the genetic autoregulatory network model.

parameter	inference result	real value
$\alpha$	$(1.6 \pm 0.5) \times 10^{-3}$	$2 \times 10^{-3}$
$\beta$	$(6.5 \pm 1.0) \times 10^{-5}$	$6 \times 10^{-5}$
$\gamma$	$(4.9 \pm 0.7) \times 10^{-4}$	$5 \times 10^{-4}$
$\delta$	$(7.7 \pm 1.2) \times 10^{-5}$	$7 \times 10^{-5}$
$x_c$	$20.4 \pm 1.6$	20

## 10.7 Mean Field Approximation to Posterior MJPs

In this section we will discuss a type of approximation where the minimization of the KL divergence (10.10) which characterizes the exact posterior is carried out in a restricted class of tractable probability distributions [Opper and Sanguinetti, 2007]. We will choose this approximating process  $q$  to factorize into products of path probabilities for individual species. It can easily be seen that this *mean field ansatz* translates into the following conditions on marginals and posterior transition rates

$$q(X, t) = \prod_{i=1}^D q_{it}(x_i) \quad g_t(X'|X) = \sum_{i=1}^d \prod_{j \neq i} \delta_{x'_j, x_j} g_{it}(x'_i|x_i). \quad (10.41)$$

The KL divergence between the approximating process and the posterior process is then found to be

$$\begin{aligned} \text{KL}[q \parallel p_{\text{post}}] &= \log Z - \sum_{l=1}^N \left\langle \log \hat{p}(Y_l|X(t_l)) \right\rangle_q + \int_0^T dt \sum_i \sum_{x_i} q_{it}(x_i) \sum_{x'_i: x'_i \neq x_i} \\ &\quad \left\{ g_{it}(x'_i|x_i) \log \frac{g_{it}(x'_i|x_i)}{\hat{f}_i(X'|X)} + \tilde{f}_i(X'|X) - g_{it}(x'_i|x_i) \right\} \end{aligned} \quad (10.42)$$

where we have defined

$$\begin{aligned} \hat{f}_i(X'|X) &= \exp \left( \left\langle \log f_i(X'|X : x'_j = x_j, \forall j \neq i) \right\rangle_{X \setminus i} \right) \\ \tilde{f}_i(X'|X) &= \left\langle f_i(X'|X : x'_j = x_j, \forall j \neq i) \right\rangle_{X \setminus i}. \end{aligned} \quad (10.43)$$

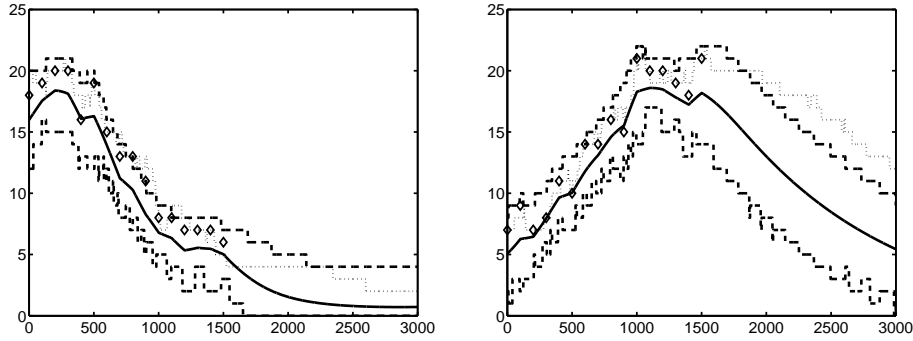
Again, we will minimize the KL divergence (10.42) with respect to the marginals  $q_{it}(x_i)$  and the rates  $g_{it}(x'_i|x_i)$  taking the Master equation (10.2) into account. As before we introduce Lagrange multiplier functions  $\lambda_i(x_i, t)$  and find that  $r_i(x_i, t) = e^{-\lambda_i(x_i, t)}$  fulfils the system of linear differential equations

$$\frac{\partial}{\partial t} r_i(x_i, t) = \sum_{X' \neq X} \left( \tilde{f}_i(X'|X) r_i(x_i, t) - \hat{f}_i(X'|X) r_i(x'_i, t) \right). \quad (10.44)$$

The Lagrangian (10.13) can be optimized iteratively. Starting with an initial guess for  $q_{it}(x_i)$  and selecting a species  $i$ , we compute  $\hat{f}_i(X'|X)$  and  $\tilde{f}_i(X'|X)$ . Using these, (10.44) is solved backwards starting from the condition

$$r_i(x_i, T) = 1 \forall x \quad (10.45)$$

(*i.e.*, the constraint becomes void at the final time.). As before, the observation probabilities will translate into jump conditions on the  $r_i$ s as in (10.18). Solving (10.44) allows us to update our estimate of the rates  $g_{it}(x'_i|x_i)$  using equation (10.45), which can then be used to solve the master equation (10.2) and update the guess of  $q_{it}(x_i)$ . This procedure is followed sequentially for all the species and



**Figure 10.5** Inference results for the Lotka-Volterra process: approximate posterior distribution of preys (left) and predators (right). The true parameter values are  $\alpha = 0.0005$ ,  $\beta = 0.0001$ ,  $\gamma = 0.0005$ ,  $\delta = 0.0001$ ,  $x_1(0) = 19$ ,  $x_2(0) = 7$ , and  $\sigma = 1$ . Vertical axis indicates discrete numbers in the populations, horizontal axis is time. Symbols denote the noisy observations, thick lines show the posterior mean, and the 95% confidence region is surrounded by dashed lines. For comparison the original process is plotted as a thin dotted line.

iterated until a stationary point of the Lagrangian is found. There are some minor numerical stability issues with the procedure, analogous to those arising in the forward-backward algorithm for discrete time HMMs. First of all, the independent variables in (10.44) are positive quantities, being the exponentials of the Lagrange multipliers. However, the jump conditions (10.18) mean that  $r_i(x'_i|x_i)$  can become extremely small when  $x_i$  is far from the observed value. For this reason, it is better to solve the equation in log space by deriving an analogous equation for the ratios  $r_i(x'_i|x_i)/r_i(x''_i|x'_i)$ . Secondly, the mean field rates (10.43) involve computing the logarithm of the process rates for the prior process. As some process rates may be very small or zero, this can lead to numerical instabilities. In practice, we regularize the mean field rates by replacing  $\log(0)$  with a large (but finite) negative number.

Since  $\text{KL}[q \| p_{\text{post}}] \geq 0$ , we obtain, as useful by-product of the MF approximation, a tractable variational lower bound on the log-likelihood of the data  $\log Z = \log p(Y_1, \dots, Y_N)$  from (10.42). As usual [see *e.g.* Opper and Saad, 2001] such a bound can be used in order to optimize model parameters using a variational EM algorithm. In the case of the Lotka-Volterra model, this results in intuitive fixed points equations

$$\alpha = \frac{\int_0^T \langle g_{\text{preyt}}(x+1|x) \rangle_{\text{preyt}}}{\int_0^T \langle x \rangle_{\text{preyt}}}, \quad \beta = \frac{\int_0^T \langle g_{\text{preyt}}(x-1|x) \rangle_{\text{preyt}}}{\int_0^T \langle x \rangle_{\text{preyt}} \langle y \rangle_{\text{predatort}}}, \quad (10.46)$$

$$\gamma = \frac{\int_0^T \langle g_{\text{predatort}}(y-1|y) \rangle_{\text{predatort}}}{\int_0^T \langle y \rangle_{\text{predatort}}}, \quad \delta = \frac{\int_0^T \langle g_{\text{predatort}}(y+1|y) \rangle_{\text{predatort}}}{\int_0^T \langle y \rangle_{\text{predatort}} \langle x \rangle_{\text{preyt}}}. \quad (10.47)$$

**Table 10.3** Type II maximum likelihood estimates of reaction constants based on 15 observations. Mean values and standard deviations have been obtained by averaging over 25 Lotka-Volterra processes using equal reaction constants and  $x_1(0) = 19$ ,  $x_2(0) = 7$  as initial conditions.

parameter	inference result	real value
$\alpha$	$(4.6 \pm 3.4) \times 10^{-4}$	$5 \times 10^{-4}$
$\beta$	$(1.2 \pm 0.3) \times 10^{-4}$	$1 \times 10^{-4}$
$\gamma$	$(7.1 \pm 2.1) \times 10^{-4}$	$5 \times 10^{-4}$
$\delta$	$(1.2 \pm 0.2) \times 10^{-4}$	$1 \times 10^{-4}$

To show the performance of the approximation, we again test it on the Lotka-Volterra process with the same initial conditions and parameter values as in Figure 10.1. The maximum number of predators and preys allowed was capped at 200, which greatly exceeds the maximum numbers of individuals the system could achieve in the time frame under consideration (exponential growth of the preys in the absence of predators).

As both our observations and our process are discrete, a Gaussian noise model would not be appropriate. We chose the following observation noise model

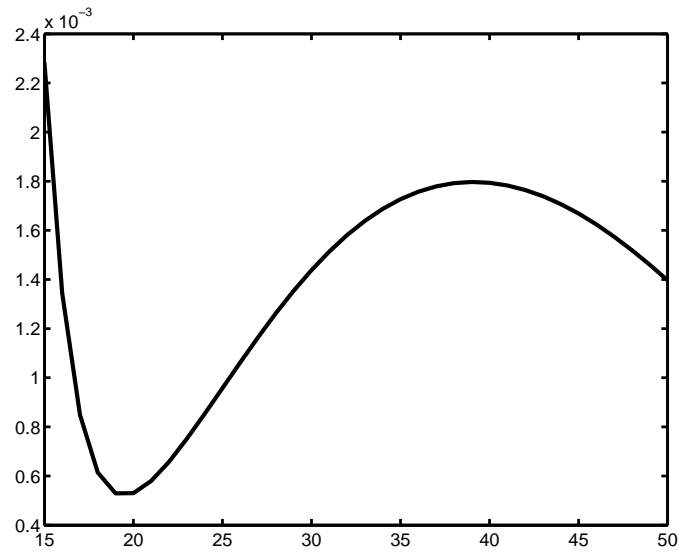
$$\hat{p}_i(y_{il}|x_i(t_l)) \propto \left[ \frac{1}{2^{|y_{il}-x_i(t_l)|}} + 10^{-6} \right] \quad (10.48)$$

where the small uniform term  $10^{-6}$  was added to increase numerical stability in the backward equation. We generated fifteen equally spaced counts from a Lotka-Volterra process and then corrupted them with the noise model (10.48).

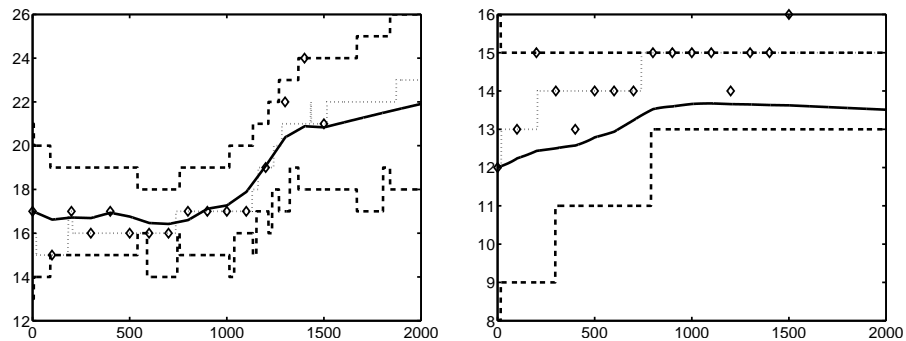
The results of the inference are shown in Figure 10.5. The figure shows also the posterior prediction for 1500 time steps after the last observation. The approximate posterior is shown to explain well the data and to overall give a good prediction also after the last observation.

Parameter inference is less accurate than in the weak noise case. Table 10.3 shows the mean and standard deviations in parameter estimates over 25 independent samples from the same Lotka-Volterra process. While the estimates are compatible with the true values, the standard deviations are much higher than the ones reported in Table 10.1, possibly reflecting the higher levels of noise. Computationally, the mean field approach still involves solving a large number of differential equations, and is hence much slower than the weak noise approach (a typical run takes approximately 10 minutes).

There are however some advantages in retaining the discrete nature of the process. Both the weak noise and mean field approximation predict with high confidence that the preys will reach extinction towards the end of the time window we consider. However, in Figure 10.6 we take a close look at the marginal distribution of preys at the final time ( $t = 3000$ ) obtained using the mean field approximation. This reveals a small secondary mode (accounting for approximately 0.5% of the total probability) at about 40 individuals. This accounts for the small probability that



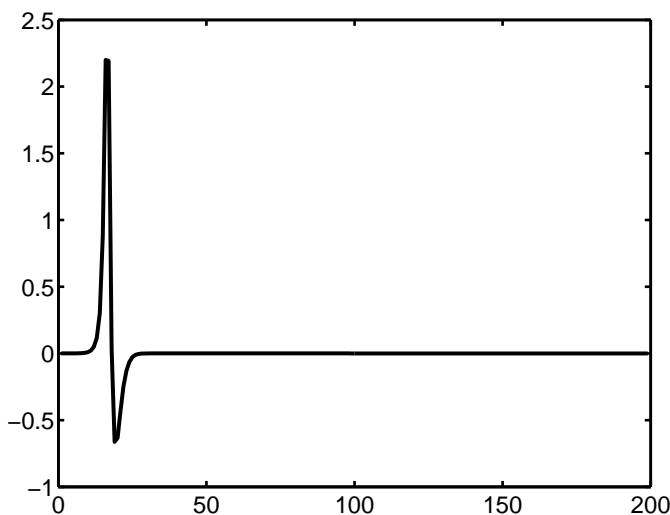
**Figure 10.6** Marginal prey distribution at the final time  $t = 3000$ , revealing a secondary mode at about 40 individuals. This accounts for the possibility that the preys might have recovered from near extinction. Horizontal axis denotes the number of preys, vertical axis the associated marginal at the final time.



**Figure 10.7** Inference results for the autoregulatory network process: approximate posterior distribution for proteins (left) and mRNAs (right). True parameter values are  $\alpha = 0.002$ ,  $\beta = 6 \times 10^{-5}$ ,  $\gamma = 5 \times 10^{-4}$ ,  $\delta = 7 \times 10^{-5}$ ,  $x_1(0) = 12$ ,  $x_2(0) = 17$ , and  $\sigma = 1$ . Symbols denote the noisy observations, thick lines show the posterior mean, and the 95% confidence region is surrounded by dashed lines. For comparison the original process is plotted as dotted line. Vertical axis indicates discrete numbers in the populations, horizontal axis is time.

**Table 10.4** Type II maximum likelihood estimates of reaction constants based on 15 observations. Mean values and standard deviations have been obtained by averaging over 25 autoregulatory processes using equal reaction constants and  $x_1(0) = 12$ ,  $x_2(0) = 17$  as initial conditions.

parameter	inference result	real value
$\alpha$	$(0.8 \pm 0.8) \times 10^{-2}$	$1 \times 10^{-2}$
$\beta$	$(2.8 \pm 1.9) \times 10^{-4}$	$6 \times 10^{-4}$
$\gamma$	$(4.6 \pm 1.8) \times 10^{-4}$	$5 \times 10^{-4}$
$\delta$	$(4.6 \pm 2.2) \times 10^{-5}$	$7 \times 10^{-5}$
$x_c$	$18.5 \pm 1.6$	20



**Figure 10.8** Variational free energy as a function of the threshold parameter  $x_c$ .

a “lucky” sequence of births for the preys and deaths for the predators helped the prey population to recover from near extinction. Clearly this can only happen when the intrinsic discreteness of the process is retained.

As a second example we again consider the case of an autoregulatory network described in the previous section. Results of the inference task are given in Figure 10.7 and Table 10.4. One can derive fixed point update equations for the kinetic parameters  $\alpha$ ,  $\beta$ ,  $\gamma$  and  $\delta$  of this model in a similar fashion to those obtained for the Lotka-Volterra process. However, a key difficulty is the determination of the critical threshold  $x_c$ . Clearly, if the proteins never cross the critical threshold, the parameter  $x_c$  is not identifiable. In general, its identifiability will be a problem when there are only very few observations in either state of the regulatory switch, *i.e.* when the proteins cross the threshold very early or very late during the observation window. In our experiments, we found this to be a significant problem in approximately 25% of cases (with these parameter settings) In general, the dependence of the

MF bound on the critical parameter  $x_c$  is given by

$$L_{x_c} = 2 \int_0^T dt \bar{g}_t h_t + \log \left( 1 - 0.99 \frac{1}{T} \int_0^T dt h_t \right) \int_0^T dt \bar{g}_t + \text{const} \quad (10.49)$$

where  $\bar{g}_t = \langle g_{1t}(x_1 + 1|x_1) \rangle_{q_{1t}}$  and  $h_t = \sum_{x_2 \geq x_c} q_{2t}(x_2)$ . A typical plot of this function obtained during the inference task (in the case when  $x_c$  is identifiable) can be seen in Figure 10.8(b). We can determine the minimum of (10.49) by searching over the possible (discrete) values of  $x_c$ .

## 10.8 Discussion

In this chapter we have discussed some approximate inference strategies for Markov Jump Processes. The approximations both stem from an exact variational formulation of the inference problem. By assuming weak noise and taking a perturbative expansion, we derive a diffusion approximation. This is computationally efficient and surprisingly accurate, but neglects the intrinsic discreteness of the process. On the other hand, by restricting the class of distributions to factorized distributions in the variational formulation, we obtain a mean field approximation to the process. This retains its discrete nature, but replaces correlations with average effects. We demonstrate both approaches on two simple but non-trivial problems, the Lotka-Volterra process and a genetic autoregulatory network, obtaining encouraging results in both cases.

The intrinsic discreteness of many biological and biochemical processes led to MJPs being extensively used to simulate biological systems via Gillespie's algorithm [Gillespie, 1977]. Given the increasing availability of technologies that can reveal this discreteness experimentally, it is to be expected that inference tools for MJPs will become increasingly useful.

While our approach relies on ideas from statistical physics and is essentially deterministic, other groups have been approaching the problem from a sampling point of view. Golightly and Wilkinson [2005] also use a diffusion approximation to MJPs, providing an MCMC sampling scheme for state and parameter estimation. An MCMC approach that retained the discrete nature of the MJPs was proposed by Boys et al. [2008]; however, this approach involved marginalizing over all the jumping times, which made it computationally very expensive.

There are several directions in which this work can be extended and improved. An important task is to reduce the computational expense of the mean field approximation. It would be of interest to develop alternative variational distributions (besides mean field) such as tractable linear processes. Another task is to extend both inference algorithms to using partial observations. This is quite important, as it is often impossible to observe some species of interest, *e.g.* transcription factor proteins [Barenco et al., 2006, see also Chapters 6 and 8]. An interesting extension of the work would be to consider hybrid models where discrete and continuous

random variables both play a role. A simple but fundamental example in systems biology is reaction-diffusion systems, where discrete particles diffuse and interact in continuous spatial dimensions. Another example of interest would be that of a biological switch, for example a transcription factor which can rapidly transit between on and off states and regulates genes whose expression level is essentially continuous.

