Published in *EPJ B* **65**, 369-377 (2008).

EPJ manuscript No.
(will be inserted by the editor)

Dynamical inference of hidden biological populations

Dmitri G. Luchinsky[1 2 3,a], Vadim N. Smelyanskiy[1], Marko Millonas[3], and Peter V. E. McClintock[2]

[1] NASA Ames Research Center, Mail Stop 269-2, Moffett Field, CA 94035, USA
[2] Department of Physics, Lancaster University, Lancaster LA1 4YB, UK
[3] Mission Critical Technologies Inc., 2041 Rosecrans Ave. Suite 225 El Segundo, CA 90245

Abstract. Population fluctuations in a predator-prey system are analyzed for the case where the number of prey could be determined, subject to measurement noise, but the number of predators was unknown. The problem of how to infer the unmeasured predator dynamics, as well as the model parameters, is addressed. Two solutions are suggested. In the first of these, measurement noise and the dynamical noise in the equation for predator population are neglected; the problem is reduced to a one-dimensional case, and a Bayesian dynamical inference algorithm is employed to reconstruct the model parameters. In the second solution a full-scale Markov Chain Monte Carlo simulation is used to infer both the unknown predator trajectory, and also the model parameters, using the one-dimensional solution as an initial guess.

1 Introduction

Population biologists use time-series data to infer the factors that regulate natural populations [1] and to determine when populations may be at risk of extinction. Often, however, only a few of the system's variables can be measured, while the rest of the variables remain unobservable, or *hidden* [2–5]. Furthermore, models describing population dynamics are multidimensional, nonlinear, stochastic and usually are not known exactly from first principles. A classical example is the intensively studied [3,6,7] cycling behavior of populations of small rodents observed in Kilpisjärvi, Finnish Lapland [8], 1952-1992 (see Fig. 1(a)) where the number of predators could not be measured, the dynamics was fully nonlinear and subject to seasonal and random perturbations, and the model was not exactly known beforehand. In these settings, perhaps the most fundamentally difficult unsolved problem is how, and to what extent, one can reconstruct missing information and deduce both the model and the full system trajectory from a given set of noise-corrupted, incomplete, trajectory measurements. Although specifically a problem of population biology (the cited database accumulates nearly 5000 individual datasets of similar structure, collected over more then 150 years), its solution is of importance across many disciplines where similar situations arise in diverse scientific contexts. Examples range from molecular motors [9] and epidemiology [2] to coupled matter-radiation systems [10].

It was shown earlier that the Markov Chain Monte Carlo (MCMC) and particle filter approaches to dynamical inference [4,5,11,12] can be very useful in this context, especially in applications to maps. However, the techniques [11,12] developed for one-dimensional maps are not immediately applicable to flows. The reason is that the log-likelihood functions in the two cases are of different form due to different transformations between the stochastic and dynamical variables [13,14].

[a] e-mail: d.luchinsky@lancaster.ac.uk

In this paper we extend earlier results in two important respects. First, we consider continuous systems and introduce for this case the correct likelihood function of observations, taking into account the Jacobian of transformation from stochastic to dynamical variables. Secondly, we consider an extreme case of missing data, when the entire dynamical trajectory of the predator population is missing and show how to infer both the missing dynamical trajectory and unknown model parameters. Note that the case when all the dynamical variables are measured is much simpler and can be solved [15] without application of the MCMC technique. As an example we consider a model inspired by analysis of the fluctuations in a population of small rodents in Finnish Lapland in order to address the problem of inference in a realistic setting. It will be shown that an extended MCMC method can be applied to reconstruct both the model parameters and the unobserved (hidden) predator dynamics.

In Sec. 2 the model of the rodent population and its transformation to standard form are discussed. The approximation of the model by a one-dimensional integro-differential equation, and inference of the model parameters in this simplified case, are considered in Sec. 4. The MCMC algorithm for the reconstruction of the parameters and the hidden predator trajectory for the full model are discussed in Sec. 5. Finally, conclusions are drawn in Sec. 6.

2 Model

2.1 Original model

Fluctuations of population density that are nearly periodic in time, but cannot be explained by seasonal variation, have fascinated ecologists for decades. But there is still no general agreement on the reasons for these cyclic variations in abundance. Extensive studies of population cycling have been carried out for the Finnish rodents mentioned above [3,7,6] and for lemming populations in Arctic tundra [16]. These studies include in particular a remarkable data series [17] for fluctuations in vole population from Kilpisjärvi, Finnish Lapland and a long-term field study of the cycling dynamics of collared lemming from northeastern Greenland. Many features of the observed fluctuations can be explained by the predator-prey model introduced in [3,6,16]. The parameters of these models and the exact functional forms of the various terms are known only approximately, while the predator dynamics is unobservable in most cases. To gain further insight into the ecological mechanisms underlying population fluctuation it becomes essential to infer unobserved predator dynamics and to obtain model parameters from experimental time-series data. To set the development of the inferential framework in a proper mathematical context let us summarize briefly the main results of the corresponding predator-prey model.

According to [3,6,16] the cycling dynamics of the vole population in Finnish Lapland is mainly controlled by the interplay between the so-called specialist predators (weasels) and generalist predators (foxes, owls, and others). The corresponding equations for the fluctuating densities of rodents N and their predators P can [3,7] be written as

$$\dot{N} = rN\left(1 - e_1 \sin(2\pi t) + \sigma_n \xi_n(t)\right) - (r/K)N^2 - \frac{GN^2}{N^2 + H^2} - \frac{CNP}{N + D}, \tag{1}$$

$$\dot{P} = sP\left(1 - e_2 \sin(2\pi t) + \sigma_p \xi_p(t)\right) - sQ\frac{P^2}{N}.$$

The effect of the *generalist predators*, whose population does not directly depend on the number of voles, is described by a functional response of Type III with maximum rate of mortality G and half-saturation prey density H. The *vole population* is characterized by prey-carrying capacity K and by the intrinsic rate of vole population growth r, which is disturbed by seasonal and stochastic forcing with amplitudes e_1 and σ_n respectively. The *specialist predator population* is described by the intrinsic rate of weasel population growth s, maximum consumption per predator C, the half-saturation constant D, by seasonal and stochastic forcing with amplitudes e_2 and σ_p, and by the carrying capacity proportional to prey density (Q is the constant of proportionality). The measurement error is modeled by a log-normal distribution, i.e. the measured

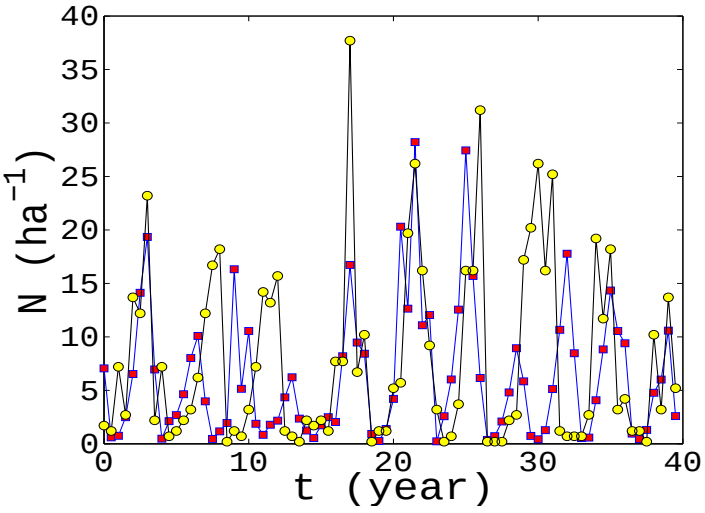

Fig. 1. The population dynamics of small rodents observed in Kilpisjärvi, Finnish Lapland, 1952-1992[3] is shown by yellow dots. The full black line is a guide to the eye. The blue line shows a simulation of the population dynamics using the model (1).

rodent density N' is related to the actual (unknown) value N via $N' = N\exp(\sigma_{obs}\eta(t))$ where $\eta(t)$ is a white Gaussian noise of unit intensity. The predator density cannot be measured and so the variable P is hidden. The values of the model parameters estimated in earlier research based on extensive field studies are summarized in Table 1. The problem is, however, that these field estimates are not related directly to the time-series data, and the range of parameter values is too broad. It is highly desirable from the point of view of understanding ecological mechanisms of population fluctuation, their prediction, and control, to develop methods for the estimation of model parameters directly from the time-series data. Below we suggest two such methods and analyze their performance using synthetic time-series.

To generate synthetic time-series data we use Eqs. (1) to obtain $N(t)$ and the measurement model to produce $ln(N'(t)) = ln(N(t)) + \sigma_{obs}\eta(t)$. The synthetic data points shown by the red squares in Fig. 1 are obtained by resampling $N'(t)$ with sampling frequency 0.5 year^{-1}. It can be seen from the figure that the model described above reproduces quite well the amplitude and frequency of the fluctuations of vole population.

Parameter	*units*	*range*
r	yr^{-1}	4 – 7
s	yr^{-1}	1 – 1.5
K	ha^{-1}	100 – 300
C	yr^{-1}weasel^{-1}	500 – 700
Q	voles weasel^{-1}	20 – 40
G	ha^{-1} yr^{-1}	70 – 125
H	ha^{-1}	11 – 16
D	ha^{-1}	5 – 6
e_1	-	0.5 – 1
e_2	-	0.5 – 1

Table 1. Values of the model parameters introduced for populations of a small rodent in Fennoscandia in [3].

The problem of dynamical inference can be stated as follows: use the 80 noise-corrupted data points to recover both the hidden dynamics of their predators $P(t)$ and the parameters of the model (1). This is a problem that could not be solved earlier because no general methods were available for its solution (but cf. discussions of the very different case when either dynamical or measurement noise is absent [18,7], or where only model parameters are estimated [4,5]).

2.2 Model transformation

To write Eqs. (1) in dimensionless variables, some known nominal values of the scaling coefficients K' and Q' are introduced. The dynamical equations for the scaled prey densities $n = N/K'$ and predator $p = \frac{Q'P}{K'}$ populations take the form

$$\begin{aligned} \dot{n} &= rn\left(1 - e_1 \sin(2\pi t + \psi_0) + \sigma_n \xi_n(t)\right) - \tilde{r} n^2 - \frac{gn^2}{n^2 + h^2} - \frac{anp}{n + d}, \\ \dot{p} &= sp\left(1 - e_2 \sin(2\pi t + \psi_0) + \sigma_p \xi_p(t)\right) - \tilde{s}\frac{p^2}{n}. \end{aligned} \tag{2}$$

The coefficients in this model are $g = G/K'$, $a = C/Q'$, $d = D/K'$, and $h = H/K'$. Note also the relationships between the original coefficients r and s and the scaled coefficients $\tilde{r} = rK/K'$ and $\tilde{s} = sQ/Q'$. They allow one to infer the carrying capacity K and the constant of proportionality of populations Q that are difficult to estimate using other methods. Note also that we have introduced into the seasonal variability terms an additional parameter ψ_0 corresponding to the unknown phase of the periodic seasonal driving.

The difficulties in applying methods of dynamical inference to Eqs. (2) stem from the following facts: (i) the noise terms are multiplicative; (ii) the predator trajectory is hidden; and (iii) the prey dynamics is measured together with some measurement noise. We overcome the first problem by making the changes of variable: $x_1(t) = \log(n)$ and $x_2(t) = \log(p)$. This set of equations can then be reduced to the form

$$\begin{aligned} \dot{x_1} &= r\left(1 - e_1 \sin(2\pi t + \psi_0)\right) - \tilde{r} e^{x_1} - \frac{g e^{x_1}}{e^{2x_1} + h^2} - \frac{a e^{x_2}}{e^{x_1} + d} + r\sigma_n \xi_n(t), \\ \dot{x_2} &= s\left(1 - e_2 \sin(2\pi t + \psi_0)\right) - \tilde{s} e^{x_2 - x_1} + s\sigma_p \xi_p(t), \\ y(t) &= x_1(t) + \sigma_{obs}\eta(t), \end{aligned} \tag{3}$$

where $y = ln(N'/K')$.

A solution of two other problems will be considered in a general Bayesian framework in Sections 3 and 5. But some very useful results can be obtained in a one-dimensional approximation (see Sec. 4), neglecting both measurement noise and noise in the second equation of (3). The latter approximation was also adopted earlier in [7] where estimation of the model parameters in (3) was performed by introducing numerically a so-called "atlas" function [19]. We show below that this approximation allows for an analytic solution of the problem at hand and provides a very useful guide to the actual values of the model parameters.

3 Bayesian inferential framework for hidden dynamical variables

The problem of dynamical inference is usually considered on a discrete time lattice ($t_k = h \cdot k, \quad k = 0, ..., K$), in which case the model (3) can be rewritten in a more general form as follows

$$\left.\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{x}_k + h\,\mathbf{f}(\mathbf{x}_k^*|\mathbf{c}) + \sqrt{h\hat{\mathbf{D}}}\mathbf{z}_k, \\ \mathbf{y}_k &= \hat{\mathbf{\Gamma}}\mathbf{x}_k + \sqrt{\hat{\mathbf{M}}}\boldsymbol{\eta}_k, \end{aligned}\right\} \tag{4}$$

Here $\mathbf{z}_k = \frac{1}{\sqrt{h}}\{\int_{t_k}^{t_{k+1}} \xi_n(t)dt, \int_{t_k}^{t_{k+1}} \xi_p(t)dt\}$, $\hat{\mathbf{D}}$ is a diagonal matrix with elements $\{(r\sigma_n)^2, (s\sigma_p)^2\}$ on the main diagonal, $\hat{\mathbf{M}}$ in our case is simply σ_{obs}^2, and $\mathbf{f}(\mathbf{x}_k^*|\mathbf{c})$ is a deterministic vector field

of the system (3) with $\mathbf{x}_k^* = \frac{\mathbf{x}_k + \mathbf{x}_{k+1}}{2}$. The measured M-dimensional time-series data $\mathcal{Y} = \{\mathbf{y}_k\}$ in our model is $\{\ln(N'_k/K')\}$, and the measurement matrix $\hat{\mathbf{\Gamma}}$ in this case is 1. The vector of unknown parameters can now combine (cf. [20,12]) a set of L-dimensional ($L > M$) hidden dynamical variables $\mathcal{X} = \{\mathbf{x}_k\}$ with model coefficients as follows

$$\mathcal{M} = \{\mathbf{c}, D_{ij}, \sigma_{obs}, \mathit{\Gamma}_{ij}, \mathcal{X}\}. \tag{5}$$

In the Bayesian approach to dynamical inference, the *posterior* probability $\rho_{\rm post}(\mathcal{M}|\mathcal{Y})$ of unknown parameters conditioned on observations $\mathcal{Y}$ is given by Bayes' theorem

$$\rho_{\rm post}(\mathcal{M}|\mathcal{Y}) = \frac{\ell(\mathcal{Y}|\mathcal{M})\,\rho_{\rm pr}(\mathcal{M})}{\int \ell(\mathcal{Y}|\mathcal{M})\,\rho_{\rm pr}(\mathcal{M}){\rm d}\mathcal{M}} \tag{6}$$

relating $\rho_{\rm post}(\mathcal{M}|\mathcal{Y})$ to the probability of observations (*likelihood*) $\ell(\mathcal{Y}|\mathcal{M})$ conditioned on the $\mathcal{M}$ and to the given *prior* probability $\rho_{\rm pr}(\mathcal{M})$ which is independent of observation. Accordingly the main problem of the Bayesian approach is to find the likelihood function $\ell(\mathcal{Y}|\mathcal{M})$ and to optimize the posterior distribution with respect to the parameters $\mathcal{M}$. We emphasize that the $\ell(\mathcal{Y}|\mathcal{M})$ is in fact the likelihood of the observed variables $\mathcal{Y}$ alone conditioned on the set of unknown parameters $\mathcal{M}$ that includes both model parameters and trajectories of hidden variables (see (5)).

To find an analytic form of $\ell(\mathcal{Y}|\mathcal{M})$, we notice, following earlier work (see e.g. [14,21,15] and references therein), that the likelihood can be factorized

$$\ell(\mathcal{Y}|\mathcal{M}) = \rho(\mathcal{Y}|\mathcal{X})\rho(\mathcal{X}|\mathcal{M}'), \tag{7}$$

where $\mathcal{M}'$ is the reduced set of parameters that does not include hidden variables (cf. (5))

$$\mathcal{M}' = \{\mathbf{c}, D_{ij}, \sigma_{obs}, \mathit{\Gamma}_{ij}\}.$$

The conditional probabilities $\rho(\mathcal{Y}|\mathcal{X})$ and $\rho(\mathcal{X}|\mathcal{M}')$ can be found using known distributions for the independent sources of white Gaussian noise $\mathbf{z}_n$ and $\boldsymbol{\eta}_k$ in (4)

$$\mathcal{P}[\mathbf{z}_n] = \prod_{n=0}^{N-1} \frac{{\rm d}\mathbf{z}_n \exp\left(-\mathbf{z}_n^{\rm T}\,\frac{\hat{\mathbf{D}}^{-1}}{2h}\,\mathbf{z}_n\right)}{\sqrt{(2\pi h)^L|\hat{\mathbf{D}}|}}, \qquad \mathcal{P}[\boldsymbol{\eta}_n] = \prod_{n=0}^{N-1} \frac{{\rm d}\boldsymbol{\eta}_n \exp\left(-\boldsymbol{\eta}_n^{\rm T}\,\frac{\hat{\mathbf{M}}^{-1}}{2h}\,\boldsymbol{\eta}_n\right)}{\sqrt{(2\pi h)^M|\hat{\mathbf{M}}|}}. \tag{8}$$

Using (4) to transform from stochastic variables $\mathbf{z}_n$ and $\boldsymbol{\eta}_k$ to dynamical variables $\mathbf{x}_k$ and $\mathbf{y}_k$ we obtain

$$\begin{aligned}\rho(\mathcal{X}|\mathcal{M}') &= \rho_{\rm st}(\mathbf{x}_0)\,J(\{\mathbf{x}_n\}) \\ &\times \prod_{n=0}^{N-1} \frac{1}{\sqrt{(2\pi h)^L|\hat{\mathbf{D}}|}} \exp\left(-\frac{h}{2}\left[\dot{\mathbf{x}}_n - \mathbf{f}(\mathbf{x}_n^*;\mathbf{c})\right]^{\rm T}\hat{\mathbf{D}}^{-1}\left[\dot{\mathbf{x}}_n - \mathbf{f}(\mathbf{x}_n^*;\mathbf{c})\right]\right), \end{aligned} \tag{9}$$

$$\rho(\mathcal{Y}|\mathcal{X}) = \prod_{n=0}^{N-1} \frac{1}{\sqrt{(2\pi h)^M|\hat{\mathbf{M}}|}} \exp\left(-\frac{1}{2}[\mathbf{y}_n - \hat{\mathbf{\Gamma}}\mathbf{x}_n]^{\rm T}\,\hat{\mathbf{M}}^{-1}\,[\mathbf{y}_n - \hat{\mathbf{\Gamma}}\mathbf{x}_n]\right), \tag{10}$$

where $\rho_{\rm st}(\mathbf{x})$ signifies the stationary distribution of $\mathbf{x}(t)$, and the Jacobian of the transformation is given by

$$J(\{\mathbf{x}_n\}) = \left|\left\{\frac{\partial z_{ln}}{\partial x_{l'n'}}\right\}\right| \simeq \prod_{n=1}^{N}\prod_{l=1}^{L}\left[1 - \frac{h}{2}\,\frac{\partial f_l(\mathbf{x}_n^*;\mathbf{c})}{\partial x_{ln}}\right] \simeq \exp\left[-\frac{h}{2}\sum_{n=1}^{N}\nabla.(\mathbf{f}(\mathbf{x}_n)|\mathbf{c})\right], \tag{11}$$

Using (7) – (11) (see also [14,21,15]) the minus logarithm of the likelihood to observe $\mathcal{Y}$ can be factorized and written in the form

$$-\frac{2}{K}\,\ln\ell(\mathcal{Y}|\mathcal{M}) = \ln|\hat{\mathbf{D}}| + \ln|\hat{\mathbf{M}}| + \frac{1}{K}\sum_{k=0}^{K}[\mathbf{y}_k - \hat{\mathbf{\Gamma}}\mathbf{x}_k]^{\rm T}\,\hat{\mathbf{M}}^{-1}\,[\mathbf{y}_k - \hat{\mathbf{\Gamma}}\mathbf{x}_k]$$

$$+\frac{h}{K}\sum_{k=0}^{K-1}\left\{\nabla.(\mathbf{f}(\mathbf{x}_k)|\mathbf{c})+\left[\dot{\mathbf{x}}_k-\mathbf{f}(\mathbf{x}_k^*|\mathbf{c})\right]^{\mathrm{T}}\hat{\mathbf{D}}^{-1}\left[\dot{\mathbf{x}}_k-\mathbf{f}(\mathbf{x}_k^*|\mathbf{c})\right]\right\}+2L\ln(2\pi h). \quad (12)$$

It is important to note that this likelihood function is asymptotically exact in the limit $h \to 0$ and $K \to \infty$ while $T = Kh$ remains constant. The term $\nabla.(\mathbf{f}(\mathbf{x}_k)|\mathbf{c})$ in the (12) term emerges in the path integral presentation of ℓ as the Jacobian of the transformation from noise variables to dynamical variables [22,23] and provides optimal compensation for the noise-induced errors of inference [14,21,15].

Now we are ready to discuss optimization of the posterior distribution using e.g. the MCMC method. But first we consider a one-dimensional approximation of (3).

4 One-dimensional approximation

4.1 1D model

A very useful guide to the actual values of the model parameters in (3) can be obtained if we assume (cf [7]) that the noise terms in the measurement equation and predator dynamics are negligible. We note that this approximation corresponds, in fact, to the original model suggested by Hanski and Turchin [3]. In their paper they comment that inference of the model parameters is prevented by the fact that the predator population was not observed. Later [7] they performed estimation of the model parameters in (3) by excluding predator dynamics and introducing numerically a so-called "atlas" function (that relates prey population at the time step k to its values at the time steps $k-1$ and $k-2$ see [19]).

We notice, however, that in this case the predator population is uniquely determined by the population of prey for a given set of the parameters in the second equation. This fact allows us to reduce the inference problem to one dimension and to infer both predator dynamics and the model parameters. Furthermore, many of the model parameters can be estimated in this case analytically using Bayesian method described in the previous section.

Indeed, dividing the second equation in (2)

$$\dot{p} = sp\left(1 - e_2\sin(2\pi t + \psi_0)\right) - \tilde{s}\frac{p^2}{n}, \quad (13)$$

by p^2, assuming for simplicity $\psi_0 = 0$ in this equation, and introducing variable $z = 1/p$ we have

$$\dot{z} = \left(s_1 + s_2\sin(2\pi t)\right)z + s_3\frac{1}{n},$$

where $s_1 = s$, $s_2 = -se_2$, $s_3 = -\tilde{s}$. This equation can be integrated to obtain the one-dimensional approximation of the problem (3) in the form

$$\dot{x_1} = r\left(1 - e_1\sin(2\pi t + \psi_0)\right) - \tilde{r}e^{x_1} - \frac{ge^{x_1}}{e^{2x_1}+h^2} - \frac{az^{-1}}{e^x+d} + r\sigma_n\xi_n(t), \quad (14)$$

$$z = e^{-s_1 t - \frac{s_2}{2\pi}\cos(2\pi t)}\left(c_0 + s_3\int_{t_0}^{t}\frac{d\tau}{n(\tau)}e^{s_1\tau + \frac{s_2}{2\pi}\cos(2\pi\tau)}\right). \quad (15)$$

4.2 Dynamical inference algorithm

To infer parameters of the model (21) we can apply earlier results [14,13] by introducing the following parametrization of the vector field

$$\mathbf{f}(\mathbf{x}|\mathbf{c}) = \hat{\mathbf{F}}(\mathbf{x})\,\mathbf{c}, \quad (16)$$

where matrixes $\hat{\mathbf{F}}$ have the form

$$\hat{\mathbf{F}} = \left[\begin{pmatrix} \phi_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \phi_1 \end{pmatrix} \dots \begin{pmatrix} \phi_F & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \phi_F \end{pmatrix} \right]. \tag{17}$$

Choosing a Gaussian prior PDF for $\mathbf{c}$ and a uniform prior PDF for $\hat{\mathbf{D}}$, we obtain [14] the following equations for updating the dynamical model parameters

$$\langle \mathbf{c} \rangle = \left(h \sum_{k=0}^{K-1} \hat{\mathbf{F}}_k^{\mathrm{T}} \hat{\mathbf{D}}^{-1} \hat{\mathbf{F}}_k \right)^{-1} \left(h \sum_{k=0}^{K-1} \left[\hat{\mathbf{F}}_k^{\mathrm{T}} \, \hat{\mathbf{D}}^{-1} \, \dot{\mathbf{x}}_k - \frac{\mathbf{v}(\mathbf{x}_k)}{2} \right] \right), \tag{18}$$

$$\langle \hat{\mathbf{D}} \rangle = \frac{h}{K} \sum_{k=0}^{K-1} \left[\dot{\mathbf{x}}_k - \hat{\mathbf{F}}_k \, \mathbf{c} \right] \left[\dot{\mathbf{x}}_k - \hat{\mathbf{F}}_k \, \mathbf{c} \right]^{\mathrm{T}}, \tag{19}$$

where $\hat{\mathbf{F}}_k \equiv \hat{\mathbf{F}}(\mathbf{x}_k)$, and the components of the vector $\mathbf{v}(\mathbf{x})$ related to the Jacobian of transformation in (9) are

$$v_m(\mathbf{x}) = \sum_{l=1}^{L} \frac{\partial F_{lm}(\mathbf{x})}{\partial x_l}, \quad m = 1, \dots, F. \tag{20}$$

Eqs. (18), (19) provide the solution for the inference problem in the case when the minus log-likelihood (12) is a quadratic function of the model parameters $\mathbf{c}$. Note that these equations have to be applied iteratively: first we calculate $\hat{\mathbf{D}}$ for given initial values of $\mathbf{c}$; next we calculate $\mathbf{c}$ using the value of $\hat{\mathbf{D}}$ obtained at the previous step; these two steps are then repeated until convergence is finally achieved. Usually 2-3 iterations are sufficient for convergence.

For our specific model eq. (21) can now be written as follows

$$\dot{x_1} = \{\phi_n\} \left(\mathbf{c}^l \right)^T + r\sigma_n \xi_n(t), \tag{21}$$

with the base functions

$$\{\phi_n\} = \left\{ 1, \sin(2\pi t), \cos(2\pi t), e^x, \frac{e^x}{e^{2x} + h^2}, \frac{z^{-1}}{e^x + d} \right\}. \tag{22}$$

Here $\phi_5 = \phi_5(h)$ and $\phi_6 = \phi_5(d, s_1, s_2, s_3)$ are nonlinear functions of some of the model parameters

$$\{c_n\} = \{\mathbf{c}^l, \mathbf{c}^{nl}\} = \{\{r, -re_1 \cos(\psi_0), -re_1 \sin(\psi_0), -\frac{rK'}{K}, -g, -a\}, \{h, d, s_1, s_2, s_3\}\}. \tag{23}$$

Therefore the algorithm (18), (19) has to be extended to infer nonlinear parameters $\mathbf{c}^{nl} = \{h, d, s_1, s_2, s_3\}$.

4.3 Conjugate gradient search in nonlinear parameter space

A number of algorithms can be employed to infer nonlinear model parameters, including the MCMC considered in Sec. 5. In general, it is useful to compare the performance of various algorithms, since none of them can guarantee a convergence. Here we consider the conjugate gradient search in the space of nonlinear parameters. To do so we write the cost function as an abbreviated minus log-likelihood function (12) that includes only the dependence on the nonlinear parameters $\mathbf{c}^{nl}$ given in (23) as follows

$$g(\mathbf{c}^l, \mathbf{c}^{nl}) = \mathbf{c}^{l\,T} \mathbf{b}_s \left(\hat{\mathbf{D}}, \mathbf{c}^{nl} \right) + \frac{1}{2} \mathbf{c}^{l\,T} \hat{\mathbf{H}}_s \left(\hat{\mathbf{D}}, \mathbf{c}^{nl} \right) \mathbf{c}^l, \tag{24}$$

with the following definitions of $\mathbf{b}_s\left(\hat{\mathbf{D}}, \mathbf{c}^{nl}\right)$ and $\hat{\mathbf{H}}_s\left(\hat{\mathbf{D}}, \mathbf{c}^{nl}\right)$ in the one-dimensional case (21)

$$\mathbf{b}_s\left(\hat{\mathbf{D}}, \mathbf{c}^{nl}\right) = \frac{h}{2}\sum_{k=0}^{K-1}\left[\overrightarrow{\phi'}_k - \frac{2}{D}\dot{x}_k\overrightarrow{\phi}_k\right], \qquad \hat{\mathbf{H}}_s\left(\hat{\mathbf{D}}, \mathbf{c}^{nl}\right) = h\sum_{k=0}^{K-1}\overrightarrow{\phi}_k\frac{1}{D}\overrightarrow{\phi}^T_k, \tag{25}$$

where the base functions ϕ_k are given in (22). The resultant dependence of the cost function $g(\mathbf{c}^l, \mathbf{c}^{nl})$ on the values of the nonlinear parameters (while the values of $\mathbf{c}^l$ are fixed) has a well-pronounced nearly quadratic minimum. Therefore, we can apply a conjugate gradient method to optimize the cost function with respect to the nonlinear parameters. Below we consider as an example convergence of the $g(\mathbf{s})$ in the space of the predator parameters $\mathbf{s} = \{s_1, s_2, s_3\}$ keeping all other parameters fixed.

To find the gradient of the cost function (24) we note that there is only one function that depends on the predator parameters. It is $\phi_6(t_k) = z^{-1}(\mathbf{s}, t_k)/(e^{x_k} + d)$, which depends on the predator parameters via (15). Therefore, we can write

$$\nabla g(\mathbf{s}) = hc_6\sum_{k=0}^{K-1}\frac{1}{(e^{x_k}+d)}\left[-\frac{e^{x_k}}{2(e^{x_k}+d)} + \frac{1}{D}\left(-\dot{x}_k + \sum_{m=1}^{6}c_m\phi_{k,m}\right)\right]\nabla p_k(\mathbf{s}). \tag{26}$$

Taking into account the expression (15) we obtain

$$\frac{\partial g(\mathbf{s})}{\partial s_1} = hc_6\sum_{k=0}^{K-1}F_1(t_k)p_k(\mathbf{s})\left[t - s_3F_2(t_k)\int_{t_0}^{t}\frac{\tau d\tau}{x(\tau)}e^{s_1\tau + \frac{s_2}{2\pi}\cos(2\pi\tau)}\right], \tag{27}$$

$$\frac{\partial g(\mathbf{s})}{\partial s_2} = hc_6\sum_{k=0}^{K-1}F_1(t_k)p_k(\mathbf{s})\left[\frac{\cos(2\pi\tau)}{2\pi} - s_3F_2(t_k)\int_{t_0}^{t}\frac{\cos(2\pi\tau)d\tau}{2\pi x(\tau)}e^{s_1\tau + \frac{s_2}{2\pi}\cos(2\pi\tau)}\right], \tag{28}$$

$$\frac{\partial g(\mathbf{s})}{\partial s_3} = hc_6\sum_{k=0}^{K-1}F_1(t_k)p_k(\mathbf{s})\left[-F_2(t_k)\int_{t_0}^{t}\frac{d\tau}{x(\tau)}e^{s_1\tau + \frac{s_2}{2\pi}\cos(2\pi\tau)}\right]. \tag{29}$$

Here

$$F_1(t_k) = \frac{1}{(e^{x_k}+d)}\left[-\frac{e^{x_k}}{2(e^{x_k}+d)} + \frac{1}{D}\left(-\dot{x}_k + \sum_{m=1}^{6}c_m\phi_{k,m}\right)\right]$$

and

$$F_2(t_k) = \frac{1}{c_0 + s_3\int_{t_0}^{t_k}\frac{d\tau}{x(\tau)}e^{s_1\tau + \frac{s_2}{2\pi}\cos(2\pi\tau)}}.$$

Finally, to infer the predator parameters, we use the following conjugate gradient algorithm:

- Initialize values of the parameters $\mathbf{s}_0 = \{s_1^{(0)}, s_2^{(0)}, s_3^{(0)}\}$;
- Find the initial direction of the search: $\mathbf{d}_0 = -\nabla_s g(\mathbf{s}_0)$;
- Update values of the initial guess for the coefficients according to the rule $\mathbf{s}_1 = \mathbf{s}_0 + \alpha\mathbf{d}_0$;
- Find the new direction of the search (conjugate to the previous direction) according to the rule: $\mathbf{d}_1 = -\nabla_s g(\mathbf{s}_1) + w_1\mathbf{d}_0$, where $w_1 = ||\nabla_s g(\mathbf{s}_1)||^2/||\nabla_s g(\mathbf{s}_0)||^2$;
- Go back to the previous step, and iterate until convergence is achieved.

The step α in the conjugate direction is found by a line-optimization procedure.

4.4 Inference results

Examples of the one-dimensional inference of linear coefficients of the model (21) are shown in Fig. 2 and summarized in Table 2. It can be seen from both the figure and the table that the relative error of the inference of the linear parameters is less than 2% except in the case of e_1, for

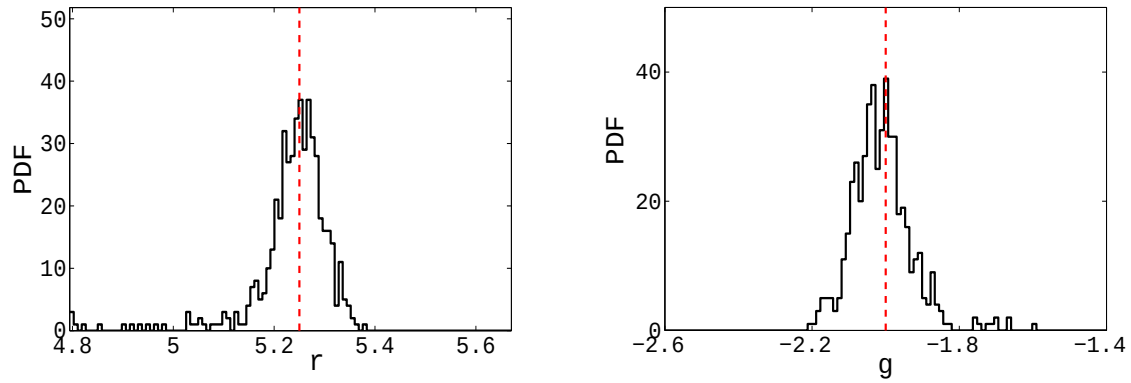

Fig. 2. Example of the inference of the linear parameters r and g in model (21). The distributions are obtained using 1000 trajectories with 128000 points each and a sample interval of 0.001. The coefficients $\{\mathbf{s}, h, d, e_2\}$ are assumed to be known.

which it is 3.75%. We note, however, that such a small error was achieved for densely sampled data, when the nonlinear parameters were set to their correct values. Even for more realistic settings, however, the method provides a useful initial guess for the linear model parameters and for the hidden predator trajectory.

Furthermore, using the conjugate gradient technique one can also estimate values of the nonlinear parameters of the problem. The corresponding results are illustrated in Fig. 3 and summarized in Table 3. Fig. 3 shows the convergence of nonlinear parameters in the projections of the cost functions $g(\mathbf{s})$ on the hyperplanes defined by conditions $s_3 = \text{const}$ in Fig. 3(a) and $s_2 = \text{const}$ in Fig. 3 (b).

We, therefore, conclude that the one-dimensional approximation (21) of the predator-prey dynamics (3) provides a useful guide to the values of the model parameters and for the hidden predator trajectory (15). The values obtained in this analysis can be used as an initial guess for more general and numerically extensive searches such as the MCMC technique described in the following section.

5 The Markov Chain Monte Carlo (MCMC) technique

We now consider the problem of inferring both the parameters of the nonlinear stochastic model, and the latent dynamical variables, simultaneously. The MCMC approach [24] to such inference was adopted recently in [11,12] for one-dimensional maps. We extend these results to flows by including the correct prefactor term into the likelihood function (12) as discussed in the introduction (see also [13,21,14]), and by considering an extreme case of missing data when the entire predator trajectory is missing.

parameter	r	s	a	g	h	d	e_1	e_2	$\cos(\psi_0)$	D
actual value	5.25	1.25	-15	-2	0.4	0.04	0.38	0.8	1.0	0.0625
inferred value	5.2	1.25	-14.9	-1.99	0.4	0.04	0.38	0.8	0.99	0.063
$\sqrt{\sigma^2}$	0.08	-	0.28	0.035	-	-	0.015	-	0.011	0.0004

Table 2. Values of the parameters in model (3) that were used to obtain the sample of synthetic data shown in Fig. 1. The third-from-top row shows values of the inferred parameters using 1000 prey trajectories with 128000 points each and a sample interval of 0.001. The bottom row shows values of the corresponding standard deviations. Missing standard deviations indicate that coefficients ($\{s, h, d, e_2\}$) are assumed to be known.

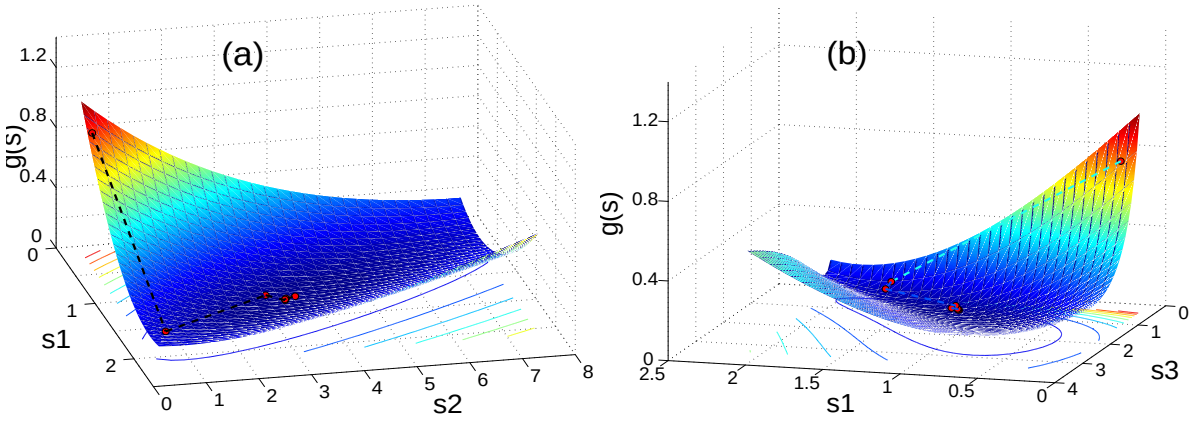

Fig. 3. (a) Hyperplane of the cost function defined by the condition $s_3 = s_3^*$, where s_3^* is the optimal value of the parameter s_3. Red dots show evolution of the solution of the optimization problem starting from the initial values of the parameters $s_1^{in} = 0.08$ and $s_2^{in} = 0.1$. (b) Hyperplane of the cost function are defined by the condition $s_2 = s_2^*$, where s_2^* is the optimal value of the parameter s_2. Red dots show evolution of the solution of the optimization problem starting from the initial values of the parameters $s_1^{in} = 0.08$ and $s_3^{in} = 0.3$.

parameter	s_1	s_2	s_3	*cost*
initial value	0.1	2.2	2.5	-1355
2^{nd} iteration	1.161	2.384	2.152	-1467
4^{th} iteration	1.225	2.744	1.262	-1479
8^{th} iteration	1.1534	3.344	1.308	-1482.5
actual value	1.25	4	1.25	-1482.9

Table 3. Values of the nonlinear parameters in model (21) inferred using conjugate gradient method in 8 steps. The actual values are shown in the bottom line.

For the sake of simplicity we assume the noise intensities to be fixed and introduce an abbreviated vector of the unknown parameters $\tilde{\mathcal{M}} = \{\mathbf{c}, \{\mathbf{x}_k\}\}$. The desired probability of the model parameters is then $p(\tilde{\mathcal{M}}|\mathcal{Y}) \propto \rho_{post}(\mathcal{M}|\mathcal{Y}) = \text{const} \times \exp(S)$, where $S = -\ell(\mathcal{Y}|\mathcal{M})$ is

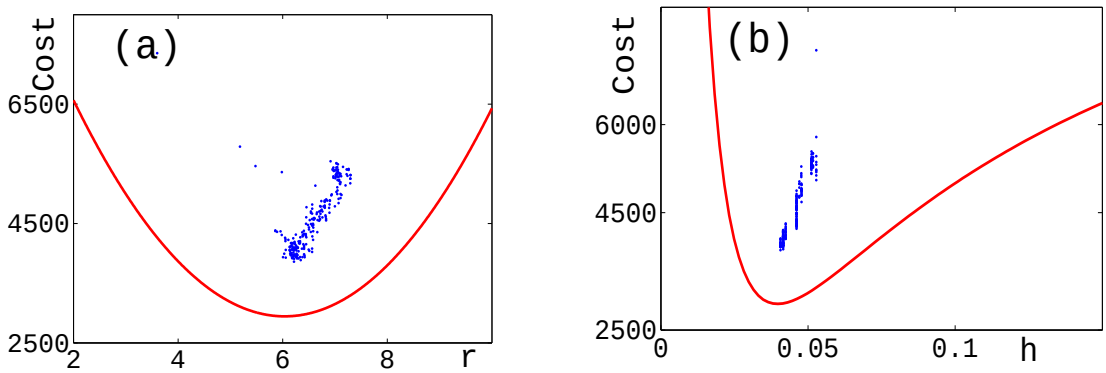

Fig. 4. Results of the MCMC calculations. (a) Iterations of the coefficient r are shown by blue dots. The local curvature of the cost function is shown by the red solid line. (b) Iterations of the coefficient h are shown by blue dots. The local curvature of the cost function is shown by the red solid line.

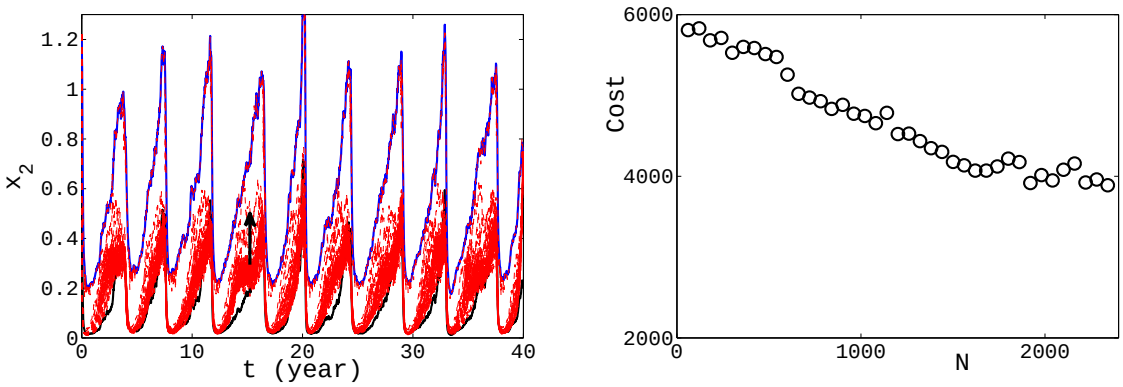

Fig. 5. Results of the MCMC calculations. (left) Convergence of the unknown predator trajectories from an initial guess (solid black line at the bottom of the figure) to the actual trajectory (solid blue line at the top of the figure) is shown by dashed red lines. The arrow indicates the direction of convergence as a function of number of iterations. (right) The changes of the cost function at each iteration step corresponding to (left) and Fig. 4 are shown by the black circles.

given in (12). We analyze the convergence only in the space of parameters $\mathbf{c}$ and dynamical trajectories. With this function the MCMC algorithm can be briefly summarized as follows

(1) Take an initial guess for $\tilde{\mathcal{M}}^{(0)} = \{\mathbf{c}^{(0)}, \{\mathbf{x}_k^{(0)}\}\}$.
(2) Sample a trajectory from $p(x_k|x_{k-1}, x_{k+1}, \tilde{\mathcal{M}}, \hat{\mathbf{D}}, \sigma_{obs}, y_t)$ for $k = 0, ..., K$ using Gibbs sampler with Metropolis-Hastings (M-H) steps [25];
(3) Sample model parameters from $p(\tilde{\mathcal{M}}|\{x_t\}, \hat{\mathbf{D}}, \sigma_{obs}, \{y_t\})$ using M-H algorithm or possibly directly using eq. (18);
(4) Repeat steps (2), (3) until convergence is achieved.

Note that this algorithm takes into account that the coordinate $\mathbf{x}_k$ enters only in two terms of the sum (12). This fact considerably simplifies the MCMC calculations of hidden dynamical variables (cf [11,12]). The initial values of the model parameters are drawn using uniform distributions from intervals (see Table 4) that overlap with and extend the field-study-based estimations of ecological parameters shown in Table 1. Once the parameters are known, the initial guess for the trajectory is found using (15). The initial guess for the predator trajectory is shown by the black solid line at the bottom of Fig. 5(a). A Gibbs sampler is used to update the trajectory. At each step k the new coordinate x_k is drawn sequentially from the distribution $p(x_k|x_{k-1}, x_{k+1}, \tilde{\mathcal{M}}, \hat{\mathbf{D}}, \sigma_{obs}, y_t)$ using M-H algorithm. For a given set of model parameters the trajectory is updated a number of times to achieve local convergence. Once the trajectory is updated, new model parameters are drawn from the posterior distribution using M-H algorithm. These steps are repeated until global convergence is achieved.

To improve the convergence we:

parameter	r	s	a	g	h	d	e_1	e_2
actual value	6.0	1.2	-15	-1	1.0	0.04	0.07	0.21
initial range	1:10	0.3:2	-1:-25	-0.01:-2	0.2:2	0.01:1	-0.5:1.3	-0.5:1.3
inferred value	6.5	1.17	-16.4	-0.96	0.82	0.05	0.22	0.38
$\sqrt{\sigma^2}$	0.5	0.09	1.3	0.4	0.5	0.004	0.1	0.015

Table 4. Convergence of some of the model parameters in the MCMC calculations. Notice that, unlike in Table 2, all the model parameters except noise intensities are unknown. The initial values of the model parameters are drawn according to a uniform distribution from the intervals shown in the third raw.

(i) Keep parameters within the box of values specified by the initial range for each parameter shown in Table 4;
(ii) Scale the noise of the MCMC simulations by a factor proportional to the curvature of the cost function for each parameter;
(iii) Increase the number of trajectory iterations up to 20 at each MCMC step.

Examples of the convergence of the model parameters are shown in Fig. 4. The convergence of trajectories is illustrated in the Fig. 5 together with the dynamics of the cost function. The estimates of the parameter values obtained in these simulations are shown in Table 4.

6 Conclusions

In conclusion, we have considered the problem of dynamical inference of latent state variables and parameters of nonlinear stochastic dynamical models, and done so for the extreme case of missing data, when an entire trajectory is missing. As an example of how to solve a long-standing ecological problem, we inferred an unobservable predator trajectory, and parameter values, for a predator-prey model by analysis of measurements of the prey dynamics that were (as is typical) corrupted by noise. We proposed a solution of this problem based on the MCMC method with a Gibbs sampler and M-H steps to draw parameters from non-Gaussian distributions. This solution extends earlier results [11,12] obtained for one-dimensional maps to multidimensional flows for the case when only partial information about the system dynamics is available. To obtain an initial guess for the model parameters and unobservable predator trajectory, we introduced a one-dimensional approximation of the predator-prey model, neglecting the noise in the predator dynamics. It was shown that the MCMC method converges both in the state space and the parameter space. The work is still in progress and can be further improved in a number of ways. In particular, information about the gradient of the cost function can be included in the MCMC simulations. As an immediate extension of this work, we plan to apply our results to an analysis of the population dynamics of small rodents in Finnish Lapland [3, 7,6]. Details of this analysis will be published elsewhere, but the preliminary results are very promising and indicate that the methods developed in the present research can successfully infer both hidden dynamics of the predator populations and the unknown model parameters from the time-series data of prey dynamics observed [3] in Kilpisjärvi, 1952-1992.

We emphasize that the results obtained are of importance across many disciplines. As discussed in the Introduction, the method will also be applicable wherever similar situations arise, including scientific contexts as diverse as molecular motors [9] and aerospace applications [26].

References

1. N.C. Stenseth, Science **269**, 1061 (1995)
2. J.L. Aron, I.B. Schwartz, J. Theor. Biol. **110**, 665 (1984)
3. P. Turchin, I. Hanski, American Naturalist **149**, 842 (1997)
4. E.L. Ionides, C. Breto, A.A. King, Proceedings of the National Academy of Sciences **103**(49), 18438 (2006)
5. S. Cauchemez, N.M. Ferguson, Journal of The Royal Society Interface **5**(25), 885 (2008), 10.1098/rsif.2007.1292
6. I. Hanski, H. Henttonen, E. Korpimaäki, L. Oksanen, P. Turchin, Ecology **82**(6), 1505 (2001)
7. P. Turchin, S.P. Ellner, Ecology **81**(11), 3099 (2000)
8. NERC Centre for Population Biology, Imperial College (1999), *The Global Population Dynamics Database*, http://www.sw.ic.ac.uk/cpb/cpb/gpdd.htm
9. K. Visscher, M.J. Schnitzer, S.M. Block, Nature **400**, 184 (1999)
10. J. Christensen-Dalsgaard, Rev. Mod. Phys. **74**(4), 1073 (2002)
11. R. Meyer, N. Christensen, Physical Review E **62**, 3535 (2000)
12. C. Calder, M. Lavine, P. Müller, J.S. Clark, Ecology **84**(6), 1395 (2003)
13. V.N. Smelyanskiy, D.G. Luchinsky, A. Stefanovska, P.V.E. McClintock, Physical Review Letters **94**(9), 098101 (4) (2005)

14. V.N. Smelyanskiy, D.G. Luchinsky, D.A. Timucin, A. Bandrivskyy, Physical Review E **72**(2), 026202 (12) (2005)
15. D.G. Luchinsky, V.N. Smelyanskiy, M. Millonas, P.V.E. McClintock, in *Noise in Complex Systems and Stochastic Dynamics III*, edited by L.B. Kish, K. Lindenberg, Z. Gingl (2005), Vol. 5845 of *Proc. of SPIE)*, conf. on Noise in Complex Systems and Stochastic Dynamics III MAY 24-26, 2005 Austin, TX. SPIE, Bellingham, 2005, pp. 173–181.
16. O. Gilg, I. Hanski, B. Sittler, Science **302**, 866 (2003)
17. K. Laine, H. Henttonen, Oikos **40**(3), 407 (1983)
18. H.U. Voss, J. Timmer, J. Kurths, Int. J. Bifurc. and Chaos **14**, 1905 (2004)
19. C.W. Tidd, L.F. Olsen, W.M. Schaffer, Proc R Soc Lond B **254**(1341), 257 (1993)
20. R. Meyer, N. Christensen, Phys. Rev. E **65**, 016206 (2001)
21. D.G. Luchinsky, V.N. Smelyanskiy, J. Smith, in *Unsolved Problems of Noise and Fluctuations*, edited by L. Reggiani, C. Pennetta, V. Akimov, E. Alfinito, M. Rosini (2005), Vol. 800 of *AIP Conference Proceedings*, pp. 539–545, 4th International Conference on Unsolved Problems of Noise and Fluctuations in Physics, Biology and High Technology June 06-10, 2005 Gallipoli, ITALY
22. R. Graham, *Tracts in Modern Physics* (Springer-Verlag, New York, 1973), Vol. 66, chap. Quantum Statistics in Optics and Solid-State Physics
23. E. Gozzi, Phys. rev. D **28**(8), 1922 (1983)
24. W.R. Gilks, S. Richardson, D.J. Spiegelhalter, *Markov Chain Monte Carlo in Practice* (Chapman and Hall, New York, 1996)
25. W.K. Hastings, Biometrika **57**(1), 97 (1970)
26. V.V. Osipov, D.G. Luchinsky, V.N. Smelyanskiy, D.A. Timucin, **AIAA 2007-5823** (2007), 43rd AIAA/ASME/SAE/ASEE Joint Propulsion Conference & Exhibit, 8 - 11 July 2007, Cincinnati, OH