Recovering "Lost" Information in the Presence of Noise: Application to Population Ecology

V.N. Smelyanskiy[1], D.G. Luchinsky[2,3], M. Millonas[2], and P.V.E. McClintock[3]

[1] *NASA Ames Research Center, MS 269-2, Moffett Field, CA 94035, USA*
[2] *Mission Critical Technologies Inc., 2041 Rosecrans Ave. Suite 225 El Segundo, CA 90245 and*
[3] *Department of Physics, Lancaster University, Lancaster LA1 4YB, UK*
(Dated: June 4, 2008)

A Hamiltonian approach to reconstruction of a trajectory and model of complex stochastic dynamics from noisy measurements is introduced. The method converges even when the entire trajectory components are unobservable and the conventional Monte Carlo technique fails. The method is applied to reconstruct the nonlinear models of predator-prey oscillations. We found that the projected (incomplete) character os measurements results in the likelihood distribution with two very different scales: it is strongly localized in the vicinity of a hyperplane in the joint parameters-trajectory space. This reflects the intrinsic tradeoff between the system parameters and hidden trajectory components.

Measurements of complex (multidimensional, nonlinear, stochastic) systems often provide only partial information about the system dynamics. Missing information about the state and the model of the system is normally considered as "lost" in the case of historic time series where measurements cannot be repeated. Classical examples include climate evolution [1], epidemic outbursts [2], and fluctuations in predator-prey communities [3, 4]. How to reconstruct such missing information and deduce both the model and the full system trajectory from incomplete, noise-corrupted measurements is an unresolved scientific conundrum of long standing.

Aspects of this problem have received a considerable attention recently. When measurement noise is absent, attractor reconstruction [5] in time-delayed variables was used [4] to infer model parameters; when dynamical noise is absent, shooting and recursive approaches were applied to reconstruct both parameters and trajectory [6]. A simplified problem, when all the dynamical variables are measured directly, was considered in [7–10].

In this Letter a solution of the general problem is found by introducing a novel Hamiltonian formalism within the log-likelihood minimization framework. The formalism is derived from first-principles Bayesian statistics [11, 12] and stochastic dynamics [13–15]. It allows to efficiently analyze the complex structure of the likelihood and provides a robust convergence, even in the cases where other general methods like Markov Chain Monte Carlo (MCMC) [9, 10] fail.

The formalism is applied to an analysis of archetypical problem in nonlinear dynamics and statistics: noise driven population dynamics in a rodent-predator community [3, 4, 17], where the population of predators could not be *practically* observed. We show that, contrary to earlier belief [3, 4, 17], noise-corrupted measurements of prey dynamics can be used as a basis for computing both the likelihood of the unmeasured predator population and the ecological model parameters. To validate the
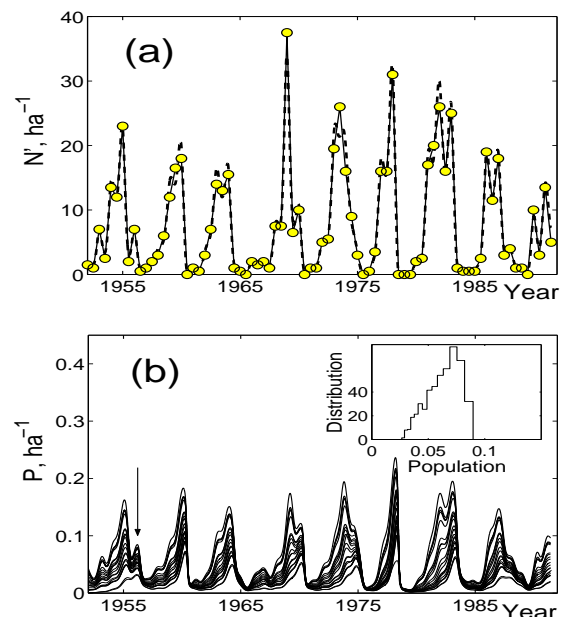

FIG. 1: (a) Measured population dynamics of small rodents in Kilpisjärvi, Finnish Lapland, 1952-1992 [16] is (yellow dots). The full black line is a guide to the eye and the dashed line shows the solution of the optimization problem. (b) Recovered hidden dynamics of the specialist predator population, for different values of parameters r and rK'/K in the model (1). The inset shows a cross-section of the weighted distribution of dynamical trajectories for 1956 (arrow in main figure).

method we apply it to prey data from another rodent-predator community [17] where, unusually, it was possible to record *both* populations, enabling us to compare the measured and inferred predator densities. We anticipate that the formalism will be useful in many different disciplines and contexts.

Time-series data $N(t)$ for rodent (prey) population density oscillations in Finnish Lapland [16] are shown by the yellow points in Fig. 1(a). The dynamical inference problem is to use this limited number (80) of noise-corrupted measurements to recover both the unmeasured time-dependence of the population density of their predators $P(t)$ and the nonlinear stochastic dynamical model underlying the oscillations. The general problem could not be solved earlier especially since the measurement error is significant [4]. Our first step in tackling it is to write down a stochastic model accumulating expert knowledge about predator-prey dynamics in this community [3, 17]

$$\dot{N} = rN\left[1 - e_1 \sin(2\pi t) + \sigma_n \xi_n(t)\right]$$
$$- (r/k)N^2 - \frac{g\,N^2}{N^2 + h^2} - \frac{c\,NP}{N + d}, \quad (1)$$

$$\dot{P} = sP\left[1 - e_2 \sin(2\pi t) + \sigma_p \xi_p(t)\right] - s\,q\frac{P^2}{N}. \quad (2)$$

Parameters $e_{1,2}$ in (1),(2) are dimensionless amplitudes of periodic (seasonal) forcing. $\xi_n(t)$ and $\xi_p(t)$ in (1) are mutually uncorrelated zero-mean white Gaussian dynamical noise sources. Parameters r and s are intrinsic rates of prey and predator growth, respectively; k is prey carrying capacity that determines the saturation of prey population in the absence of predators. The effect of so-called generalist predators (e.g., foxes) whose total population does not appreciably change with N is encapsulated into the term $\propto N^2$ with coefficient g (maximum mortality rate). The "specialist" predators (small mustelids) most likely maintain the oscillatory prey dynamics and their effect is described by the term $\propto NP$ with maximum killing rate c. However their population is notoriously difficult to study in a field. The corresponding density $P(t)$ represents an unmeasured (hidden) variable in our model. The measured rodent density $N'(t)$ is related to the actual (unknown) value $N(t)$ via $N' = Ne^{\sigma_{\mathrm{obs}}\eta(t)}$ where $\eta(t)$ is white Gaussian noise of unit intensity.

The precise functional forms are known neither for predation nor for the predator response, and some modifications of equations (1),(2) have been considered [4]. For this and other predator-prey models there exists a change of variables $x_1(t) = \log(N(t)/k')$, $x_2(t) = \log(q'P/k')$ (for known nominal values of the scaling coefficients k' and q') that transforms the predator-prey and the measurement equations into a simpler form with noise entering the equations only additively

$$\dot{x}_i(t) = K_i(\mathbf{x}(t), \mathbf{c}) + \xi_i(t), \quad (3)$$

$$y_k(t) = \sum_{i=1}^{L} B_{ki}\,x_i(t) + \beta_k(t), \quad (4)$$

where $L=2$ corresponds to the present case. The above set of equations represents a generalization that can be applied to a great variety of stochastic dynamical model reconstruction problems with "hidden variables". The state-vector $\mathbf{x}(t)=\{x_1(t), \ldots, x_L(t)\}$ is governed by an L-dimensional drift vector field with components K_i

and a white Gaussian process with zero-mean components $\xi_i(t)$ characterized by an $L \times L$ correlation matrix $\hat{\mathbf{D}}$. Each vector field component $K_i(\mathbf{x}, \mathbf{c})$ depends on the state vector $\mathbf{x}$ and a set of (unknown) model parameters $\mathbf{c}=\{c_1, \ldots\}$. The measurement equations (4) are described by an $M \times L$ measurement matrix $\hat{\mathbf{B}}$, by a white Gaussian process with zero-mean components $\{\beta_m(t)\}$, and by an $M \times M$ measurement noise matrix N_{nm} ($M \leq L$). The model (3), (4) is characterized by the unknown parameters $\mathcal{M} \equiv \{\mathbf{c}, \hat{\mathbf{B}}, \hat{\mathbf{D}}, \hat{\mathbf{N}}\}$. Existence of hidden variables implies that $M < L$. For the problem at hand the vector of unknown parameters $\mathcal{M} = \{r, s, e_1, e_2, k, g, c, q, h, d, \sigma_n, \sigma_p, \sigma_{\mathrm{obs}}\}$, measurement matrix $B_{ij} = \delta_{i1}\delta_{j1}$ and vector fields K_i can be directly obtained using (1),(2) and the change of variables from N, P to x_1, x_2 described above.

We now turn to the general formulation of the inference problem of finding unknown model parameters $\mathcal{M}$ and unobserved trajectory components ($x_2(t)$ in our case). Because of the dynamical and measurement noise, this problem must be cast in probabilistic terms. A key statistical quantity is the so-called likelihood probability density functional (LPDF) $\mathcal{P}_{\mathcal{Y}}[\mathbf{x}(t); \mathcal{M}]$. It represents the joint probability density that the system trajectory is $\mathbf{x}(t)$ and that the system parameter values are $\mathcal{M}$, conditioned on the observed time-series $\mathcal{Y} = \{\mathbf{y}(t_m), t_m = m\,h, m = 1 : \mathcal{K}\}$. It can be written as $\mathcal{P}_{\mathcal{Y}}[\mathbf{x}(t); \mathcal{M}] = A_{\mathcal{Y}} \exp\left(-S_{\mathcal{Y}}[\mathbf{x}(t); \mathcal{M}]\right)$, where $A_{\mathcal{Y}}$ is a normalization constant. A negative log-likelihood functional $S_{\mathcal{Y}}[\mathbf{x}(t); \mathcal{M}]$ can then be found using the path-integral approach to fluctuational dynamics [14, 15], as

$$S_{\mathcal{Y}} = \frac{1}{2}\int_0^T dt\,(\mathcal{L}_{yx} + \mathcal{L}_x) + \frac{\mathcal{K}}{2}\ln\det\hat{\mathbf{D}}\hat{\mathbf{N}}, \quad (5)$$

$$\mathcal{L}_{yx} = [\mathbf{y} - \hat{\mathbf{B}}\,\mathbf{x}]^T\hat{\mathbf{N}}^{-1}[\mathbf{y} - \hat{\mathbf{B}}\,\mathbf{x}],$$

$$\mathcal{L}_x = [\dot{\mathbf{x}} - \mathbf{K}(\mathbf{x}, \mathbf{c})]^T\hat{\mathbf{D}}^{-1}[\dot{\mathbf{x}} - \mathbf{K}(\mathbf{x}, \mathbf{c})] + \frac{\partial\mathbf{K}(\mathbf{x}, \mathbf{c})}{\partial\mathbf{x}}.$$

Here $\mathbf{x}=\mathbf{x}(t)$, $\mathbf{y}=\mathbf{y}(t)$. Also $T = \mathcal{K}h$ where $\mathcal{K}$ is the number of data points, h is the sampling time, and dimension L of the state vector $\mathbf{x}$ is greater than the dimension ML of the observation vector $\mathbf{y}$, implying the existence of hidden variables. Despite the hidden dynamical variables not being measured directly, the functional $S_{\mathcal{Y}}$ (5) contains the dynamical coupling between the variables via the force fields $K_i(x_1, x_2)$.

If the measurements provide information sufficient to pin down both the key model parameters of the system and its trajectory, the joint LPDF $P_{\mathcal{Y}}(\mathbf{x}(t), \mathcal{M})$ is well-localized in the vicinity of its maximum, thus revealing the most probable trajectory $\mathbf{x}^*(t)$ and parameter values $\mathcal{M}^*$ that provide minimum of functional $S_{\mathcal{Y}}(\mathbf{x}(t), \mathcal{M})$ for a given set of measurements $\mathcal{Y}$. Paragraph is removed

We solve this variational problem by introducing a new paradigm in which $S_{\mathcal{Y}}(\mathbf{x}(t), \mathcal{M})$ is viewed as the mechanical action of an auxiliary Hamiltonian system with coordinate $\mathbf{x}$, momentum $\mathbf{p} = \hat{\mathbf{D}}^{-1}(\dot{\mathbf{x}}(t) - \mathbf{K})$ and a Hamil-

tonian function $H(\mathbf{x}, \mathbf{p})$

$$H(\mathbf{x}, \mathbf{p}) = -\frac{1}{2}\left(\mathbf{y} - \hat{\mathbf{B}}\,\mathbf{x}\right)^T \hat{\mathbf{N}}^{-1}\left(\mathbf{y} - \hat{\mathbf{B}}\,\mathbf{x}\right)$$
$$-\frac{1}{2}\frac{\partial \mathbf{K}}{\partial \mathbf{x}} + \mathbf{K}\mathbf{p} + \frac{1}{2}\mathbf{p}^T \hat{\mathbf{D}}\,\mathbf{p}. \qquad (6)$$

The extremum of $S_{\mathcal{Y}}(\mathbf{x}(t), \mathcal{M})$ in the joint space $(\mathbf{x}(t), \mathcal{M})$ is then found by solving the coupled variational problems

$$\delta S/\delta \mathbf{x}(t) = 0, \quad \partial S/\partial \mathcal{M} = 0. \qquad (7)$$

The first condition corresponds to a solution of the boundary value problem (BVP) for the Hamiltonian equations

$$\dot{\mathbf{x}} = \mathbf{K} + \hat{\mathbf{D}}\mathbf{p},$$
$$\dot{\mathbf{p}} = \frac{1}{2}\frac{\partial^2 \mathbf{K}}{\partial \mathbf{x}^2} - \frac{\partial \mathbf{K}}{\partial \mathbf{x}}\mathbf{p} - \left(\mathbf{y} - \hat{\mathbf{B}}\,\mathbf{x}\right)^T \hat{\mathbf{N}}^{-1}\hat{\mathbf{B}} \qquad (8)$$

that satisfy the boundary conditions $\mathbf{p}(0) = \mathbf{p}(t) = 0$. This BVP can be solved very efficiently in many cases using, e.g., results of [18]. An interesting property of this solution seen from the structure of Eqs. (8) is that measured data $\mathbf{y}(t)$ effectively act as a "control force" of an amplitude inversely proportional to the measurement noise intensity (the difference between $\mathbf{y}(t)$ and $\mathbf{x}(t)$ is fed back in the last term of the second equation in (8)). This control force drives the $\mathbf{x}(t)$ towards the most probable one (at a fixed $\mathcal{M}$). We then fix the inferred trajectory $\mathbf{x}(t)$ and update the values of the parameters in the set $\mathcal{M}$, using an analytic solution [8] of the second variational problem, $\frac{\delta S}{\delta \mathcal{M}} = 0$. This procedure is iterated until the desired convergence is achieved. The outcome of this algorithm is the most probable system trajectory $\mathbf{x}^*(t)$ and set of model parameters $\mathcal{M}^*$ providing a minimum to the action functional $S_{\mathcal{Y}}$. It corresponds to a point $p_*=(\mathbf{x}_*(t), \mathcal{M}_*)$ in the joint trajectory-parameter space.

Note however, that in the presence of hidden variables slight changes in the model parameters can lead to significant changes in the system's trajectory giving rise to a complex shape of the $S_{\mathcal{Y}}$. This proved to be the case for the problem at hand because dynamical model (1) with periodic driving force undergoes the onset of bifurcations and the transition into a chaos when the system parameters vary [4, 19]. To investigate the shape of $S_{\mathcal{Y}}$ we define a window of possible values for each model parameter from the set using the approximate plausible ranges of parameter elucidated in earlier research [3, 4, 17]. For many parameters the windows are very broad. We show some of them as rectangular planes in Figs. 2. Starting from a random point inside of the predefined range we find a point $p_*=(\mathbf{x}_*(t), \mathcal{M}_*)$ corresponding to one of the solutions of a dual variational problem (7). Random restarts are needed if there is (as in our case) more then one solution p_*. Projections of the points p_* onto the 2D parameter plains and corresponding 3D plots of LPDF are shown in Figs. 2.
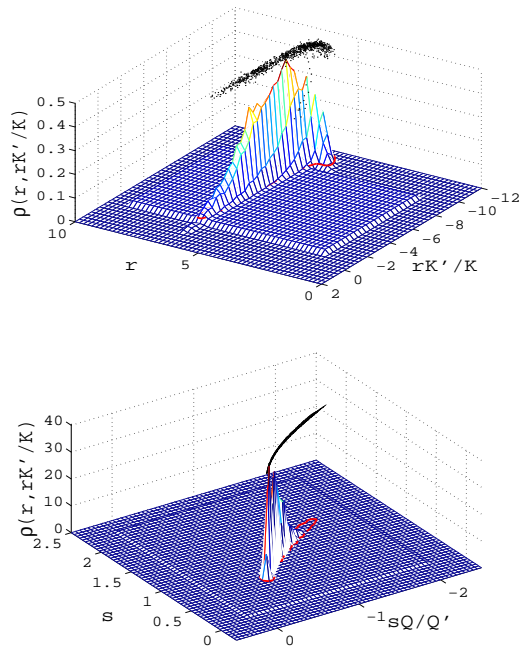

FIG. 2: Weighted distributions of the inferred values (a) of the model parameters r and rK'/K, and (b) of the model parameters s and sQ/Q'. For some other parameters their likelihood distributions are well-localized around their expected values $r = 5.69 \pm 0.49$, $k = 76 \pm 17$ $s = 1.08 \pm 0.31$, $q = 43 \pm 22$, $g = 0.12 \pm 0.3$, $e_1 = 1.4 \pm .4$, $e_2 = 1 \pm .5$ which are close to the values considered in the earlier ecological research. The projections of the points p_* onto the 2D parameter plains are shown at the top of both figures by black dots.

One can infer from these figures that LPDF has a form of a very steep ridge localized in the vicinity of a ridge top forming a "line" $\mathcal{R}=\{p_*^\ell\}=\{(\mathbf{x}_*^\ell(t), \mathcal{M}_*^\ell\}$ in the joint parameter-trajectory space that can be seen in the density plots in Figs. 2. Fluctuations along the ridge top indicated by the superscript ℓ above are large and non-Gaussian as can be seen from Figs. 2. However fluctuations in the direction transverse to the ridge top are strongly suppressed. The mechanical action along the $\mathcal{R}$ is varying slightly and each point at $\mathcal{R}$ (approximately) corresponds to a solution of (7). This difference is scales and a quasi-degenerate form of the $S_{\mathcal{Y}}$ and LPDF can be attributed to the **projective** character of the measurements with $M < L$ that are incapable of resolving certain tradeoffs between the parameters and hidden trajectory components. We note that different points along R correspond to quite different time traces of the hidden trajectory component (predator oscillations) shown in Fig. 1(b) but they all fit very well to the the set of prey population observations. Such tradeoffs are intrinsic to a given dynamic and measurement models and their understanding is crucial to the domain experts for model discovery and data interpretation.
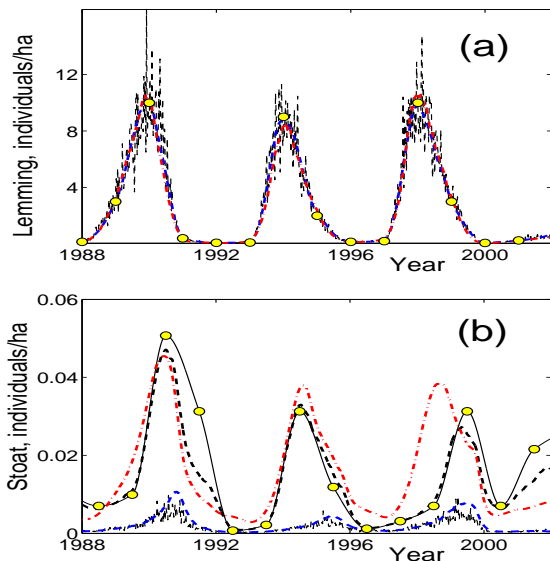
FIG. 3: Populations of (a) lemmings and (b) stoats (individuals/ha) observed in the high-Arctic tundra (1988–2002) [20] are shown by yellow ellipses. Jagged thin solid lines show the initial guesses for each population. Dash-dot red lines show the population dynamics inferred from (1) using the Hamiltonian approach and assuming that *only* the lemming population was determined, with a measurement error of 0.1, while the stoat dynamics remained hidden. Dashed blue lines are from MCMC after 40000 iterations of the whole trajectory, under the same assumption.

Next, we validate the approach by applying it to the inference of population dynamics in a stoat-lemming community [17] and by comparing it to Markov Chain Monte Carlo (MCMC) method. The corresponding data (Fig. 3) are similar to the earlier set [?] but with an important difference: the dynamics of *both* populations, prey $N(t)$ and predator $P(t)$, were recorded. We assume initially that only the prey population was measured, and and we subsequently compare the inferred predator population with the measured one. The dynamical model [17] for this community is rather elaborate. However, given the close similarity between the two communities it is in-

structive to perform such a comparison using a simplified model (1). The corresponding results are shown in Fig. 3. It is evident that the predator trajectory is reproduced, including the correct slope and scale of the oscillations, unlike the earlier simulations [17]. It can also be seen that the MCMC method gives a poor result because it takes so long to converge and because it tends to get stuck in local minima close to the original guesses. We found that these local minima correspond to non-smooth time functions $P_*(t)$ that are excluded from the start in our Hamiltonian approach (6)-(8).

In Conclusion, we presented a method based on Hamiltonian dynamics for inference of parameters and hidden (unobserved) trajectory components of a nonlinear dynamical model in the presence of noise and measurement errors. We applied the method to ecological problem where the most important conclusion is that a lack of observational data for predator populations need no longer constitute a fundamental obstacle to the inference of ecological parameters from noisy measurements [3, 4, 17, 21]. The method allowed us to uncover a quasi-degeneracy of the log-likelihood along certain hyperplane $\mathcal{R}$ in the joint parameters-trajectory space. The existence and shape of $\mathcal{R}$ reflects the tradeoffs between parameters r and k'/k (s and q/q') on one hand and and hidden trajectory components on the other. An important feature of our procedure is that, unlike MCMC, it avoids sampling in the trajectory space. Instead, it relies on continuous in time solution of the deterministic auxiliary Hamiltonian problem (8), thereby exploiting recent advances in [18] the BVP problem. The method will also be applicable quite generally to cases where some state variables were not, or could not be, recorded, e.g. those mentioned in the introduction [1, 2]. It can be particularly useful in the context of physiological measurements relating difficult-to-access parameters to noninvasively-measured data [22], and to cases where dynamical variables are intrinsically hidden and cannot be measured experimentally, even in principle: e.g. flow parameters in aerospace applications [23], susceptible and exposed populations in epidemiology [2, 24].

[1] S. Rahmstorf, Nature **419**, 207 (2002), 10.1038/nature01090.

[2] D. J. D. Earn, S. A. Levin, and P. Rohani, Science **290**, 1360 (2000).

[3] I. Hanski, *et al*, Ecology **82**, 1505 (2001).

[4] P. Turchin and S. P. Ellner, Ecology **81**, 3099 (2000).

[5] C. W. Tidd, L. F. Olsen, and W. M. Schaffer, Proc R Soc Lond B Biol Sci. **254**, 257 (1993).

[6] H. U. Voss, J. Timmer, and J. Kurths, Int. J. Bifurc. and Chaos **14**, 1905 (2004).

[7] A. A. Poyton, M. S. Varziri, K. B. McAuley, P. J. McLellan, and J. O. Ramsay, Computers & Chemical Engineering **30**, 698 (2006).

[8] V. N. Smelyanskiy, D. G. Luchinsky, D. A. Timucin, and A. Bandrivskyy, Physical Review E **72**, 026202 (pages 12) (2005).

[9] R. Meyer and N. Christensen, Phys. Rev. E **65**, 016206 (2001).

[10] C. Calder, M. Lavine, P. Müller, and J. S. Clark, Ecology **84**, 1395 (2003).

[11] P. Congdon, *Bayesian statistical modelling*, Wiley series in probability and statistics (Wiley, Chichester, 2001).

[12] G. D'Agostini, *Bayesian reasoning in high-energy physics: principles and applications* (CERN, Geneva,

1999).

[13] M. Freidlin and A. D. Wentzel, *Random Perturbations in Dynamical Systems* (Springer, New-York, 1984).

[14] R. Graham, Z. Phys. B **26**, 397 (1977).

[15] M. I. Dykman, Phys. Rev. A **42**, 2020 (1990).

[16] NERC Centre for Population Ecology, Imperial College *The Global Population Dynamics Database* Record 3987. http://www.sw.ic.ac.uk/cpb/gpdd.html. (1999).

[17] O. Gilg, I. Hanski, and B. Sittler, Science **302**, 866 (2003).

[18] U. M. Ascher, R. M. M. Mattheij, and R. D. Russel, *Numerical solutions of Boundary Value problems for ordinary differential equations*, Classics in Applied Mathematics (SIAM, Philadelphia, 1995), 2nd ed.

[19] T. P. K. E. Hanski, I. and H. Henttonen, Nature **364**, 232 (1993).

[20] O. Gilg, I. Hanski, and B. Sittler, Science **302**, 866 (2003).

[21] P. Turchin and I. Hanski, American Naturalist **149**, 842 (1997).

[22] V. N. Smelyanskiy, D. G. Luchinsky, A. Stefanovska, and P. V. E. McClintock, Physical Review Letters **94**, 098101 (pages 4) (2005).

[23] V. V. Osipov, D. G. Luchinsky, V. N. Smelyanskiy, and D. A. Timucin, *Proc. AIAA/ASME/SAE/ASEE Joint Propulsion Conf. and Exhibit*, *AIAA Conference Proceedings* (AIAA, Cincinnati, OH, 2007), AIAA 2007-5823 (2007).

[24] J. L. Aron and I. B. Schwartz, J. Theor. Biol. **110**, 665 (1984).