A new cost function for model fitting: the length-scale of dynamical consistency

A. B. Campbell

Queensland Department of Primary Industries and Fisheries
Email: alex.campbell@dpi.qld.gov.au

Abstract: This article introduces a cost function for model fitting that exploits geometrical properties of the models' dynamics to improve the estimation of uncertainty. The cost function can be thought of as an operational way of asking the following question: are the observations consistent with the dynamics of the model to within the precision granted by a prescribed level of uncertainty? Model parameters which allow for low levels of uncertainty while retaining consistency are taken to provide a better fit than parameters for which the level of uncertainty must be increased if consistency is to be preserved. The minimum level of uncertainty is the statistic which is returned by the cost function for a given set of parameters, similar to the way sums of squared errors is returned by a maximum likelihood function with Gaussian errors. Unlike maximum likelihood however, this output can be interpreted directly in physical (or biological) terms: it is the smallest length scale of the system variable (e.g. biomass) at which predictability can be maintained; that is, forecasts can be made within plus or minus this length scale. We term this cost function the 'length-scale of dynamical consistency' (LSDC).

All techniques for model fitting must at some point, directly or indirectly, measure a discrepancy (that is, a distance) between model predictions and reality (the observations). A critical methodological step taken in this paper is that we are no longer measuring this distance in the Euclidean space of the observations, but rather in the (almost always) non-Euclidean space of the dynamics.

The cost function is validated with a numerical experiment where the true dynamics are known and noise is artificially added. The experiment confirms that the technique is accurate with respect to converging on the true parameter values, but also that it provides additional information about parameter uncertainty compared to a reference least squares approach. In particular it clearly identifies a small region of parameter space with minimal length scale, the shape of which reflects the parameter confounding set up in the experiment, and for surrounding regions the length scale jumps sharply. In contrast the least squares approach displays a relatively flat solution surface. The cost function is also applied to a real data set, a series of catch and effort data from Australia's northern prawn fishery. The LSDC results and those of a reference maximum likelihood implementation differ in a non-trivial fashion and further work is needed to interpret this fully.

While the theoretical framework is explained in some detail, the results presented here are preliminary in nature and numerical investigations are ongoing.

Keywords: Consistent nonlinear dynamics, parameter estimation, surplus production model.

1 INTRODUCTION

Insofar as mathematical models codify our understanding of the world, the central task of science is to develop models which have some basis in reality. Models make contact with reality through data, and assessing how well a given model 'fits' the data is of fundamental importance. In ecological and biological domains, and in particular in population biology, this assessment is hampered by the paucity of data available relative to the complexity of the system under study. Harvested populations (such as fisheries) are particularly important modelling domains because they need careful management if they are to be both productive and sustainable, and this is only possible through successful models. They are also particularly complex in that the underlying rhythms of the ecological system are perturbed, or forced, by human activity.

Least squares (Björck, 1996) and total least squares (van Huffel and Vandewalle, 1991) are methods to determine how well a model fits a given data set. They are tried and tested approaches, however for nonlinear and data sparse problems like population dynamics they have limitations. In particular, it can be shown that in numerical experiments where the 'true' model is known, least squares will reject the true model. This is because the true model by definition operates on noise-free values, and a model that operates on noisy values exploits this additional variation to minimise forecast error beyond what the true model could (Kantz and Schreiber, 2002, p. 253). A different approach to modelling is to ask the following question: given an inherent 'noise

floor' of uncertainty, below which we shall not enquire, are the observations *consistent* with the dynamics of the model? This question is the basis of Consistent Nonlinear Dynamics (CND) devised by McSharry and Smith (McSharry and Smith, 2004).

The advantages of incorporating dynamics into the model validation process are several, including: preferring models that can reproduce the observations over models with the lowest forecast error, the former arguably more valuable if we are primarily interested in *understanding* the system; being able to identify *regions* of state space where the model is deficient; and providing insight into the natural length scales on which the data/phenomena should be investigated. However, CND is not exactly what we would want for model fitting in many real world situations. The first issue is topological in nature: our observations usually lie in a different 'space' to that of the model. This issue is relegated to a footnote in McSharry and Smith (2004), where it is stated that although in general the model-state space will differ from the true state space of the system, 'the projection from one space to another poses several foundational difficulties' (McSharry and Smith, 2004, p. 2). It is a major purpose of this article to point out that this is a difficulty that needs overcoming in order to apply CND style analysis to population biology style problems, and to propose a solution. A second issue with CND which we address is that in its current form it is suitable only for autonomous systems, that is, systems without any forcing (this rules out harvested populations). Finally we incorporate some ideas from Pecora et al. (2007) to make the technique intelligently sensitive to the amount of data.

The cost function is developed in Section 2. Sections 3 and 4 detail the evaluation of the cost function on synthetic and real data sets respectively. Section 5 concludes and discusses future work. A significantly expanded version of this work, including more background theory and numerical results, can be found in Campbell (2009).

2 METHODOLOGY

Consider a functional relationship between two variables, $y = f(x)$, and a set of observations of these variables, $\{s_n, x_n\}_{n=1}^N$, where $s_n = y_n + \eta$ is a noisy measurement of y and x_n a corresponding noise-free measurement of x. We wish to find a model $y' = g(x')$ such that the relationship between y' and x' under the action of g is in some sense 'close' to the relationship between x and y under the action of f. If we have some idea of the functional form for g in terms of parameters a (or even if we don't), a standard approach is to find the parameter values $\hat{a}$ that minimise the prediction error when g is used on the observed data. If the measurement errors η are assumed to be independent and normally distributed then a provably optimal approach is to minimise the familiar 'least squares' cost function:

$$\mathcal{C}_{LS}(a) = \sum_n (y_n - \hat{y}_n)^2 \tag{1}$$

where $\hat{y}_n$ is a prediction of y_n under the model, that is $\hat{y}_n = g(s_n; a)$.

Even if the η are reasonably well behaved, the assumption that the so called independent variable x is measured without error is often problematic. In population biology for example we are often most interested in functional relationships of the form

$$x_{i+1} = f(x_i) \tag{2}$$

where x is some measure of population, and i has replaced n as an index to emphasise temporal structure. If $s = x + \eta$ as before then the values obtained for both sides of the equation by definition cannot be noise free, and a least squares approach (ie $(s_i - \hat{s}_i)^2$) will be biased even with an infinite amount of data (McSharry and Smith, 1999). More generally we desire to understand and model relationships of the form

$$x_{i+1} = f(x_i, z_i) \tag{3}$$

where z is some exogenous forcing variable (for example catch in a fisheries model).

The extension of the least squares approach to the time series model fitting problem leads to an improved 'total least squares' cost function that takes into account the geometry of the dynamics:

$$\mathcal{C}_{TLS}(a) = \sum_n ||\vec{s}_i - \tilde{\vec{x}}|| \tag{4}$$

where $\vec{s}_i \triangleq [s_i, s_{i+1}] \in \mathbb{R}^2$ and $\tilde{\vec{x}} \triangleq [\tilde{x}, g(\tilde{x}; a)] \in \mathbb{R}^2$ denote points in the two-dimensional space required to unfold the dynamics of g such that all points $[\tilde{x}, g(\tilde{x}; a)]$ lie on a single-valued curve, and $||\cdot||$ denotes the Euclidean norm. Figure 1 illustrates the distinction between least squares and total least squares for a system of the form of equation 2. Note that the true x_i are unknown, only the curve $[\tilde{x}, g(\tilde{x}; a)]$ is known (and then only for a given value of a).

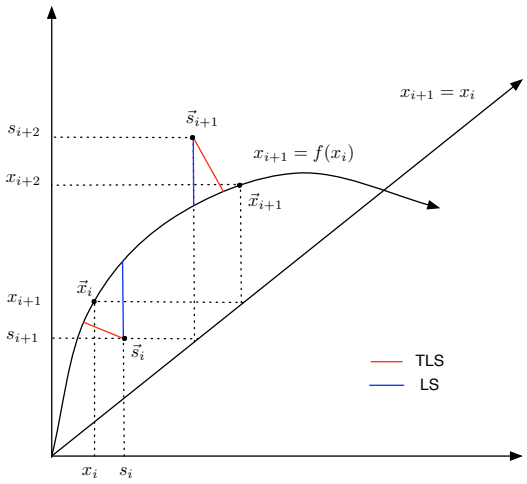


Figure 1: Standard least squares minimises the vertical distances (blue); total least squares minimises the perpendicular distances (red).

Least squares and total least squares both minimise the prediction error - least squares the one-step error, and total least squares the error perpendicular to the geometry of the dynamics. However, for noisy observations of nonlinear systems, minimising the prediction error is not equivalent to finding the best model, and in fact it can lead to a known true model being rejected. This is because the true model by definition operates on noise-free values, and a model that operates on noisy values exploits this additional variation to minimise forecast error beyond what the true model could (Kantz and Schreiber, 2002). A different approach to modelling is to ask the following question: given an inherent 'noise floor' of uncertainty, below which we shall not enquire, are the observations *consistent* with the dynamics of the model? This question is the basis of Consistent Nonlinear Dynamics (CND) devised by McSharry and Smith (McSharry and Smith, 2004) and an illustration of the central idea is given in Figure 2. Informally, the observations are deemed consistent with the dynamics if, for all $\vec{s}_i$, the one-step dynamical evolution of its 'ϵ-ball' has a non-zero intersection with an ϵ-ball centered on $\vec{s}_{i+1}$. This overlap implies the existence of at least one trajectory that is consistent with both model and data. Multi-step prediction horizons are a natural extension.

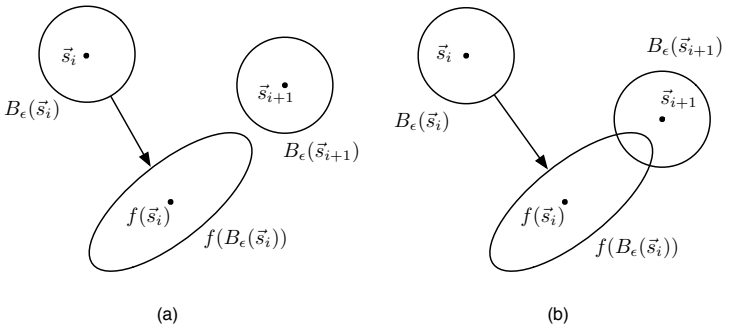


Figure 2: Inconsistent (a) and consistent (b) dynamics (McSharry and Smith, 2004).

The advantages of incorporating dynamics into the model validation process are several, including: preferring models that can reproduce the observations over models with the lowest forecast error, the former arguably more valuable if we are primarily interested in *understanding* the system; being able to identify *regions* of state space where the model is deficient; and providing insight into the natural length scales on which the data/phenomena should be investigated. However, CND is not exactly what we would want for model fitting in many real world situations. The first issue is topological in nature: our observations usually lie in a different 'space' to that of the model. For example, our observations, $\vec{s}_i$ in Figure 1 lie in $\mathbb{R}^2$, whereas the model state-space is the one-dimensional curve defined by $y = g(x; a)$. This issue is relegated to a footnote in McSharry and Smith (2004), where it is stated that although in general the model-state space will differ from the true state space of the system, 'the projection from one space to another poses several foundational difficulties' (McSharry and Smith, 2004, p. 2). It is a major purpose of this article to argue that this is a difficulty that needs overcoming in order to apply CND style analysis to population biology style problems, and to propose a solution. A second issue with CND is that in its current form it is suitable only for autonomous systems, that is, systems of the form of equation 2. We extend its applicability to systems of the form of equation 3 (essential for harvested population analysis). Finally we incorporate some ideas from Pecora et al. (2007) to make the technique intelligently sensitive to the amount of data.

2.1 Projecting into model space

The solution to the projection problem is to realise that an ϵ-ball level of uncertainty is valid in any space, whether that be the Euclidean space arising from delayed instances of an observable (e.g. $[s_i, s_{i+1}] \in \mathbb{R}^2$) or the curved manifold of model space. Furthermore, this ball is required to be of a minimum radius if individual observations are even to be *potentially* consistent: that is, this ball must intersect the manifold. Potential

consistency is a pre-requisite for dynamical consistency; if this has been attained, the region of the manifold that intersects with the ϵ-ball (this intersection is well-defined regardless of the different dimensionality of the manifold to observation space) is taken to be the true $B_\epsilon(\vec{s}_i)$. The other aspects of CND follow naturally as illustrated in Figure 3 (although see Campbell (2009) for more details on observation space vs model space uncertainty).

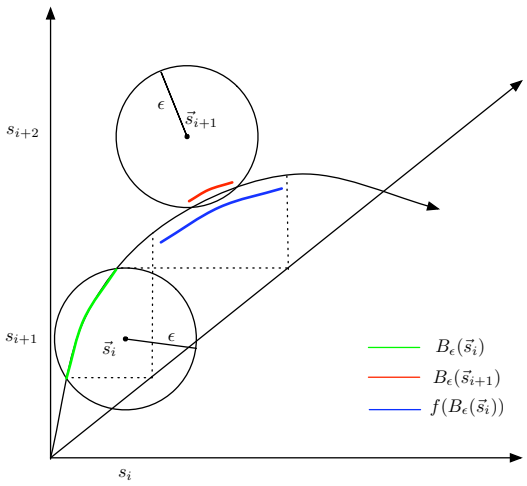


Figure 3: Model-space projection.

This projection step has the following consequences. Firstly, observations which are far from the manifold require that ϵ is large if they are to even have potential consistency. This is intuitively satisfactory. Secondly, for a given ϵ, $\vec{s}$ that are further from the manifold than others leave a smaller ' ϵ-ball footprint' - this means that the dynamics on the manifold must be more consistent with the observations in this region than if $\vec{s}$ were closer and had a larger footprint. Thus there is a trade-off between geometrical error (distance from the manifold) and dynamical error (dynamics on the manifold). Essentially this means that a lone irregular (large error) observation can still be considered consistent (assuming it has potential consistency) if its temporal neighbours were close to the manifold, giving them bigger ϵ-footprints and thus a greater chance of intersection with the relatively small ϵ-footprint of the irregular observation. This also is intuitively satisfactory and moreover it emphasises that the technique treats error in some sense holistically - consistency cannot be determined without the context of its temporal neighbourhood (this notion becomes particularly relevant when considering multiple step forecast horizons). In this sense the projection step maximises the 'efficiency' of the data with respect to the information they provide to the modelling process.

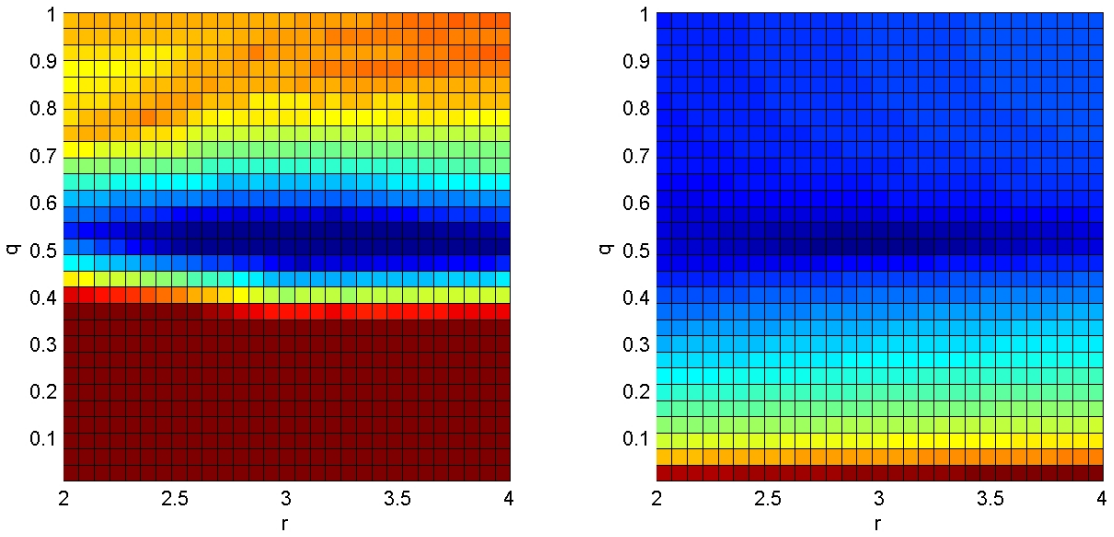
Model-space projection is especially important in situations where there may be a significant discrepancy between observation space and model (manifold) space. This discrepancy is obviously a function of the phenomena under study, the data collection process, and the models put forward to explain the phenomena. We posit that in many fields of study this discrepancy will be non-trivial, and that the model-space projection step is especially important in areas such as ecology and population biology in particular where observation-model space mismatch is combined with sparsity of data.

2.2 Extension to non-autonomous dynamics

If the system is of the form of equation 2, $B_\epsilon(\vec{s}_{i+1})$ should have non-zero intersection with $f(B_\epsilon(\vec{s}_i))$. However with a non-autonomous system of the form of equation 3, with z as a forcing variable, then $\vec{s}_i$ and $\vec{s}_{i+1}$ arose in different contexts (due to z_i and z_{i+1} respectively), and are not directly comparable. The solution is to calculate $\vec{s}_i'$: what $\vec{s}_i$ would be *if the forcing value were* z_{i+1} *instead of* z_i, and test for the intersection of $f(B_\epsilon(\vec{s}_i'))$ with $B_\epsilon(\vec{s}_{i+1})$. Implementation-wise this is relatively straight-forward.

2.3 Data quantity dependence

One issue that has not been addressed thus far is: for a given data set, how many point-pairs must be checked for consistency, and do we require *all* of them to be consistent? The first question is just trade off between accuracy and computation, and in data sparse ecological situations it is both more important and computationally easier to check all point pairs (and this is what we do in Sections 3 and 4). The answer to the second question is definitely no: a useful way to think about ϵ is that it represents the 95% (or 99% or 99.9% etc) isoline (contour) of an unbounded noise distribution; if one wanted all points to be consistent under this scenario one would theoretically always need an infinite ϵ. McSharry and Smith (2004) use the fraction of inconsistent predictions (along with their location in state-space) as a diagnostic. This is a good potential candidate for our cost function (the best parameter value for a given model is that which minimises the fraction of inconsistent predictions), however we instead borrow a clever trick from Pecora et al. (2007) and use a hypothesis test to determine overall consistency. The correct distribution for this test is the Binomial because we are counting the number of consistent predictions out of a total number of trials. For a given model $g(\cdot; a)$ a given parameter value a^*, and a given ϵ, resulting in k consistent predictions from n trials, consistency is achieved if $\binom{n}{k}p^k(1-p)^{n-k} < \alpha$



(a) length scale of dynamical consistency

(b) (logarithm of) sums of squares

Figure 4: LSDC and sum of squares measures of goodness of fit

where p is the *a priori* probability of a consistent prediction (in this paper we use 0.5) and α the probability below which we refute the null (that there were k consistent predictions from n trials just by chance; in this paper set to 0.05). The choice of $p = 0.5$ is a simple but robust choice, making it quite hard to reject the null, however ultimately the choice of values p and α should be seen as application-dependent, for more discussion on this see Pecora et al. (2007). This hypothesis test approach has the advantage that it is intelligently sensitive to the amount of data: for very small data sets (6 points or less) no inconsistent predictions are allowed, as the data set grows the number allowed increases.

This leads us to our final 'length scale of dynamical consistency' (LSDC) cost function:

$$\mathcal{C}_{\text{LSDC}}(a) = \min_{\epsilon} : \text{Binom}\left(\sum_{i=1}^{N-d+1} g(\psi(B_\epsilon(\vec{s}_i)); a) \circ \psi(B_\epsilon(\vec{s}_{i+1})), N-d+1 \right) < \alpha \tag{5}$$

where $\psi : R^d \rightarrow R^d$ is the model-space projection operator and $a \circ b$ is an 'overlap' operator that returns 1 if a and b have non-zero intersection and 0 if not.

3 NUMERICAL EVALUATION

Numerical evaluation of the cost function was carried out on the following system:

$$x_{i+1} = x_i + rx_i - rx_i^2 - qy_i x_i \tag{6}$$

This is essentially a stripped back surplus production model, with the capacity parameter dropped. It crucially retains the potential for overall 'biomass' (x) to be confounded between r and q parameters, while simplifying to allow easy visualisation of parameter space. 20 noisy observations $s_i = x_i + \eta_i \sim \text{U}(-\sigma/2, \sigma/2)$ (ie uniformly distributed with mean zero and range σ) were generated using $r = 3$, $q = 0.5$ and $\sigma = .1$. Grid-based search over the parameter values resulted in Figure 4(a) for the LSDC cost function and Figure 4(b) for a reference least squares approach.

These results a) confirm basic functionality of the LSDC cost function and b) indicate that it provides information that is not. For example, the LSDC parameter space displays much greater relative contrast between the confounded region $0.4 < q < 0.6$ and adjacent parameter regions. Note that in this example the confounding is *designed* and inevitable: we are not expecting either technique to be able to 'see through it' to the true parameter values, the point is to understand how the different techniques handle this confounding. One could think of the philosophy of the LSDC approach in relation to this as: 'given that confounding exists, let us label those confounded regions of parameter space as clearly and accurately as possible'.

4 APPLICATION TO FISHERIES DATA

The technique was applied to a time series of catch and effort data from Australia's northern prawn fishery (data obtained from Ciccosillo (2008)). This fishery has a significant historical record as well as having been well researched, making it suitable as a reference data set on which to trial new approaches. The model applied used the classic Schaefer form for the biomass dynamics (Schaefer, 1954):

$$b_{t+1} = b_t + rb_t - rb_t^2/k - \begin{cases} c_t \\ qe_tb_t \end{cases} \tag{7}$$

where b_t is biomass in year t and e_t is effort in year t. Historical catches (c_t) and an initial biomass ratio of $b_1/k = .76$ (maximum likelihood value, lognormally distributed errors) were used to generate a time series of predicted biomass, from which a value for q was estimated using the closed form equation

$$\hat{q} = e^{\frac{1}{T}\sum ln(\frac{c_t/e_t}{b_t})} \tag{8}$$

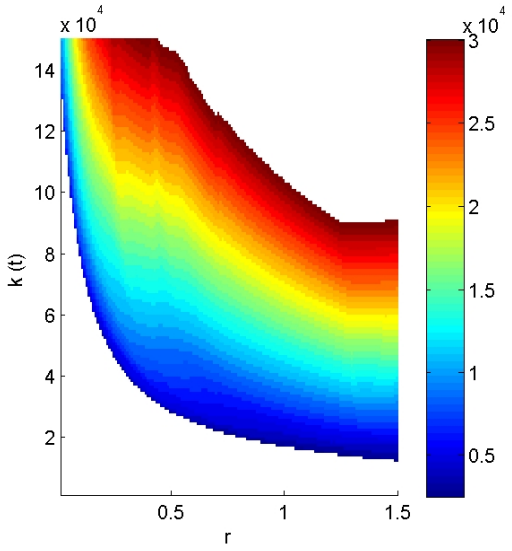


Figure 5: $\mathcal{C}_{\text{LSDC}}$

Results for the dynamical consistency cost function are given in Figure 5 and for a reference maximum likelihood approach in Figure 6. These results highlight the different behaviours of the two approaches. The maximum likelihood approach reaches a maxima at $(r \approx .4, k \approx 40000)$, with the 95% confidence region being quite elongated indicating confounding along the major (curved) axis of this region (the catch and effort trends in the data track each other very closely so this is expected). In contrast to this the LSDC cost function indicates that the best fit is in the implausible region: the length-scale continues to decrease the closer the parameters get to values for which historical catches caused negative biomass to occur (white/non-coloured region). These preliminary findings need further investigation.

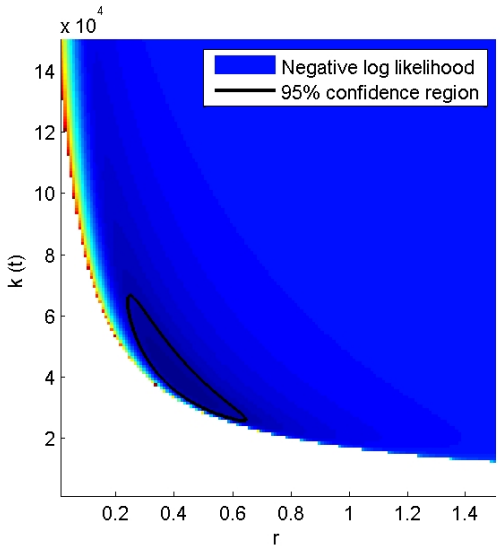


Figure 6: Maximum likelihood

One important feature of Figure 5 is that values of the cost function can be meaningfully interpreted: they are the smallest length-scale of biomass for which the model and the data are consistent. That is, the model is able to make one-step predictions with an accuracy of plus or minus this length scale at the 95% confidence level (assuming α is set to 0.05).

5 CONCLUDING REMARKS

This paper has introduced a new cost function for model fitting. Experiments using synthetic and real data have indicated that it may reveal a more detailed picture of parameter-space uncertainty than maximum likelihood. These results are preliminary and further investigation is required to guarantee its validity and fully explore its implications. The theoretical framework developed in this paper extended the applicability of Consistent Nonlinear Dynamics to real world problems by introducing a model-space projection step and allowing for non-autonomous (forced) systems.

The framework developed here can be seen as a member of the growing set of applications of nonlinear dynamics to ecological problems, starting perhaps with May (1976), and recent prominent examples of which are (Moniz et al., 2007; Nichols et al., 2005; Habeeb et al., 2005). A central theme in this work has been tackling the problems posed by ecological data sets which are both noisy and comparatively short. Because in this paper

we assume that we are given the model space *a-priori* (from domain knowledge), we do not need to rely on the reconstruction theorems (Takens, 1981; Sauer et al., 1991) and consequently the framework has no in-principle problem with short data sets. One way to interpret this is that domain knowledge, represented by the model (and hence the geometric properties of model space) is in some sense making up for the information deficit due to the short noisy data set.

There are some assumptions and limitations of the technique as presented that will be looked at in future work. These include: assumption that the forcing variable (z) has been measured without error, and the assumption that there is no 'process' error (error in the dynamics). Another consideration is the computability of the projection into model space. In the examples above, and we suspect, in quite a large number of population dynamics models, this can be done in large part analytically. Given an analytical expression for the model dynamics in terms of heights of a single valued surface, the main computational task is root finding (polynomial, which can be done efficiently even for quite high order polynomials) for the footprints and integrations to get arc lengths for the actual projection. However the approach is not limited by the requirement of being able to analytically derive model space from Euclidean space - a very interesting development in the machine learning community has been the emergence of algorithms to 'learn' a manifold from observations, see Tenenbaum et al. (2000). This implies that all we need to do is construct the model space manifold by simulating the model and applying a manifold learning algorithm to the output.

ACKNOWLEDGMENTS

The author would like to thank Mai Tanimoto, Michael O'Neill and an anonymous reviewer for constructive feedback.

REFERENCES

Björck, Å. (1996). *Numerical Methods for Least Squares Problems*. Philadelphia: SIAM.

Campbell, A. B. (2009). On the nonequivalence of dynamical and statistical uncertainty. In preparation.

Ciccosillo, R. (2008). Northern prawn fishery data summary 2007. Logbook program, Australian Fisheries Management Authority, Canberra.

Habeeb, R. L., J. Trebilco, S. Wotherspoon, and C. R. Johnson (2005). Determining natural scales of ecological systems. *Ecological Monographs 75*(4), 467–487.

Kantz, H. and T. Schreiber (2002). *Nonlinear Time Series Analysis*. Cambridge, UK: Cambridge University Press.

May, R. M. (1976). Simple mathematical models with very complicated dynamics. *Nature 261*, 459–467.

McSharry, P. E. and L. A. Smith (1999). Better nonlinear models from noisy data: Attractors with maximum likelihood. *Physical Review Letters 83*, 4285.

McSharry, P. E. and L. A. Smith (2004). Consistent nonlinear dynamics: identifying model inadequacy. *Physica D 192*, 1–22.

Moniz, L. J., J. D. Nichols, and J. M. Nichols (2007). Mapping the information landscape: Discerning peaks and valleys for ecological monitoring. *Journal of Biological Physics 33*(3), 171–181.

Nichols, J. M., L. J. Moniz, J. D. Nichols, L. M. Pecora, and E. Cooch (2005). Assessing spatial coupling in complex population dynamics using mutual prediciton and continuity statistics. *Theoretical Population Biology 67*, 9–21.

Pecora, L. M., L. Moniz, J. Nichols, and T. L. Carroll (2007). A unified approach to attractor reconstruction. *Chaos 17*(1).

Sauer, T., J. A. Yorke, and M. Casdagli (1991). Embedology. *Journal of Statistical Physics 65*(3 - 4), 579–616.

Schaefer, M. B. (1954). Some aspects of the dynamics of populations important to the management of the commercial marine fisheries. *Bulletin, Inter-American Tropical Tuna Commission 1*, 25–56.

Takens, F. (1981). Detecting strange attractors in turbulence. In D. A. Rand and L.-S. Young (Eds.), *Dynamical Systems and Turbulence*, Volume 898 of *Lecture Notes in Mathematics*, pp. 366–381. Springer.

Tenenbaum, J. B., V. de Silva, and J. C. Langford (2000). A global geometric framework for nonlinear dimensionality reduction. *Science 290*, 2319.

van Huffel, S. and J. Vandewalle (1991). *The Total Least Squares Problem: Computational Aspects and Analysis*. Philadelphia: SIAM.