Ann. Inst. Statist. Math.
Vol. 54, No. 1, 224–233 (2002)

ESTIMATING INVARIANT PROBABILITY DENSITIES FOR DYNAMICAL SYSTEMS*

DEVIN KILMINSTER[1], DAVID ALLINGHAM[1] AND ALISTAIR MEES[1,2]

[1] *Centre for Applied Dynamics and Optimization, The University of Western Australia,*
35 Stirling Highway, Nedlands, WA 6009, Australia
[2] *Department of Systems Engineering and Engineering Management,*
The Chinese University of Hong Kong, Shatin, NT, Hong Kong, China

(Received February 8, 2001; revised July 23, 2001)

Abstract. Knowing a probability density (ideally, an invariant density) for the trajectories of a dynamical system allows many significant estimates to be made, from the well-known dynamical invariants such as Lyapunov exponents and mutual information to conditional probabilities which are potentially more suitable for prediction than the single number produced by most predictors. Densities on typical attractors have properties, such as singularity with respect to Lebesgue measure, which make standard density estimators less useful than one would hope. In this paper we present a new method of estimating densities which can smooth in a way that tends to preserve fractal structure down to some level, and that also maintains invariance. We demonstrate with applications to real and artificial data.

Key words and phrases: Nonlinear dynamics, probability density, invariant measure, Radon transform.

1. Introduction

Density estimation in one dimension is an essential part of data analysis, and methods such as histograms and kernel density estimators (see, for example, Silverman (1986)) are well known. In higher dimensions, the problem is much harder, principally due to the requirement for a large amount of data. For dynamical systems, density estimation is potentially extremely useful, but the problems of practical estimation are particularly acute because the usual assumptions used to smooth the data points are not appropriate. In addition there is an "invariance" constraint that should often be applied.

In this paper we show how one can take advantage of the existence of high quality one-dimensional density estimators to construct higher-dimensional estimators with smoothing properties that are more appropriate for dynamics, and how to apply invariance constraints to these estimators. The approach is based on the Radon transform, widely used in tomography, though our application requires different algorithms to the FFT-based methods that are used in medical imaging. Allingham *et al.* (1999, 2001) have described this method and compared it with other methods, but have not shown how to estimate densities that explicitly satisfy an invariance constraint; the present paper will provide little in the way of detail on the estimation method, but will show how to make the estimated densities invariant.

*This research was partially supported by the Australian Research Council.

In Section 2, we discuss the requirements for a useful density estimator for dynamical systems. In Section 3, we describe the Radon transform and show how it can be applied to reconstruct densities. In Subsection 3.1 we discuss the computational aspects of the method, with reference to an example using experimental data in Subsection 3.2. In Section 4, we consider how to enforce invariance, and present an example that demonstrates that invariance can sometimes greatly improve dynamical behaviour.

2. Densities and dynamics

In this paper we are interested in discrete-time dynamical systems with dynamical noise. That is, if the state is $x_t \in \mathbb{R}^n$ then

$$x_{t+1} = f(x_t) + \epsilon_t(x_t) \tag{2.1}$$

where the dynamical noise ϵ_t is an i.i.d. random process, and may be identically zero. In the zero noise case, the system will in general have attractors, which may be fixed points, periodic orbits, or more complicated objects possibly having a fractal structure. In the noisy case, Chan and Tong (1994) show that, under reasonable assumptions, (2.1) will be an ergodic stochastic system. That is, there will exist a definite probability measure representing the long-term limiting distribution of the system's states. Equivalently, the probability measure can be thought of as attaching weights to different parts of phase space corresponding to how much time a "typical" trajectory spends there. Given a sufficiently smooth ϵ_t we may take this distribution to be a probability density.

In this paper we will only consider systems with dynamical noise; however, it is interesting to note that the methods we present could also be applicable to the noiseless case: Although we do not have as strong a guarantee as in the noisy, case, noiseless systems do often possess a "physically" relevant invariant measure representing the long-term behaviour of typical trajectories. More troublesome is the fact that these physically relevant invariant measures are unlikely to be densities. We might, none the less, attempt to estimate them as densities —we could then think of our estimates as being relevant for the system with a small amount of added noise. For the behaviour of dynamical systems in the presence of small noise, see Zeeman (1988), and for discussion of some of the problems of the existence and estimation of measures, see Froyland (1996).

Given a density p of the long-term distribution of states of the system, we can use it to give essentially all the useful information about the system: for example, with either known or embedded states, dynamical invariants such as Lyapunov exponents can be calculated as suitable averages (Froyland *et al.* (1995); Froyland (2001)), while with embedded data we can compute mutual information (Allingham *et al.* (2001)). Froyland *et al.* (1995) and Allingham *et al.* (2001) have discussed some of the uses of densities in dynamics. In the present paper we will concentrate on estimating the density and on applying it for simulation and prediction. For prediction, we must estimate joint densities and from them deduce conditional densities. For example, if y_t is observed and $z_t = (y_t, y_{t-1}, \ldots, y_{t-k})$ is the embedded state, we can estimate the conditional density $p(y_{t+1} \mid z_t)$ as discussed in Section 4. Thus, the estimated density p induces a process that may serve as a model of the original system.

At first, we will consider how to estimate densities from one or more orbits of the system, without concerning ourselves with invariance. We assume that a finite orbit segment x_t, $t = 1, \ldots, T$, is either known for each t, or has been re-created by embedding observed data (Stark (2001)). Later we will show how to ensure the density is *invariant.*

That is, the density p is invariant under the action of the induced process $p(y_{t+1} \mid z_t)$ —this usually ensures that p is in fact the long-term distribution exhibited by the induced process.

The question is now, how can we compute a density from an orbit segment in a space of dimension greater than 1, given that the orbit segment is nothing more than a finite set of points? We must make some extra assumptions which result in smoothing the points in some way, so as to interpolate between the points.

The natural thing to try first is a histogram, possibly one that adapts itself to the data (Fraser and Swinney (1986)). For purposes such as conditional density estimation, this approach appears to require too much data (Allingham *et al.* (2001)). An alternative is to use some sort of smoothing kernel (Silverman (1986)), but unfortunately, as was shown in Allingham *et al.* (2001), these methods suffer in more than one dimension through being unable to alter the directions in which they smooth from place to place. (There is no problem in one dimension as there is no choice about the directions in which to smooth, only the degree.) As such, it is difficult to prevent kernel methods from smoothing excessively across important structures if we are to maintain adequate smoothing in other directions.

3. The Radon transform and density estimation

Kernel estimators, and even histograms, work well in one dimension. Let us assume from now on that we can always make adequate one dimensional estimates where required. We are going to describe a method in which a large number of these one dimensional estimates are used to construct a higher dimensional estimate. If we think of a gray-scale image as being equivalent to a probability density, this approach is already well-established in medical tomography, where images such as X-rays taken from different angles are combined into a single two or three dimensional picture. Each image can be thought of as a projection, in a particular direction, of the density function of bone and tissue. The original density is to be reconstructed from the projections.

The Radon transform (Lim (1990)) describes the action of a projection on a function. Tomographic image reconstruction uses many projections to estimate the original density by inverting the Radon transform.

The Radon transform for a function $p : \mathbb{R}^D \to \mathbb{R}$ is a projection onto a line in the direction of unit vector $\hat{n}$, given by

$$\rho(t, \hat{n}) = \int \delta(t - \hat{n} \cdot x) p(x) dx. \tag{3.1}$$

Here, t is the projection of a point $x \in \mathbb{R}^D$ onto the line, $\hat{n} \cdot x$ is the inner product, and δ is the Dirac delta function. There are efficient FFT methods for computing this linear transform and its inverse: see Lim (1990).

In our application, we project data points in $\mathbb{R}^D$ onto each of many lines, and use a one-dimensional density estimator to reconstruct the density on each line. We then use the inverse Radon transform to deduce the density at points in $\mathbb{R}^D$.

The function we are trying to estimate is a probability density and is therefore non-negative. A simple inverse Radon transform will not preserve non-negativity. Although this is also true in medical imaging, tomographic reconstructions are interpreted by eye and artifacts can be ignored. For present purposes, however, the artifacts make an FFT-based reconstruction essentially useless (Allingham *et al.* (2001)). We must add the constraint $p \geq 0$ explicitly. This will force us to use less efficient methods: specifically, we will use quadratic programming, which takes time bounded by a polynomial in IJ where I and J measure the problem size (as defined in Subsection 3.1) rather than the $O(IJ\log(IJ))$ of FFT methods.

3.1 *Computational considerations*

In two dimensions the Radon transform (3.1) becomes

$$\rho(t,\theta) = \int \delta(t - x\cos\theta - y\sin\theta)p(x,y)dx\,dy. \tag{3.2}$$

We use a finite set θ_j, $j = 1,\ldots,J$, of projection angles. For each angle θ_j we project the data points onto the line making an angle of θ_j with the x-axis, and then compute a one-dimensional kernel density estimate $K(t,\theta_j)$ for that projection, evaluating it at points t_i, $i = 1,\ldots,I$.

Assume we want to evaluate p at a finite set of points (x_k, y_ℓ), possibly points on a grid. Our estimate of the density at each of these points represents the density averaged across a region that the point "owns", such as a grid box; we will call the region the (k,ℓ) *pixel*. Likewise, each point t_i on the projection line at angle θ_j corresponds to an interval that we shall call the (i,j) *interval*. Equation (3.2) discretizes to

$$\rho(t_i,\theta_j) = \sum_{k,l} \Delta(i, x_k\cos\theta_j + y_\ell\sin\theta_j)p(x_k,y_\ell) \tag{3.3}$$

where $\Delta(i, z_{jk\ell})$ is the fraction of the area of the (k,ℓ) pixel that projects onto the (i,j) interval. (As a result, $\Delta = 0$ for most choices of i, j, k and ℓ.)

Make $\rho(t_i,\theta_j)$, $i = 1,\ldots,I$, $j = 1,\ldots,J$, into a vector $R \in \mathbb{R}^m$ where $m = I \times J$ in any suitable fashion, and likewise make $p(x_k,y_\ell)$ into a vector $P \in \mathbb{R}^n$. Then (3.3) becomes a matrix equation

$$R = AP \tag{3.4}$$

where the elements of the $m\times n$ matrix A correspond to Δ from (3.3). After solving (3.4) for P, we can unpack the vector P into $p(x_k,y_\ell)$ and so obtain the desired estimate. Note that A is sparse because most (k,ℓ) pixels have zero area projection onto a given (i,j) interval: that is, $\Delta(i,z_{jk\ell}) = 0$ for most values of i, j, k and ℓ.

Since $m \neq n$ in general, (3.4) will not have a unique solution. It is usual to use enough projections so that $m > n$, and to find an approximate solution to the resulting over-determined set of equations. We have found that minimizing in either the ℓ_1 or ℓ_2 norms works well; the first gives rise to a linear program and the second to a quadratic program, for both of which there exist efficient solution packages. In either case, it is simple to add the constraint $P \geq 0$ to the solver. For the ℓ_2 case we solve

$$\text{minimize } \epsilon^T W^2 \epsilon \text{ over } P, \epsilon \text{ subject to } P \geq 0 \text{ and } \epsilon = AP - R.$$

Here we have introduced a weight matrix $W = \operatorname{diag}(w_i)$ which, as described in Allingham *et al.* (2001), places less weight on large R_i values, which are likely to have greater errors. A possible choice for w_i is $1/\max\{R_i,\delta\}$ for some small positive δ. The quadratic program is strictly convex and has a unique solution. It can be solved using either simplex-based or interior point methods, though it is important to take advantage of the sparsity of A. For the examples presented in this paper, we used the optimisation package MOSEK, (see EKA Consulting (accessed 2001) and Andersen and Andersen (2000)) to solve the resulting quadratic programs. Even with sparse matrix methods, the polynomial growth of computation time makes it difficult to get fine resolution in higher dimensions because of the large number of pixels required. Resolutions on the order of 100×100 in 2 dimensions, with a few hundred projections, are readily achievable, taking about one day of computation time on a 350MHz PC. For resolutions of only 50×50 in 2 dimensions, as is used for the examples in Subsection 4.1, the computation time is cut to about 1 hour. In higher dimensions we must either assume the density in certain pixels is zero (for example, if we believe the attractor is confined to some small region), or reduce the resolution.

3.2 *Application to neural recordings*

Part of a time series of data from a recording of a voltage-clamped squid giant axon (Mees *et al.* (1992)) is shown in Fig. 1, together with a two dimensional embedding of the full data set, which contains 400 points. This data has been modified somewhat. Firstly, the initial 100 points have been discarded, as they appear to represent a transient. Secondly the time-series has been "unfolded" according to the transformation:

$$v_t \mapsto \begin{cases} v_t, & v_{t-1} < -130 \\ v_t + 70, & \text{otherwise.} \end{cases}$$

This unfolding has the effect of changing the system from one with dynamics appearing to be "slightly more than one-dimensional" to a system with truly one-dimensional dynamics. It is clear that this transformation is invertible. Finally, the data was transformed again to remove intervals clearly having zero-density, and rescaled to lie in $[-.5, .5]$. Reference to Fig. 1 suggests a system with one-dimensional chaotic dynamics, and some degree of noise.

Using the ℓ_2 norm and the method of Subsection 3.1, we estimated a density, which is shown in Fig. 2. The estimate is $p(v_{t+1}, v_t)$ where v is the transformed axon voltage.

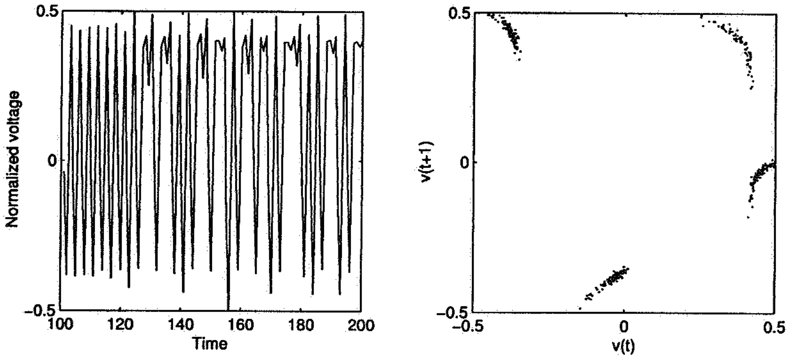

Fig. 1. A segment of the time series from the squid giant axon experiment, and a lag-1 embedding of the entire 400 point data set. The data have been transformed in order to produce a system with one-dimensional dynamics, and to compress regions where the density would be zero.

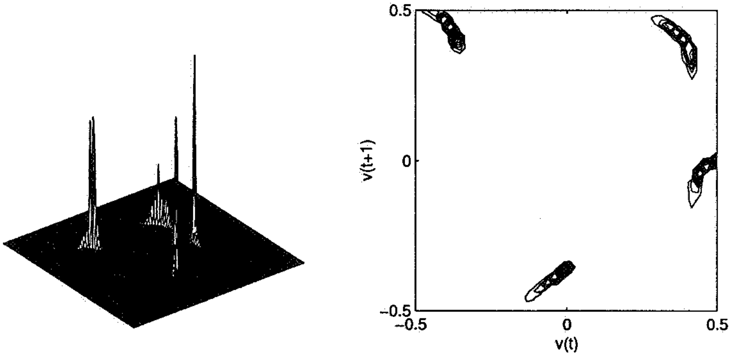

Fig. 2. A mesh plot and the corresponding contour plot of the density estimate of the embedded transformed squid data. The differential smoothing along and across the attractor is strongly apparent in the mesh plot.

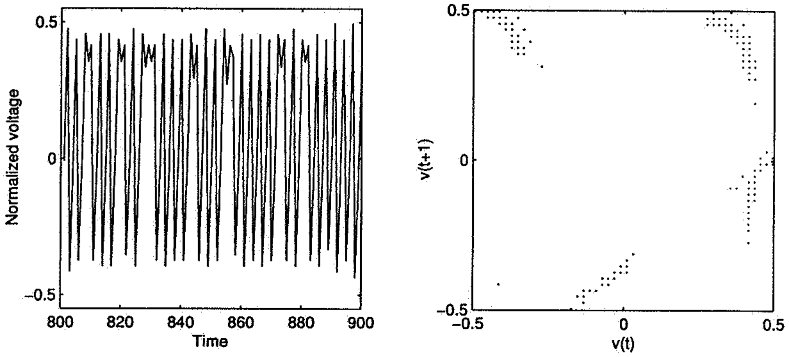

Fig. 3. A time series simulated from the two dimensional squid density estimate, by converting it to a conditional probability estimator and sampling recursively from that. The embedding makes apparent the effect of the discretization of the estimator, but the general features of the embedded data have been preserved (see Fig. 1). Note the outlier points, caused by small nonzero artifacts in our estimate. Also note that there are some points in the simulation just outside the boundaries of this figure and hence not shown here: the smoothing causes the density estimator to spread slightly too far. This is also apparent in the contour map in Fig. 2.

From this, we compute

$$p(v_{t+1} \mid v_t) = \frac{p(v_{t+1}, v_t)}{p(v_t)}, \tag{3.5}$$

so we are assuming the dynamics is one-dimensional, with the voltage as the state. We assume the original system is stationary, so the conditional density estimator can now be used to generate simulations of the system by starting with any v_0 and, whenever v_t is known, drawing v_{t+1} from the conditional distribution implied by (3.5).

A typical simulation and its embedding are shown in Fig. 3. Within the limitations of the discretized grid on which we estimated p, the dynamics has been captured relatively faithfully.

Allingham *et al.* (2001) have described other applications of the density estimate for this data, for example in computing Lyapunov exponents.

4. Invariance

So far, we have not required that the probability density, $p(x_{t-1}, x_t)$, that we estimate be invariant under the action of the induced system,

$$\text{Prob}(x_t \mid x_{t-1}, x_{t-2}, \ldots) = p(x_t \mid x_{t-1}) \quad \left(= \frac{p(x_{t-1}, x_t)}{\int p(x_{t-1}, z) dz} \right). \tag{4.1}$$

A necessary condition for this invariance is that its projection onto each axis (or onto each coordinate plane) be the same. That is, for all x,

$$\int p(x, y) dy = \int p(y, x) dy. \tag{4.2}$$

Again assuming stationarity, this must be true since every time the system enters the state x (from some state y with frequency $p(y, x)$) it must also leave the state x (to some state z with frequency $p(x, z)$). In fact (Kilminster (2002)), it turns out that this condition is also sufficient to guarantee that p is invariant under the action of the induced system associated with it. It is usually true, then, that the long-term behaviour of (4.1) will match p.

We expect invariance to be helpful since if p is an estimate of the density of our embedded data, we should expect it to be close to the true density of "embedded states" of the original system. If p is also invariant then the distribution of states generated by (4.1) will match p, and hence be close to that of the original system. Thus invariance should usually ensure a model that has behaviour similar to that of the original system.

To enforce the invariance condition for a density estimator of a two dimensional embedded system, we therefore need only add the constraint

$$\Pi_1 P = \Pi_2 P \tag{4.3}$$

where Π_1 and Π_2 sum the discretized densities onto the x and y axes respectively, with the summation expressed in terms of the packed P vector. Since this constraint is linear it is easily incorporated into the quadratic or linear program to be solved.

Our final problem definition is therefore

$$\text{minimize } \epsilon^T W^2 \epsilon \text{ over } P, \epsilon \text{ subject to } P \geq 0, \quad \epsilon = AP - R \text{ and } \Pi_1 P = \Pi_2 P.$$

The additional constraint typically makes little difference to the solution time.

4.1 *The need for invariance*

Ensuring invariance is less important than might be expected in some cases since usually the density is derived from embedded data, and for most data points (x_{t-1}, x_t), there is a corresponding point (x_t, x_{t+1}). For this reason, the estimated density often comes very close to satisfying (4.2). To demonstrate that invariance *is* sometimes important, we look at a one-dimensional map forced by noise. The orbit is given by equation (2.1), where the noise realizations ϵ_t are i.i.d. $N(0, 0.15^2)$ random variates. The map f is given by

$$f(x) = 2x \exp(-x^{64}) + 0.15 \exp(-(5x)^2)$$

and is shown in Fig. 4. Observe the slight asymmetry for x around 0. We remark that this map has a stable periodic orbit of period 5, but unstable fixed points of all periods.

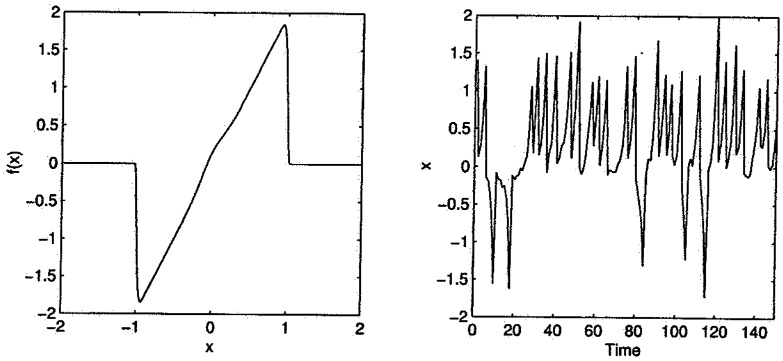

Fig. 4. The map used to demonstrate the need for invariance has a slight asymmetry around $x = 0$. We show a typical time series generated by this map according to equation (2.1). The asymmetry of the map is reflected in the time series, with x spending more time on the positive axis than on the negative axis.

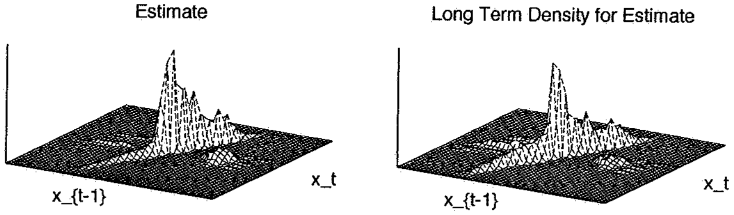

Fig. 5. Density estimate without invariance, and density of states from the induced system. There is a marked difference between the two.

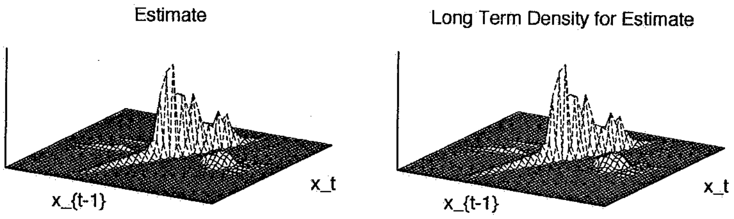

Fig. 6. Density estimate with invariance, and density of states from the induced system. The two densities are indistinguishable.

The level of the noise is sufficient to prevent the orbit from settling into the basin of the stable orbit, as seen in the time series that accompanies the map.

In Figs. 5 and 6 we show the density estimates with and without invariance for $p(x_{t-1}, x_t)$ calculated for the 150 points of the time series shown in Fig. 4, and also the long-term density of states, (x_{t-1}, x_t) generated from the system each estimate induces. In Fig. 5 we made no attempt to enforce invariance, while in Fig. 6, we added the constraint (4.3) to the quadratic program in order to generate an invariant estimate. While by eye, the two estimates appear very similar, the behaviour of the models induced by each varies markedly.

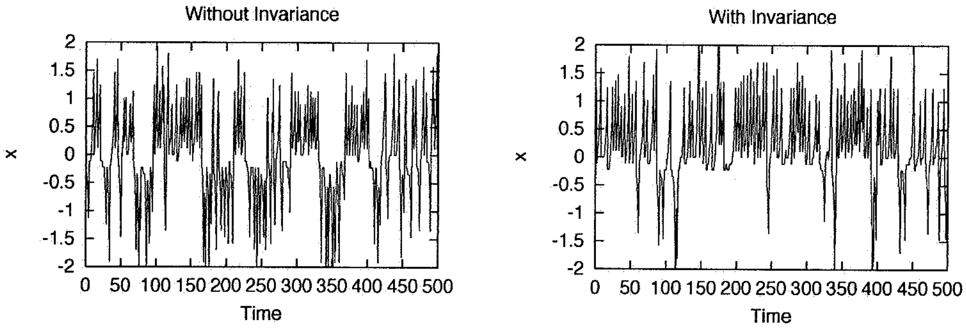

Fig. 7. Free-run simulations both with and without invariance. The invariant estimator is more faithful to the dynamics.

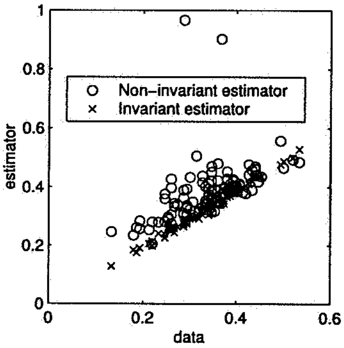

Fig. 8. The "up-down" statistic shows a strong bias in simulations from the ordinary density estimator, but none in the estimator with the invariance property.

We can see the difference in behaviour in Fig. 7 in which we have plotted free-run simulations for the estimate both with and without invariance. A fairly characteristic property of our original system (as can be seen in Fig. 4) is the greater proportion of "upward spikes" to "downward spikes". We see that the invariant estimate respects this property, whereas without invariance the ratio has become closer to 1 : 1. The invariant estimator is more faithful to the true dynamics.

We can measure the performance of models with respect to matching the proportion of "upward" and "downward" spikes by considering a simple "up-down" statistic that measures the proportion of time that $x_t < 0$ (i.e., it measures something to do with the proportion of down spikes). Figure 8 shows experimental results, plotting this statistic for each of 100 different simulation runs, again each of 150 points, against the value obtained directly from the data used to generate each density. The invariant estimator (crosses) shows no bias, and reflects the data, but the estimator without invariance (circles) shows a strong bias, with some gross errors. (The true proportion is approximately equal to 0.338.)

5. Concluding remarks

We have introduced a new density estimation method based on a weighted Radon transform, but with added non-negativity and invariance constraints that are necessary for estimating probability densities in dynamical systems, but which increase the computational complexity beyond the standard FFT-based use of Radon transforms.

The application and example in the present paper showed only the use of the estimator in simulation, by converting a joint density to a conditional density, but the estimator can also be used to find many of the standard dynamical invariants.

References

Allingham, D., Kilminster, D. and Mees, A. I. (1999). Estimating probability distributions using tomographic imaging techniques, *Proceedings of International Symposium on Nonlinear Theory and Its Applications* (*NOLTA*'99), 379–382, NOLTA, Waikoloa, Hawaii.

Allingham, D., Kilminster, D. and Mees, A. I. (2001). Estimation of probability densities for dynamical systems, Tech. Report, Centre for Applied Dynamics and Optimization, The University of Western Australia, Perth.

Andersen, E. D. and Andersen, K. D. (2000). The MOSEK interior point optimizer for linear programming: An implementation of the homogeneous algorithm, *High Performance Optimization* (eds. H. Frenk, K. Roos, T. Terlaky and S. Zhang), 197–232, Kluwer, Dordrecht.

Chan, K. S. and Tong, H. (1994). A note on noisy chaos, *J. Roy. Statist. Soc. Ser. B*, **56**(2), 301–311.

EKA Consulting (accessed 2001). World Wide Web page, `http://www.mosek.com/`.

Fraser, A. M. and Swinney, H. L. (1986). Independent coordinates for strange attractors from mutual information, *Phys. Rev. A*, **33**(2), 1134–1140.

Froyland, G. (1996). Estimating physical invariant measures and space averages of dynamical systems indicators, PhD Thesis, Department of Mathematics, The University of Western Australia, Perth.

Froyland, G. (2001). Extracting dynamical behaviour via Markov models, *Nonlinear Dynamics and Statistics* (ed. A. I. Mees), 281–321, Birkhäuser, Boston.

Froyland, G., Judd, K., Mees, A. I., Murao, K. and Watson, D. (1995). Constructing invariant measures from data, *Internat. J. Bifur. Chaos Appl. Sci. Engrg.*, **5**(4), 1181–1192.

Kilminster, D. (2002). Modelling dynamical systems via behaviour criteria, PhD Thesis, Department of Mathematics and Statistics, The University of Western Australia, Perth.

Lim, J. S. (1990). *Two-Dimensional Signal and Image Processing*, Prentice Hall Signal Processing Series, Prentice Hall, New Jersey.

Mees, A. I., Aihara, K., Adachi, M., Judd, K., Ikeguchi, T. and Matsumoto, G. (1992). Deterministic prediction and chaos in squid axon response, *Phys. Lett. A*, **169**, 41–45.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

Stark, J. (2001). Delay reconstruction: Dynamics versus statistics, *Nonlinear Dynamics and Statistics* (ed. A. I. Mees), Birkhäuser, Boston.

Zeeman, E. C. (1988). Stability of dynamical systems, *Nonlinearity*, **1**(1), 115–155.