Consistent Estimation of a Dynamical Map

Andrew Nobel
Department of Statistics
University of North Carolina, Chapel Hill

September, 1999

Abstract

Estimation of a non-linear map F governing the evolution of an observed dynamical system is considered for two specific models. In the first model, F is successively applied to a fixed initial vector in the absence of noise, so that the the observed states of the system constitute a trajectory of F. In the second, dynamical noise model, the system is perturbed by independent noise between each application of F. Estimates of F are proposed for each model, and are shown to be consistent under general conditions, in particular, when the measured states of the system are bounded and their common (or asymptotic) distribution is comparable to a known reference measure. No assumptions are made regarding mixing rates, or the regularity of the function F.

Appears in the collection *Nonlinear Dynamics and Statistics*, A.I. Mees editor, Birkhauser, Boston, 2001.

1 Introduction

The advent of modern computing and the recent interest in chaos have focussed increasing attention on deterministic systems that exhibit random behavior. While there is no universally accepted definition of chaos, phenomena termed 'chaotic' have generally been studied in the context of dynamical systems. A dynamical system is a mathematical model of a physical system. Its state is commonly described by a family of differential equations. The solution of the equations describes the time evolution of the dynamical system starting from any initial condition.

In many situations, the data arising from measurements of a physical system are obtained at discrete, equally spaced instants of time. In some cases direct measurement of the state of the system is possible. More commonly, one makes periodic measurements of some scalar function ϕ of the state of the system. It is known from Takens's Embedding Theorem [44, 2, 42] that, for generic functions ϕ and suitable integers d, one may study the dynamics of the system by means of time delay vectors that consist of d successive scalar measurements. Thus, when it is in a steady state and no noise is present, observations of the physical system can be modelled by iteration of a fixed, nonlinear map $F : \mathbb{R}^d \to \mathbb{R}^d$.

The statistical analysis of measurements from dynamical systems is complicated by the fact that such measurements can exhibit very long range dependence. Even

when a system is perturbed by independent noise, its measurements may fail to satisfy standard mixing assumptions, and existing statistical estimation theory may not apply to the analysis of these measurements. Nevertheless, it is often reasonable to assume that measurements of a dynamical system are stationary and ergodic, or when this assumption fails, that the system converges asymptotically to an ergodic steady state.

1.1 Overview

The subject of this paper is estimation of a non-linear map F whose iteration governs the behavior of an observed dynamical system. Two dynamical models are considered. In the first, F is successively applied to a fixed initial vector in the absence of noise, so that the the measured states of the system constitute a trajectory of F. In the second, often referred to as a dynamical noise model, the system is perturbed by independent noise between each application of F. Histogram estimates of F are proposed and analyzed for each model. The estimates are shown to be consistent under very general conditions, in particular, when the measurements are bounded and their common (or asymptotic) distribution is comparable to a known reference measure. No assumptions are made regarding mixing rates, or the regularity of the function F, which need not be continuous. The primary goal of the paper is to rigorously establish the existence of consistent estimates for F under very general conditons. Although in specific cases the proposed estimates can be implemented on a computer, no attempt has been made to assess their empirical performance.

The problem of estimating an iterated map governing a dynamical system has previously been considered by a number of authors, working in several fields, and the bibliography here makes no claim to completeness. Most often F is estimated with the ultimate goal of prediction, estimating Lyapunov exponents, or estimating the dimension of an attractor. Representative work and additional references can be found in the papers of Farmer and Sidorowich [14], Casdagli [7, 8], Kostelich and Yorke [24], Nychka *et al.* [36], Lu and Smith [27], and the book of Tong [45]. See also the surveys by Eckmann and Ruelle [13], Jensen [23], and Isham [22]. In most of this work iterates of the map are perturbed by observational or dynamical noise. Bosq and Guégan [6] study kernel estimates of uniformly mixing transformations in the absence of noise. Lalley [26] describes a general means of reconstructing the orbit of a smooth diffeomorphism F, acting on a hyperbolic attractor, when the iterates of F are corrupted by observation noise.

In the references above it is commonly assumed that the map under study is continuous or differentiable. Hofbauer and Keller [21], Mayer [29], and Denker and Keller [12] established central limit theorems for U-statistics and smooth functionals of noiseless dynamical systems that are generated by piecewise-monotone maps. More irregular transformations are of interest to ergodic theorists and may arise, for example, when one considers the Poincaré return map of a smooth flow to a low-dimensional set $A \subseteq \mathbb{R}^d$.

1.2 Outline

The two noise models studied in the paper are defined in the next section. A regression estimation scheme for ergodic processes $\{(X_i, Y_i)\}$ proposed by Nobel and Adams [33] is described in Section 3. Theorem A shows that the proposed scheme is consistent

when the squared error loss of the regression function can be estimated, and the distribution of the covariates X_i is comparable to a known reference measure. It is shown in Theorem 1 that the latter assumption can be weakened when suitable density estimates are available.

In Section 4 the regression estimates of Theorem A are applied to the problem of estimating a μ-preserving ergodic map $F : \mathbb{R}^d \to \mathbb{R}^d$ from one of its trajectories. It is shown in Proposition 1 that consistent estimation of F is possible if μ is comparable to a known reference measure with compact support.

The regression estimates of Theorems A and 1 are applied in Section 5 to the problem of estimating the map F governing a dynamical noise process. Density estimates of Györfi and Masry [17] are briefly discussed, and it is shown in Theorem 2 that one may estimate F provided only that the observations are stationary, ergodic, and bounded, and that the distribution of the noise has a density.

2 Two Models for Dynamical Data

Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be a fixed nonlinear map governing the one-step evolution of a dynamical system under study. Let $\mathcal{B}$ denote the Borel subsets of $\mathbb{R}^d$. Here and in what follows it is assumed that F is measurable, so that $F^{-1}A \in \mathcal{B}$ for every $A \in \mathcal{B}$.

Model I: No Noise. In the simplest model of a dynamical system, the evolution of the system is determined by successive application of F, in the absence of observational or dynamical noise. Starting from an initial vector $x \in \mathbb{R}^d$ successive observations of the system are described by a the trajectory

$$x, Fx, F^2x, \ldots \in \mathbb{R}^d \tag{1}$$

Here F^i denotes the i-fold composition of F with itself. In this case the system evolves in a purely deterministic fashion. From (complete) knowledge of F and any single observation one can, in principle, reconstruct every subsequent measurement. In analyzing the model (1) it is typically assumed that the map F is measure preserving and ergodic.

Definition: A Borel-measurable map $F : \mathbb{R}^d \to \mathbb{R}^d$ is said to preserve a probability measure μ on $(\mathbb{R}^d, \mathcal{B})$ if

$$\mu(F^{-1}A) = \mu(A) \quad \text{for every} \quad A \in \mathcal{B}\,, \tag{2}$$

and is said to be ergodic if

$$F^{-1}A = A \quad \text{implies} \quad \mu(A) = 0 \text{ or } 1,$$

or, equivalently, if

$$\frac{1}{n}\sum_{i=1}^{n} \mu(A \cap F^{-i}B) \to \mu(A)\mu(B) \quad \text{for every} \quad A, B \in \mathcal{B}. \tag{3}$$

The books of Petersen [37] and Walters [] give comprehensive introductions to ergodic theory.

Model II: Dynamical Noise. When the evolution of a system is mediated by independent noise, its state may be represented by a simple non-linear autoregression. The resulting dynamical noise model has broad application in the analysis of chaotic data.

Definition: A system is said to obey a dynamical noise model if its state evolves according to the recursion

$$X_{i+1} = F(X_i) + Z_{i+1} \quad i \geq 0 \tag{4}$$

where $X_0 \in \mathbb{R}^d$ is the (random) initial state of the system, $F : \mathbb{R}^d \to \mathbb{R}^d$ is a fixed map, and $Z_1, Z_2, \ldots \in \mathbb{R}^d$ are i.i.d., independent of X_0, and such that $EZ_i = 0$. Thus F represents the deterministic component of the dynamics, and Z_i represents a random perturbation, or noise, that influences subsequent measurements through the action of F.

Definition: A stationary process $W_1, W_2, \ldots \in \mathbb{R}^k$ is said to be ergodic if for every $l \geq 1$ and every pair of Borel sets $A, B \subseteq \mathbb{R}^{ld}$,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{P}\{W_1^l \in A, W_{i+1}^{i+l} \in B\} \to \mathbb{P}\{W_1^l \in A\}\mathbb{P}\{W_1^l \in B\} \tag{5}$$

as n tends to infinity, where $W_i^j = (W_i, \ldots, W_j)$ for $i \leq j$.

For example, if $F : \mathbb{R}^d \to \mathbb{R}^d$ preserves a measure μ and is ergodic, and if X_0 is a random vector with distribution μ, then (2) and (3) imply that the process $X_0, FX_0, F^2X_0, \ldots$ is stationary and ergodic. By contrast, due to the perturbations Z_i, ergodicity of the dynamical noise process (4) does not require that F preserve a measure on $\mathbb{R}^d$ or that F be ergodic. (To take a trivial example, let $F : \mathbb{R} \to \mathbb{R}$ be identically zero, and let $\{Z_i\}$ be any real valued i.i.d. sequence.)

The dynamical noise process (4) is a discrete time Markov chain. Results such as those in Nummelin [35], and Meyn and Tweedie [30], provide general conditions under which such chains are ergodic, or converge to an ergodic steady state. Dynamical noise models are also a special case of random dynamical systems. Many of the theoretical properties of such systems, *e.g.* topological dynamics, the existence of invariant measures, and Lyapunov exponents, have been studied. See Kifer [25] or Arnold [3] for more details.

3 Regression Estimation from Ergodic Processes

Let (X, Y) be a jointly distributed pair with $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$. The regression function of Y on X is the conditional expectation $g(x) = E(Y|X = x)$. The regression function g minimizes $E(Y - g'(X))^2$ over all functions g' of X, and is therefore an optimal predictor of Y given X under the squared error loss. The problem considered below is how to estimate g from a stationary ergodic sequence $\{(X_i, Y_i) : i \geq 1\}$ of random pairs, each distributed as (X, Y).

A regression scheme is a sequence of functions $g_n : \mathbb{R}^d \times (\mathbb{R}^d \times \mathbb{R})^n \to \mathbb{R}$ for $n \geq 1$. Given observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, the corresponding function $\hat{g}_n(x) = g_n(x; X_1, Y_1, \ldots, X_n, Y_n)$ is taken to be an estimate of g. The scheme $\{g_n\}$

is said to be strongly L_2-consistent for the process if $\int (\hat{g}_n(x) - g(x))^2 d\mu(x) \to 0$ with probability one as $n \to \infty$, where μ is the distribution of X.

The existence of regression estimates weakly consistent for any i.i.d. process was first established by Stone [43]. Beginning with the papers of Roussas [40, 41] and Rosenblatt [38], there has been a great deal of work on regression estimation from stationary, weakly dependent processes satisfying α, β, ρ, and related mixing conditions. The monographs of Györfi, Härdle, Sarda and Vieu [18], Rosenblatt [39], and Bosq [5] give an overview of kernel and histogram regression estimation from weakly dependent processes. Masry [28] studies local polynomial regression in the same setting. There is also a substantial body of work on regression estimation from stationary processes exhibiting long range dependence. For an overview of these results and additional references, see Cheng and Robinson [9], Hidalgo [19], and the book of Beran [4].

Delecroix [10], Györfi *et al.* [18], and later Delecroix and Rosa [11] established the consistency of kernel estimates of the k-step autoregression function for two-sided processes exhibiting a very mild mixing condition. Yakowitz [47, 48] studied autoregression from real-valued Markov chains under mild regularity conditions. Yakowitz *et al.* [49] and Morvai *et al.* [34] propose regression estimates that are consistent for any ergodic process whose marginal regression function is suitably regular.

In spite of these positive results, it has recently been shown by Adams [1], Nobel [32], and Yakowitz and Heyde [50] that no regression estimation scheme is weakly consistent for every stationary ergodic process. The upshot of these negative results is that restrictions must be placed on the family of possible observations in order to establish the consistency of a density or regression scheme under study. Most work to date places assumptions on either the dependence (mixing) structure of the observations, on the regularity of the unknown regression function, or both. For the models studied here it suffices to place assumptions only on the one-dimensional distribution of the observations.

3.1 Consistent Regression with Error Estimates

Fix a nested sequence $\pi_0, \pi_1, \ldots$ of finite partitions of $\mathbb{R}^d$ such that $\pi_0 = \{\mathbb{R}^d\}$, and such that for each vector $x \in \mathbb{R}^d$,

$$\lim_{k \to \infty} \operatorname{diam}(\pi_k[x]) = 0\,. \tag{6}$$

Here $\pi_k[x]$ is the unique cell of π_k containing x, and $\operatorname{diam}(A) = \sup_{u,v \in A} ||u - v||$ denotes the maximum Euclidean distance between any two points in A. The partitions π_l can be obtained by dividing $[-l, l)^d$ into cubes of side-length 2^{-l}, and letting the complement of $[-l, l)^d$ comprise a single cell.

Let $\{(X_i, Y_i) : i \geq 1\}$ be a stationary ergodic sequence of random pairs with $X_i \in \mathbb{R}^d$, $Y_i \in \mathbb{R}$, and regression function $g(x) = E(Y|X = x)$. Consider the measurements

$$(X_1, Y_1), \ldots, (X_n, Y_n) \tag{7}$$

The proposed regression estimates are histograms, obtained by dividing the covariates X_i into disjoint cells, and then averaging the corresponding response variables Y_i within each cell.

Candidate Histograms: For $k = 1, 2, \ldots$ define the ordinary histogram based on the partition π_k,

$$\phi_{k,n}(x) = \frac{\sum_{i=1}^{n} Y_i I\{X_i \in \pi_k[x]\}}{\sum_{i=1}^{n} I\{X_i \in \pi_k[x]\}}\,, \tag{8}$$

where, if a cell $\pi_k[x]$ contains no vector X_i then $\phi_{k,n}(x) = 0$. Let

$$\Delta_{k,n} = \left(\frac{1}{n}\sum_{i=1}^{n} |\phi_{k,n}(X_i) - Y_i|^2\right)^{1/2} \tag{9}$$

be the empirical loss of $\phi_{k,n}$. An estimate $\hat{g}_n$ of g is chosen from among the candidates $\{\phi_{k,n} : k \geq 1\}$ by selecting a suitable partition index k_n, based on the available data (7), and on prior information concerning the process from which the data is obtained. Recall that the support of a distribution μ on $\mathbb{R}^d$ is the smallest closed set $\Lambda \subseteq \mathbb{R}^d$ such that $\mu(\Lambda) = 1$.

Definition: Let μ_0 be a reference probability measure defined on the Borel subsets of $\mathbb{R}^d$, and let $M \geq 1$ be a constant. Let $\mathcal{D}(\mu_0, M)$ be the family of all probability measures μ on $(\mathbb{R}^d, \mathcal{B})$ such that $\mu \equiv \mu_0$ and

$$\int \frac{d\mu_0}{d\mu}\, d\mu_0 \;\leq\; M^2\,. \tag{10}$$

The condition $\mu \equiv \mu_0$ means that $\mu(A) = 0$ if and only if $\mu_0(A) = 0$. This implies that the densities $d\mu/d\mu_0$ and $d\mu_0/d\mu$ exist, and that μ and μ_0 have the same support.

Example: Let λ denote Lebesgue measure on $\mathbb{R}^d$, let Λ be a compact subset of $\mathbb{R}^d$ with $\lambda(\Lambda) > 0$, and let $\mu_0(A) = \lambda(A)/\lambda(\Lambda)$ be normalized Lebesgue measure on Λ. The family $\mathcal{D}(\mu_0, M)$ contains every probability measure μ having a density $f = d\mu/d\lambda$ such that $\{x : f(x) > 0\} = \Lambda$ and $\int_\Lambda (1/f(x)dx \leq M^2$. Alternatively, μ_0 might be Hausdorff measure on some low-dimensional subset of $\mathbb{R}^d$, such as a smooth manifold.

Definition: Let $\Gamma_n : (\mathbb{R}^d \times \mathbb{R})^n \to \mathbb{R}$, $n \geq 1$, be a sequence of measurable functions. In defining $\hat{g}_n$ the number

$$\hat{\Gamma}_n = \Gamma_n(X_1, Y_1, \ldots, X_n, Y_n) \tag{11}$$

is used as an estimate of the error $E(Y - g(X))^2$.

Description of Estimates: Suppose that a reference measure μ_0, constant M, and functions Γ_n have been specified in advance of the data. Let $\epsilon_1, \epsilon_2, \cdots$ be any sequence of positive numbers tending monotonically to zero. Define k_n to be the largest integer $k \geq 1$ such that

$$\int |\phi_{l,n} - \phi_{j,n}| d\mu_0 \;\leq\; 2M(\Delta_{j,n}^2 - \hat{\Gamma}_n)_+^{1/2} \;+\; 2(1+M)\epsilon_j \;\text{ for }\; 1 \leq j \leq l \leq k \tag{12}$$

and define the histogram

$$\hat{g}_n(x) \;=\; \phi_{k_n,n}(x)\,. \tag{13}$$

Note that each of the quantities appearing in (12) is either specified in advance of the data, or may be evaluated once the data is obtained. Thus $\hat{g}_n$ is well defined. In general, the partition index k_n will not increase monotonically with the sample size n, nor need it grow at any prespecified rate. The consistency of $\{\hat{g}_n\}$ is established in the following theorem, due to Nobel and Adams [33].

Theorem A *Let $\{(X_i, Y_i) : i \geq 1\}$ be any stationary ergodic sequence such that Y is bounded and X has distribution $\mu \in \mathcal{D}(\mu_0, M)$. Let $g(x) = E(Y|X = x)$ be the regression function of Y on X. If $\hat{\Gamma}_n \to E(Y - g(X))^2$ with probability one, then*

$$\int (\hat{g}_n(x) - g(x))^2 d\mu(x) \to 0$$

with probability one as $n \to \infty$.

Remark: Theorem A shows that for ergodic sequences $\{(X_i, Y_i)\}$ with suitable one-dimensional distributions estimating the regression function of Y on X is no more difficult than estimating its squared error loss.

3.2 Use of Auxiliary Density Estimates

It is assumed in Theorem A that the common distribution μ of the X_i is comparable to a known reference measure μ_0 and, in particular, that μ and μ_0 have the same support. These assumptions can be dropped if L_1-consistent estimates of the density $f = d\mu/d\mu_0$ are available. A μ_0-density estimation scheme is a sequence of non-negative functions

$$f_n : \mathbb{R}^d \times \mathbb{R}^{nd} \to \mathbb{R} \quad n \geq 1$$

such that $\int f_n(x; x_1, \ldots, x_n)\, d\mu_0(x) = 1$ for every n and every choice of $x_1, \ldots, x_n \in \mathbb{R}^d$. Using $\{f_n\}$ one may define alternative regression estimates $\tilde{g}_n$ as follows. Given $(X_1, Y_1), \ldots, (X_n, Y_n)$ let

$$\hat{f}_n(x) \;=\; f_n(x; X_1, \ldots, X_n)$$

act as an estimate of $f = d\mu/d\mu_0$. Let s_n be the greatest integer $s \geq 1$ such that

$$\int |\phi_{l,n} - \phi_{j,n}| \cdot \hat{f}_n \, d\mu_0 \;\leq\; 2(\Delta_{j,n}^2 - \hat{\Gamma}_n)_+^{1/2} \;+\; 5\epsilon_j \quad \text{for} \quad 1 \leq j \leq l \leq s\,, \tag{14}$$

where $\phi_{l,n}$, $\Delta_{j,n}$, and $\hat{\Gamma}_n$ are defined by equations (8), (9), and (11), as before. Define the regression estimate

$$\tilde{g}_n(x) \;=\; \phi_{s_n,n}(x)\,. \tag{15}$$

using the new partition index s_n. A routine modification of the proof of Theorem 1 in Nobel and Adams [33] yields the following result.

Theorem 1 *Let $\{(X_i, Y_i) : i \geq 1\}$ be a stationary ergodic sequence such that Y_i is bounded and the distribution μ of X_i has density f with respect to μ_0. Let $g(x) = E(Y|X = x)$ be the regression function of Y on X. If*

$$\int |\hat{f}_n - f| d\mu_0 \to 0 \quad \textit{and} \quad \hat{\Gamma}_n \to E(Y - g(X))^2$$

with probability one, then

$$\int (\tilde{g}_n(x) - g(x))^2 d\mu(x) \to 0$$

with probability one as $n \to \infty$.

4 Ergodic Systems without Noise

The estimates of Theorem A may be readily applied to the problem of estimating an ergodic map $F : \mathbb{R}^d \to \mathbb{R}^d$ in the noiseless model (1) where the available measurements form a trajectory $x, Fx, F^2x, \ldots$ of F. Estimation in this setting is complicated by the fact that the measurements are purely deterministic and will, in general, fail to satisfy standard mixing assumptions. In particular, the sample averages $n^{-1}\sum_{i=0}^{n-1} h(F^i x)$ of a bounded function $h : \mathbb{R}^d \to \mathbb{R}$ may converge arbitrarily slowly to the limit $\int h d\mu$ guaranteed by the ergodic theorem.

If the measurements are grouped into pairs (x, Fx), (Fx, F^2x), $(F^2x, F^3x), \ldots$ then each pair is a point on the graph of F. When F is continuous and $d = 1$, connecting neighboring points with straight lines will give pointwise consistent estimates of F on the support of μ. Similar piecewise linear estimates may be used in higher dimensions. Of interest here is the case when F may be highly irregular so that some sort of local averaging is necessary to obtain good estimates. The following consequence of Theorem A appears in Nobel and Adams [33]. Let $||\cdot||$ denote the Euclidean norm on $\mathbb{R}^d$.

Proposition 1 *Let* μ_0 *be a reference measure on* $\mathbb{R}^d$ *with compact support and let* $M > 1$. *There exist functions* $F_n : \mathbb{R}^d \times \mathbb{R}^{nd} \to \mathbb{R}^d$, $n \geq 1$, *such that for every* $\mu \in \mathcal{D}(\mu_0, M)$, *every* μ*-preserving ergodic transformation* $F : \mathbb{R}^d \to \mathbb{R}^d$, *and* μ*-almost every* $x \in \mathbb{R}^d$, *the estimates* $\hat{F}_n(u) = F_n(u; x, Fx, \ldots, F^{n-1}x)$ *are such that* $\int ||\hat{F}_n - F||^2 \, d\mu \to 0$ *as* $n \to \infty$.

Proof: The conditions of the theorem ensure that the support Λ of μ is bounded. As F preserves μ it follows that $\mu(\Lambda \setminus F^{-1}\Lambda) = 0$. Therefore for μ-almost every $x \in \mathbb{R}^d$ (equivalently μ-almost every $x \in \Lambda$) the trajectory of F starting at x is contained in Λ.

Fix $1 \leq l \leq d$, and let $g_l : \mathbb{R}^d \to \mathbb{R}$ be the l'th component function of F. If X_0 is a random vector with distribution μ, then the sequence $X_0, FX_0, F^2X_0, \ldots$ is stationary and ergodic. Moreover the sequence $\{(X_i, Y_i) = (X_i, g_l(X_i)) : i \geq 0\}$, obtained by pairing X_i with the l'th component of X_{i+1}, is stationary and ergodic, with regression function g_l and associated error $E(g_l(X) - Y)^2 = 0$. As Λ is bounded the response variables Y_i are bounded with probability one. Let $\hat{g}_{l,n-1}$ be the estimate obtained by applying the regression scheme of Section 3.1 to $\{(X_i, Y_i) : 0 \leq i \leq n-2\}$ with error estimate $\Gamma_n \equiv 0$. (The reduction in sample size is due to the fact that from $X_0, \ldots, X_{n-1}$ one obtains only $n-1$ pairs $(X_0, Y_0), \ldots, (X_{n-2}, Y_{n-2})$.) It follows from Theorem A that $\int (\hat{g}_{l,n-1} - g_l)^2 d\mu \to 0$ with probability one.

Each sample sequence of $\{(X_i, Y_i)\}$ is in 1:1 correspondence with a sample sequence of $\{X_i\}$, and each sample sequence of the latter process is a trajectory of F determined by the initial value $X_0 = x$. It follows that for μ-almost every $x \in \mathbb{R}^d$ the estimates $\hat{g}_{l,n-1}$ derived from $x, Fx, \ldots, F^{n-1}x$ are such that $\int (\hat{g}_{l,n-1} - g_l)^2 d\mu \to 0$. If

the component estimates are combined by setting $\tilde{F}_n(u) = (\hat{g}_{1,n-1}(u), \ldots, \hat{g}_{d,n-1}(u))$ then for μ-almost every $x \in \mathbb{R}^d$, $\int ||\tilde{F}_n - F||^2 d\mu \to 0$.

A slightly simpler estimate may be obtained as follows. Each component estimate $\hat{g}_{l,n-1}$ is a histogram, based on a partition index $k_{l,n-1}$ that is chosen based on the available data. Define $k_n^* = \min\{k_{l,n-1} : 1 \le l \le d\}$. Evidently $k_n^* \to \infty$, and the component estimates defined using partition $\pi_{k_n^*}$ are consistent as before. Combining these estimates one obtains the vector valued histogram

$$\hat{F}_n(u) \;=\; \frac{\sum_{i=0}^{n-1} F^{i+1}x \cdot I\{F^i x \in \pi_{k_n^*}[u]\}}{\sum_{i=0}^{n-1} I\{F^i x \in \pi_{k_n^*}[u]\}}$$

As each component function of $\hat{F}_n$ is consistent, $\int ||\hat{F}_n - F||^2 d\mu \to 0$ for μ-almost every $x \in \mathbb{R}^d$. ♣

Remarks: Bosq and Guégan [6] consider estimation of continuous maps F under uniform mixing assumptions using kernel estimators. No conditions are placed here on the regularity of F or on its mixing properties. When consistent estimates of the invariant density $f = d\mu/d\mu_0$ are available, the assumption that $\mu \in \mathcal{D}(\mu_0, M)$ can be dropped (see Section 3.2 above). In the special case when $d = 1$ and μ_0 equals Lebesgue measure, the density estimates of Nobel, Morvai, and Kulkarni [34] are consistent almost every trajectory of an ergodic μ-preserving transformation, provided that $f = d\mu/d\lambda$ satisfies a variation condition.

The family of transformations consistently estimated by a fixed sequence $\{F_n\}$ may be quite large. To illustrate this, let $d = 1$ and let μ_0 be Lebesgue measure on $\Lambda = [0, 1]$. Setting $M = 1$, the following corollary of Proposition 1 is immediate.

Corollary 1 *There exist functions $F_n : [0,1]^{n+1} \to [0,1]$, $n \ge 1$, such that for every ergodic Lebesgue measure preserving transformation $F : [0,1] \to [0,1]$, for almost every $x \in [0,1]$ the estimates $\hat{F}_n(\cdot) = F_n(\cdot; x, Fx, \ldots, F^{n-1}x)$ are such that $\int_0^1 |\hat{F}_n - F|^2 du \to 0$ as $n \to \infty$.*

Remark: Among the ergodic Lebesgue measure preserving transformations of $[0, 1]$ there is an uncountable subfamily with the property that no two transformations in the subfamily are isomorphic. The functions F_n of the corollary give consistent estimates of each such transformation from almost every one of its trajectories.

5 Ergodic Systems with Dynamical Noise

Recall that a system is said to obey a dynamical noise model if its state evolves according to the recursion

$$X_{i+1} = F(X_i) + Z_{i+1} \tag{16}$$

where $X_0 \in \mathbb{R}^d$ is the initial state of the system, $F : \mathbb{R}^d \to \mathbb{R}^d$ is a fixed map, and

$$Z_1, Z_2, \ldots \in \mathbb{R}^d \text{ are i.i.d., independent of } X_0, \text{ with } EZ_i = 0. \tag{17}$$

The regression estimates described in Section 3 are applied below to the problem of estimating the map F from the observations $X_0, X_1, \ldots$. These and several related results are described in the next two subsections.

5.1 Arbitrary Noise Distributions

Consider first the case where $\{X_i\}$ is stationary and ergodic, and the distribution of X_i is comparable to a known reference measure.

Proposition 1 *Let μ_0 be a reference measure on $\mathbb{R}^d$ with compact support and let $M > 1$. There exist functions $F_n : \mathbb{R}^d \times \mathbb{R}^{nd} \to \mathbb{R}^d$, $n \geq 1$, such that for every stationary ergodic process $\{X_i\}$ obeying the noise model (16)–(17) and such that X_i has distribution $\mu \in \mathcal{D}(\mu_0, M)$ the estimates $\hat{F}_n(u) = F_n(u : X_0, \ldots, X_{n-1})$ are such that*

$$\int (\hat{F}_n - F)^2 d\mu \to 0 \tag{18}$$

with probability one as $n \to \infty$.

Remark: Proposition 1 may be established using Theorem A and Lemma 1 below. Its proof is similar to that of Theorem 2 in the next section and is therefore omitted. The estimates $\hat{F}_n$ of the proposition are vector valued histograms of the form

$$\hat{F}_n(u) \;=\; \frac{\sum_{i=0}^{n-1} X_i I\{X_i \in \pi_{l_n}[u]\}}{\sum_{i=0}^{n-1} I\{X_i \in \pi_{l_n}[u]\}} \,,$$

where the index l_n depends only on μ_0, M, and $X_0, \ldots, X_{n-1}$.

Non-stationary measurements: It may happen that a system evolving according to (16)–(17) fails to be stationary, but nevertheless converges to an ergodic steady state. This happens, for example, if the system has an attracting stationary distribution that differs from the distribution of X_0. The conclusions of Theorem A and Lemma 1 rely only on the asymptotic behavior of the measurements (X_i, Y_i), and do not require that the measurements be strictly stationary. In order to establish (18) it is sufficient that for every bounded measurable function $h : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$,

$$\frac{1}{n} \sum_{i=0}^{n-1} h(X_i, X_{i+1}) \;\to\; Eh(X, F(X) + Z) \tag{19}$$

with probability one as $n \to \infty$, where X has distribution μ, Z is distributed as Z_1, and X and Z are independent.

5.2 Absolutely Continuous Noise

Suppose now that $\{X_i\}$ obeys the dynamical noise model (16)–(17) and is stationary. Let μ be the common distribution of the observed X_i and let ν be the common distribution of the perturbations Z_i. The relation (16) implies that

$$\mu \;=\; (\mu \circ F^{-1}) * \nu \tag{20}$$

where $*$ denotes convolution. Let λ denote Lebesgue measure on $\mathbb{R}^d$. It follows from (20) that if the noise distribution ν has a density with respect to λ (*i.e.* is absolutely continuous) then the same is true of the distribution μ of X_i. In other words, if $d\nu/d\lambda$ exists, so does $d\mu/d\lambda$. The assumption of absolutely continuous noise has several useful consequences, which are considered below.

Delecroix [10] considers estimation of the auto-regression $R(x) = E(X_1|X_0 = x)$ from ergodic processes $\{X_i : -\infty < i < \infty\}$ such that, for each $k \geq 1$, the conditional distribution of X_k given $X_0, X_{-1}, \ldots$ has a continuous density. His estimates are of the form

$$\hat{R}_n(x) \;=\; \frac{\sum_{i=0}^{n-1} X_{i+1} K((x - X_i)/h_n}{\sum_{i=0}^{n-1} K((x - X_i)/h_n}$$

where $K(\cdot)$ is a bounded, Lipschitz continuous kernel with compact support, and h_n is a sequence of bandwidths tending to zero at an appropriate rate that does not depend on the observed X_i. Györfi *et al.* ([18] Theorem 3.5.1) state a version of Delecroix's result under slightly weaker hypotheses. Verifying the conditions of their theorem in the case of the dynamical noise model is straightforward, and yields the following result.

Theorem B *Let $\{X_i : i \geq 0\}$ be a bounded, stationary ergodic sequence obeying the dynamical noise model (16)–(17) and such that X_i has distribution μ. If F is continuous and the distribution ν of the noise has a continuous density, then $f = d\mu/d\lambda$ exists and for any compact set $\Lambda \subset \mathbb{R}^d$ such that $f(x) > 0$ for $x \in \Lambda$,*

$$\sup_{x \in \Lambda} |\hat{R}_n(x) - F(x)| \to 0$$

with probability one. In particular, $\int (\hat{R}_n - F)^2 \, d\mu \to 0$ with probability one.

Györfi [16] and later Györfi and Masry [17] consider recursive kernel density estimates of the form

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=0}^{n-1} \frac{1}{h_i^d} K\left(\frac{x - X_i}{h_i}\right) \tag{21}$$

where $K : \mathbb{R}^d \to \mathbb{R}$ is a non-negative kernel such that $\int K dx = 1$, and with the property that for each $x \in \mathbb{R}^d$ the function $K(cx)$ is non-increasing in $0 < c < \infty$. The bandwidth sequence h_i is proportional to i^{-b} for any $0 < b < 1/d$. It is shown in Theorem 3.2 of Györfi and Masry [17] that the estimates $\hat{f}_n$ are L_1-consistent for every ergodic process $\{X_i : -\infty < i < \infty\}$ such that the conditional distribution of X_1 given $X_0, X_{-1}, \ldots$ is absolutely continuous with probability one. The next theorem follows directly from this result.

Theorem C *Let $\{X_i : i \geq 0\}$ be a stationary ergodic sequence obeying the dynamical noise model (16)–(17), and let μ be the distribution of X_i. If the distribution ν of Z_i has a density then $f = d\mu/d\lambda$ exists and $\int |\hat{f}_n - f| dx \to 0$ with probability one.*

In conjunction with the density estimates of Theorem C and the error estimates of Lemma 1 below, Theorem 1 may be used to obtain L_2 consistent estimates of F under weaker conditions than those of Theorem B. In particular, one may drop the assumption that F and the density of ν are continuous.

Theorem 2 *There exist functions $F_n : \mathbb{R}^d \times \mathbb{R}^{nd} \to \mathbb{R}^d$, $n \geq 1$, such that for every stationary ergodic sequence $\{X_i : i \geq 0\}$ obeying the dynamical noise model (16)–(17) the estimates $\hat{F}_n(u) = F_n(u : X_0, \ldots, X_{n-1})$ are such that*

$$\int (\hat{F}_n - F)^2 \, d\mu \to 0,$$

provided only that X_i is bounded and the distribution of Z_i has a density. Here μ is the distribution of X_i.

The proof of Theorem 2 is given in the next section. The following proposition, showing that one may estimate the variance of the dynamical noise, is obtained as a by-product of the proof, and may be of independent interest.

Proposition 2 *There exist functions $G_n : \mathbb{R}^{nd} \to \mathbb{R}$ such that for every bounded sequence $\{X_i : i \geq 0\}$ obeying the dynamical noise model (16)–(17) the estimates $\hat{G}_n = G_n(X_0, \ldots, X_{n-1}) \to E||Z_1||^2$ with probability one if $\{X_i\}$ is ergodic, or if the weaker condition (19) holds.*

5.3 Proof of Theorem 2

Let $X_0, X_1, \ldots \in \mathbb{R}^d$ be a stationary ergodic sequence of random vectors and let $\xi_1, \xi_2, \ldots \in \mathbb{R}$ be a bounded, i.i.d. sequence of random variables with $E\xi_i = 0$. The processes $\{X_i\}$ and $\{\xi_i\}$ need not be independent. For example $\{\xi\}$ may be an innovations process generating $\{X_i\}$ as in the dynamical noise model considered above. It is assumed in what follows that

$$\text{for each } \; i \geq 1, \; \xi_{i+1} \text{ is independent of } X_0, \ldots X_i \tag{22}$$

Let $g : \mathbb{R}^d \to \mathbb{R}$ be any measurable function such that $g(X_i)$ is bounded with probability one, and define a process $\{(X_i, Y_i)\}$ with

$$(X_i, Y_i) \;=\; (X_i,\, g(X_i) + \xi_{i+1}) \;\in\; \mathbb{R}^d \times \mathbb{R} \tag{23}$$

Under assumption (22) each pair (X_i, Y_i) has regression function g and associated error $E(Y_i - g(X_i))^2 = E\xi_i^2$.

Our immediate goal is to consistently estimate $E\xi_1^2$ using the observations (23). To this end define, for $n \geq 2$,

$$r_n \;=\; \max\left\{ r \geq 0 : |\pi_r| \leq \frac{n^{1/3}}{\log n} \right\}$$

Given observations $(X_0, Y_0), \ldots, (X_{n-1}, Y_{n-1})$ from (23) define the histogram

$$\theta_n(x) \;=\; \frac{\sum_{i=0}^{n-1} Y_i I\{X_i \in \pi_{r_n}[x]\}}{\sum_{i=0}^{n-1} I\{X_i \in \pi_{r_n}[x]\}} \tag{24}$$

and the error estimate

$$\hat{\Gamma}_n \;=\; \frac{1}{n} \sum_{i=0}^{n-1} \left(\theta_n(X_i) - Y_i\right)^2 \; . \tag{25}$$

Note that $\hat{\Gamma}_n$ depends only on $(X_0, Y_0), \ldots, (X_{n-1}, Y_{n-1})$. The proof of the following Lemma is given after the proof of Theorem 2 below.

Lemma 1 *For every stationary ergodic sequence $X_0, X_1, \ldots \in \mathbb{R}^d$, every function $g : \mathbb{R}^d \to \mathbb{R}$ such that $g(X_i)$ is bounded with probability one, and every bounded i.i.d. sequence $\xi_1, \xi_2, \ldots \in \mathbb{R}$ such that $E\xi_i = 0$ and (22) holds, the estimates $\hat{\Gamma}_n$ converge to $E\xi_1^2$ with probability one.*

Proof of Theorem 2 Let $\{X_i : i \geq 0\}$ be any bounded stationary ergodic sequence obeying the dynamical noise model (16)–(17) and such that the distribution of Z_i has a density. Fix $l \in \{1, \ldots, d\}$. Let g_l be the l'th component function of F and let $\xi_{l,i}$ be the l'th component of the random vector Z_i. Consider the measurements $X_0, \ldots, X_{n-1}$. Pairing X_i with the l'th component of X_{i+1} gives a new sequence of the form

$$(X_i, Y_i) \;=\; (X_i,\, g_l(X_i) + \xi_{l,i+1}) \;\; 0 \leq i \leq n-2 \tag{26}$$

Let $\hat{\Gamma}_{l,n-1}$ be the estimate (25) of $E\xi_{l,1}^2$ based on the measurements (26). As each X_j is bounded with probability one, the same is true of $g_l(X_i)$ and $\xi_{l,i}$. Moreover $EZ_i = 0$ implies $E\xi_{l,i} = 0$, and (16)–(17) imply (22). It then follows from Lemma 1 that

$$\hat{\Gamma}_{l,n-1} \;\rightarrow\; E\xi_{l,1}^2 \tag{27}$$

with probability one as $n \to \infty$. As noted in Section 5.2 the assumption that Z_i is absolutely continuous implies that the distribution μ of X_i has a density, f say, with respect to Lebesgue measure. Let (21) be the recursive kernel estimate of f based on $X_0, \ldots, X_{n-1}$. It follows from Theorem C that

$$\int |\hat{f}_n - f|\, dx \to 0 \tag{28}$$

with probability one as $n \to \infty$.

Let $\tilde{g}_{l,n-1}$ be the histogram estimate of g_l, defined as in (14)–(15), employing $\hat{f}_n$ and $\hat{\Gamma}_{l,n-1}$. In conjunction with (27) and (28), Theorem 1 implies that $\int (\tilde{g}_{l,n-1} - g_l)^2\, d\mu \to 0$ with probability one as $n \to \infty$. If the component estimates are combined by setting $\tilde{F}_n(u) = (\hat{g}_{1,n-1}(u), \ldots, \hat{g}_{d,n-1}(u))$, then $\int ||\tilde{F}_n - F||^2 d\mu \to 0$ with probability one. Alternatively, this same property is shared by the vector-valued histogram

$$\hat{F}_n(u) \;=\; \frac{\sum_{i=0}^{n-2} X_{i+1}\, I\{X_i \in \pi_{s_n^*}[u]\}}{\sum_{i=0}^{n-2} I\{X_i \in \pi_{s_n^*}[u]\}}$$

where $s_n^* = \min\{s_{l,n-1} : 1 \leq l \leq d\}$ and $s_{l,n-1}$ is the partition index selected for the component estimate $\tilde{g}_{l,n-1}$ according to (14). ♣

Proof of Lemma 1: Fix $L < \infty$ such that $|g(X_i)| \leq L$ and $|\xi_i| \leq L$ for each $i \geq 1$ with probability one. Note that $\theta_n(x) = U_n(x) + V_n(x)$, where

$$U_n(x) = \frac{\sum_{i=0}^{n-1} \xi_{i+1} I\{X_i \in \pi_{r_n}[x]\}}{\sum_{i=0}^{n-1} I\{X_i \in \pi_{r_n}[x]\}} \quad \text{and} \quad V_n(x) = \frac{\sum_{i=0}^{n-1} g(X_i) I\{X_i \in \pi_{r_n}[x]\}}{\sum_{i=0}^{n-1} I\{X_i \in \pi_{r_n}[x]\}}.$$

Expanding the square in the definition of $\hat{\Gamma}_n$ and collecting terms, one finds that

$$\left| \hat{\Gamma}_n - \frac{1}{n} \sum_{i=1}^{n} \xi_i^2 \right| \;\leq\; |\Theta_{1,n}| \,+\, |\Theta_{2,n}| \,+\, |\Theta_{3,n}|,$$

where

$$\Theta_{1,n} \;=\; \frac{1}{n} \sum_{i=0}^{n-1} (V_n(X_i) - g(X_i))^2 \qquad \Theta_{2,n} \;=\; \frac{2}{n} \sum_{i=0}^{n-1} \xi_{i+1} (V_n(X_i) - g(X_i))$$

$$\Theta_{3,n} \;=\; \frac{1}{n}\sum_{i=0}^{n-1} U_n(X_i)\,[U_n(X_i) + 2(V_n(X_i) - g(X_i)) + 2\xi_{i+1}]$$

The law of large numbers ensures that $n^{-1}\sum_{i=1}^{n}\xi_i^2 \to E\xi_1^2$ with probability one, and it is therefore enough to show that $\Theta_{j,n} \to 0$ with probability one for $j = 1, 2, 3$.

Let μ be the common distribution of the X_i. The conditional expectation of $g(X_i)$ given π_r is

$$(g \circ \pi_r)(x) = \frac{1}{\mu(\pi[x])}\int_{\pi[x]} g(u)d\mu(u)$$

if $\mu(\pi[x]) > 0$ and is zero otherwise. If

$$\hat{g}_{r,n}(x) = \frac{\sum_{i=0}^{n-1} g(X_i)I\{X_i \in \pi_r[x]\}}{\sum_{i=0}^{n-1} I\{X_i \in \pi_r[x]\}}.$$

is obtained by averaging the values of $g(X_i)$ within the cells of π_r, then it follows readily from the ergodic theorem that

$$\max\{\,|\,g_{n,r}(x) - (g \circ \pi_r)(x)\,|\,\,:\, x \in \mathbb{R}^d\} \;\to\; 0 \tag{29}$$

with probability one for each $r \geq 0$.

Consider the term $\Theta_{1,n}$. For any sequence of numbers $c_1, \ldots, c_n$ the sum $\sum_{i=1}^{n}(c_i - c)^2$ is minimized by setting $c = n^{-1}\sum_{i=1}^{n} c_i$. As the partitions π_r are nested, this fact implies that

$$\Theta_{1,n} \;\leq\; \frac{1}{n}\sum_{i=1}^{n}(g_{r,n}(X_i) - g(X_i))^2 \;\;\text{ when }\;\; r_n \geq r.$$

As $r_n \to \infty$, the relation (29) implies that

$$0 \;\leq\; \limsup_{n\to\infty}\Theta_{1,n} \;\leq\; E((g \circ \pi_r)(X) - g(X))^2$$

with probability one for each $r \geq 1$. It follows from the martingale convergence theorem and the diameter condition (6) that the right hand side above tends to zero as $r \to \infty$, and therefore $\lim_n \Theta_{1,n} = 0$ with probability one. By the Cauchy-Schwartz inequality,

$$\begin{aligned}
|\Theta_{2,n}|^2 \;&\leq\; 4\left[\frac{1}{n}\sum_{i=0}^{n-1}(V_n(X_i) - g(X_i))^2\right]\;\left[\frac{1}{n}\sum_{i=0}^{n-1}\xi_i^2\right] \\
&\leq\; \frac{4L^2}{n}\sum_{i=0}^{n-1}(V_n(X_i) - g(X_i))^2 \;=\; 4L^2\Theta_{1,n}.
\end{aligned}$$

Thus $|\Theta_{2,n}| \to 0$ with probability one as $n \to \infty$.

In order to analyze $\Theta_{3,n}$, partition the cells of π_{r_n} according to their (random) occupation counts. Let

$$\mathcal{A}_n = \left\{A \in \pi_{r_n} : \sum_{i=1}^{n} I\{X_i \in A\} \geq n^{2/3}\right\}$$

and let $\mathcal{B}_n = \pi_{r_n} \setminus \mathcal{A}_n$. Then as $|U_n(X_i) + 2(V_n(X_i) - g(X_i)) + 2\xi_i| \leq 6L$, one obtains the bound

$$\begin{aligned} |\Theta_{3,n}| &\leq \frac{6L}{n} \sum_{i=1}^{n} |U_n(X_i)| \\ &\leq 6L \left(\max_{x \in \cup \mathcal{A}_n} |U_n(x)| \right) + \frac{6L}{n} \sum_{i=0}^{n-1} |U_n(X_i)| \cdot I\{X_i \in \cup \mathcal{B}_n\}. \end{aligned} \tag{30}$$

Here $\cup \mathcal{A}_n = \cup_{A \in \mathcal{A}_n} A$ and $\cup \mathcal{B}_n$ is defined similarly. Since $|U_n(X_i)| \leq L$ the second term in (30) is at most

$$\begin{aligned} \frac{6L^2}{n} \sum_{i=0}^{n-1} I\{X_i \in \cup \mathcal{B}_n\} &= \frac{6L^2}{n} \sum_{A \in \mathcal{B}_n} \sum_{i=0}^{n-1} I\{X_i \in A\} \\ &\leq \frac{6L^2}{n} |\pi_{r_n}| n^{2/3} \leq \frac{6L^2}{\log n}. \end{aligned} \tag{31}$$

Fix constants $\delta_1 = 1$ and $\delta_n = 1/\log n$ for $n \geq 2$, and consider the probability that the first term in (30) exceeds δ_n.

$$\begin{aligned} \mathbb{P}\left\{ \max_{x \in \cup \mathcal{A}_n} |U_n(x)| > \delta_n \right\} &= \mathbb{P}\left\{ \max_{A \in \mathcal{A}_n} \left| \frac{\sum_{i=0}^{n-1} \xi_{i+1} I\{X_i \in A\}}{\sum_{i=0}^{n-1} I\{X_i \in A\}} \right| > \delta_n \right\} \\ &\leq \mathbb{P}\left\{ \max_{A \in \mathcal{A}_n} \left| \frac{\sum_{i=0}^{n-1} \xi_{i+1} I\{X_i \in A\}}{n^{2/3}} \right| > \delta_n \right\} \\ &\leq \mathbb{P}\left\{ \max_{A \in \pi_{r_n}} \left| \sum_{i=0}^{n-1} \xi_{i+1} I\{X_i \in A\} \right| > \delta_n n^{2/3} \right\} \\ &\leq \sum_{A \in \pi_{r_n}} \mathbb{P}\left\{ \left| \sum_{i=0}^{n-1} \xi_{i+1} I\{X_i \in A\} \right| > \delta_n n^{2/3} \right\}, \end{aligned}$$

where the last inequality follows from the union bound. For each $A \in \pi_{r_n}$ the condition (22) ensures that $\xi_1 I\{X_0 \in A\}, \ldots, \xi_n I\{X_{n-1} \in A\}$ form a bounded martingale difference sequence. By an application of Hoeffding's inequality for martingale difference sequences ([20, 15]) the last sum above is at most

$$\sum_{A \in \pi_{r_n}} \exp\left\{ -\frac{\delta_n^2 n^{1/3}}{L^2} \right\} = |\pi_{r_n}| \cdot \exp\left\{ -\frac{n^{1/3}}{L^2 \log^2 n} \right\} \leq n^{1/3} \exp\left\{ -\frac{n^{1/3}}{L^2 \log^2 n} \right\}$$

It follows that

$$\sum_{n=1}^{\infty} \mathbb{P}\left\{ \max_{x \in \cup \mathcal{A}_n} |U_n(x)| > \delta_n \right\} < \infty$$

and as $\delta_n \to 0$ the Borel Cantelli Lemma implies that $\max_{x \in \cup \mathcal{A}_n} |U_n(x)| \to 0$ with probability one. This and (30) show that $\Theta_{3,n} \to 0$ with probability one. ♣

Acknowledgement:

This work was supported in part by NSF Grant DMS-9501926 and Grant DMS-9971964.

References

[1] Adams, T.M. (1997). Families of ergodic processes without consistent density or regression estimates. Preprint.

[2] Aeyels, D. (1981). Generic observability of differential systems. *SIAM J. Control and Optimization*, 19:595–603.

[3] Arnold, L. (1995). Random dynamical systems. In *Dynamical Systems*, R.Johnson, editor, Springer Lecture Notes in Mathematics 1609.

[4] Beran, J. (1994) *Statistics for Long-Memory Processes*. Chapman and Hall, New York.

[5] Bosq, D. (1998). *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction*, 2nd edition. Springer Lecture Notes in Statistics, v.110.

[6] Bosq, D. and Guégan, D. (1995). Nonparametric estimation of the chaotic function and the invariant measure of a dynamical system. *Stat. and Prob. Let.*, 25:201–212.

[7] Casdagli, M. (1989). Nonlinear prediction of chaotic time series. *Physica D*, 35:335–356.

[8] Casdagli, M. (1992). Chaos and deterministic *versus* stochastic non-linear modeling. *J. R. Stat. Soc. B*, 54:303–328.

[9] Cheng, B.C. and Robinson, P.M. (1991). Density estimation in strongly dependent non-linear time series. *Stat. Sinica*, 1:335–359.

[10] Delecroix, M. (1987) *Sur l'estimation et la prévision non-paramétrique des processus ergodiques*. Ph.D. Thesis, University of Lille Flandres Artois, Lille, France.

[11] Delecroix, M. and Rosa, A.C. (1996). Nonparametric estimation of a regression function and its derivatives under an ergodic hypothesis. *Nonparam. Statistics*, 6:367–382.

[12] Denker, M., and Keller, G. (1986). Rigorous statistical procedures for data from dynamical systems. *J. Stat. Phys.*, 44:67–93.

[13] Eckmann, J.-P., and Ruelle, D. (1985). Ergodic theory of chaos and strange attractors. *Rev. Mod. Phys.*, 57(3):617–656.

[14] Farmer, J.D. and Sidorowich, J.J. (1987). Predicting chaotic time series. *Phys. Rev. Let.*, 59:845–848.

[15] Grimmet, G.R. and Stirzaker, D.R. (1992) *Probability and Random Processes*. Oxford University Press, New York.

[16] Györfi, L. (1981). Strongly consistent density estimate from ergodic sample. *J. Multivariate Analysis*, 11:81–84.

[17] Györfi, L. and Masry, E. (1990). The L_1 and L_2 strong consistency of recursive kernel density estimation from dependent samples. *IEEE Trans. Inform. Theory*, 36:531–539.

[18] Györfi, L., Härdle, W., Sarda, P., and Vieu, P. (1989). *Nonparametric Curve Estimation from Time Series*. Springer-Verlag, Berlin.

[19] Hidalgo, J. (1997). Non-parametric estimation with strongly dependent time multivariate time series. *J. Time Series Anal.*, 18(2):95–122.

[20] Hoeffding, W. (1963) Probability inequalities for sums of bounded random variables. *J. American Stat. Assoc.*, 58:13–30.

[21] Hofbauer, F., and Keller, G. (1982). Ergodic properties of invariant measures for piecewise monotonic functions. *Mathematische Zeitschrift*, 180:119–140.

[22] Isham, V. (1993) Statistical aspects of chaos: a review. In *Networks and Chaos – Statistical and Probabalistic Aspects*, Barndorff-Nielsen, O.E., Jensen, J.L., and Kendall, W.S. eds, Chapman and Hall, London.

[23] Jensen, J.L. (1993) Chaotic dynamical systems with a view torwards statistics: a review. In *Networks and Chaos – Statistical and Probabalistic Aspects*, Barndorff-Nielsen, O.E., Jensen, J.L., and Kendall, W.S. eds, Chapman and Hall, London.

[24] Kostelich, E.J. and Yorke, J.A. (1990). Noise reduction: finding the simplest dynamical system consistent with the data. *Physica D*, 41:183-196.

[25] Kifer, Y. (1986). *Ergodic Theory of Random Transformations*. Birkhäuser, Boston.

[26] Lalley, S.P. (1999). Beneath the noise, chaos. Preprint.

[27] Lu, Z.-Q., and Smith, R. L. (1997). Estimating local Lyapunov exponents. *Fields Institute Communications*, 11:135–151.

[28] Masry, E. (1996). Multivariate local polynomial regression for time series: uniform strong consistency and rates. *J. Time Sers. Anal.*, 17:571–599.

[29] Mayer, D.H. (1984). Approach to equilibrium for locally expanding maps in R^k. *Communications in Mathematical Physics*, 95:1–15.

[30] Meyn, S.P. and Tweedie, R.L. (1993). *Markov Chains and Stochastic Stability* Springer Verlag, London.

[31] Morvai, G., Kulkarni, S., and Nobel, A.B. (1999). Regression estimation from an individual stationary sequence. To appear in *Statistics*.

[32] Nobel, A.B. (1999). Limits to classification and regression estimation from ergodic processes. *Annals of Statistics* 27:262-273.

[33] Nobel, A.B., and Adams, T.M. (1999). On regression estimation from ergodic samples with additive noise. Submitted for publication.

[34] Nobel, A.B., Morvai, G., and Kulkarni, S. (1998). Density estimation from an individual numerical sequence. *IEEE Trans. Info. Theory* 44:537–541.

[35] Nummelin, E. (1984). *General Irreducible Markov Chains and Non-Negative Operators*. Cambridge Univ. Press.

[36] Nychka, D., Ellner, S., Gallant, A.R., and McCaffrey, D. (1992). Finding chaos in noisy systems. *J. R. Stat. Soc. B*, 54:399-426.

[37] Petersen, K. (1989). *Ergodic Theory*. Cambridge Univ. Press.

[38] Rosenblatt, M. (1970). Density estimates and Markov sequences. In *Nonparametric Techniques in Statistical Inference*, M. Puri editor, Cambridge Univ. Press, London, 199–210.

[39] Rosenblatt, M. (1991). *Stochastic Curve Estimation*. NSF-CBMS Regional Conference Series in Probability and Statistics, Inst. Math. Stat., Hayward, CA..

[40] Roussas, G. (1967). Nonparametric estimation in Markov processes. *Ann. Inst. Statist. Math.*, 21:73–87.

[41] Roussas, G. (1969). Nonparametric estimation of the transition distribution function of a Markov process. *Ann. Math. Stat.*, 40:1386–1400.

[42] Sauer, T., Yorke, J.A., and Casdagli, M. (1991). Embedology. *J. Stat. Phys.*, 65:579–616.

[43] Stone, C. (1977). Consistent nonparametric regression. *Ann. Stat.*, 5:595–620.

[44] Takens, F. (1980) Detecting strange attractors in turbulence In *Dynamical Systems and Turbulence, Warwick 1980*, D.A. Rand and L.-S. Young, editors. Springer Lecture Notes in Mathematics 898.

[45] Tong, H. (1990). *Non-linear Time Series: a Dynamical System Approach*. Oxford University Press.

[46] Walters, P. (1981). *An Introduction to Ergodic Theory*. Springer, New York.

[47] Yakowitz, S. (1989). Nonparametric density and regression estimation for Markov sequences without mixing assumptions. *J. Multivar. Anal.*, 30:124–136.

[48] Yakowitz, S. (1993). Nearest neighbor regression estimation for null-recurrent Markov time series. *Stoc. Proc. Appl.*, 48:311–318.

[49] Yakowitz, S., Györfi, L., Kieffer, J., and Morvai, G. (1997) Strongly-consistent nonparametric estimation of smooth regression functions for stationary ergodic sequences. Under revision, *J. Multivar. Anal.*.

[50] Yakowitz, S. and Heyde, C. (1998). Long range dependency effects with implications for forecasting and queuing inference. Preprint, submitted for publication.