Estimating a Function from Ergodic Samples with Additive Noise

Andrew B. Nobel,* Department of Statistics
University of North Carolina, Chapel Hill, NC 27514
Email: nobel@stat.unc.edu
and
Terrence M. Adams, Department of Math and CS
Rhode Island College, Providence, RI 02908.

February 23, 2001

Abstract

We study the problem of estimating an unknown function from ergodic samples corrupted by additive noise. It is shown that one can consistently recover an unknown measurable function in this setting if the one dimensional distribution of the samples is comparable to a known reference distribution, and the noise is independent of the samples and has known mixing rates. The estimates are applied to deterministic sampling schemes, in which successive samples are obtained by repeatedly applying a fixed map to a given initial vector, and it is then shown how the estimates can be used to reconstruct an ergodic transformation from one of its trajectories.

Appear in *IEEE Transactions on Information Theory*, vol. 47, pp.2895-2902, 2001.

Keywords: regression estimation, ergodic process, deterministic sampling, signal recovery.

*The work of the first author was supported in part by NSF Grants DMS-9501926 and DMS-9971964.

I Introduction

The subject of this paper is the estimation of an unknown irregular function from stationary ergodic samples that are corrupted by additive noise. Let $\mathcal{S}$ be a family of stationary ergodic processes $\{X_i : i \geq 1\}$ with $X_i \in \mathrm{I\!R}^d$, let $\mathcal{N}$ be a family of zero mean dependent processes $\{Z_i : i \geq 1\}$ with $Z_i \in \mathrm{I\!R}$, and assume that the families $\mathcal{S}$ and $\mathcal{N}$ are independent. Suppose that the values of an unknown measurable function $f : \mathrm{I\!R}^d \to \mathrm{I\!R}$ are sampled according to a process $\{X_i\} \in \mathcal{S}$, in the presence of additive noise $\{Z_i\} \in \mathcal{N}$, resulting in observations of the form

$$(X_i, Y_i) = (X_i, f(X_i) + Z_i) \quad i \geq 1. \tag{1}$$

We are interested in estimating f from these noisy samples.

Let (X, Y, Z) denote a random triple, independent of $\{(X_i, Y_i, Z_i)\}$, with the same distribution as (X_1, Y_1, Z_1). Our goal is to define an estimation scheme $\{\hat{f}_n : n \geq 1\}$ with the following property: for each $\{X_i\} \in \mathcal{S}$, each $\{Z_i\} \in \mathcal{N}$, and every measurable function f with $Ef(X)^2 < \infty$, the estimate $\hat{f}_n$ produced from n observations of the form (1) is such that $E(f(X) - \hat{f}_n(X))^2 \to 0$ with probability one. In particular, the consistency of the scheme should not depend on the mixing rates of the sampling processes, or on the regularity of the sampled functions. Our conclusion is that consistent estimation is possible under these conditions if (i) the one-dimensional distribution of each sampling process is comparable to a known reference distribution, and (ii) the correlations of the noise processes tend to zero at a known rate. Since the sampling and noise processes are independent, the function $f(x)$ is a version of the regression function $E(Y|X = x)$, and therefore estimating f is a special case of regression estimation from ergodic processes.

A Statement of Results

A family of sampling processes. Let μ_0 be a fixed, reference probability distribution on the Borel subsets $\mathcal{B}$ of $\mathrm{I\!R}^d$, and let $\alpha \in (0, 1)$. Define $\mathcal{S}(\mu_0, \alpha)$ to be the family of all stationary ergodic processes $X_1, X_2, \ldots \in \mathrm{I\!R}^d$ such that the distribution μ of X_1 satisfies the inequality

$$\alpha \;\leq\; \frac{d\mu}{d\mu_0} \;\leq\; \beta. \tag{2}$$

Here $\beta \in (0, \infty)$ is a constant that may depend on μ. The condition (2) implies in particular that μ and μ_0 have common support, and that the derivative $d\mu_0/d\mu$ is well-defined.

Membership in $\mathcal{S}(\mu_0, \alpha)$ depends only on the distribution of X_1; no conditions are imposed on the joint distributions of $(X_1, \ldots, X_k)$ for $k \geq 2$.

Examples: If the reference distribution μ_0 has a density h_0 with respect to d-dimensional Lebesgue measure then $\mathcal{S}(\mu_0, \alpha)$ contains every ergodic process whose marginal density h is such that $\alpha h_0 \leq h \leq \beta h_0$. for some $\beta < \infty$. For example, let Λ be a subset of $\mathbb{R}^d$, such as the unit cube, with d-dimensional Lebesgue measure $\lambda(\Lambda) \in (0, \infty)$, and such that the boundary $\overline{\Lambda} \setminus \Lambda^o$ of Λ has measure zero. . If $\mu_0(A) = \lambda(A \cap \Lambda)/\lambda(\Lambda)$ is normalized Lebesgue measure on Λ, then $\mathcal{S}(\mu_0, \alpha)$ contains every ergodic process having a bounded marginal density h such that $h(x) \geq \alpha/\lambda(\Lambda)$ for $x \in \Lambda$ and $h(x) = 0$ for $x \in \Lambda^c$. In general μ_0 need not be absolutely continuous. For example, μ_0 might be normalized Hausdorff measure on a bounded, low-dimensional subset of $\mathbb{R}^d$.

A family of noise processes. A real-valued stationary process $\{Z_i\}$ is said to be weakly mixing (in the ergodic theory sense) if for every $k \geq 1$ and every pair A, B of k-dimensional Borel sets,

$$\frac{1}{n} \sum_{i=1}^{n} \left| \mathbb{P}\{Z_1^k \in A, Z_{i+1}^{i+k} \in B\} - \mathbb{P}\{Z_1^k \in A\} \mathbb{P}\{Z_1^k \in B\} \right| \to 0 \quad \text{as} \quad n \to \infty. \tag{3}$$

This mild mixing condition is slightly stronger than ergodicity (which is obtained by removing the absolute values). Fix a function $\kappa : \{0, 1, 2, \ldots\} \to [0, \infty)$ such that $\kappa(s)$ tends monotonically to zero as s tends to infinity, and let $\mathcal{N}(\kappa)$ be the family of all weakly mixing stationary processes $\{Z_i\}$ such that $EZ_i = 0$, $EZ_i^2 < \infty$, and

$$|\operatorname{Cov}(Z_i, Z_j)| \leq EZ^2 \cdot \kappa(|i - j|) \tag{4}$$

for each $i, j \geq 1$. The covariance condition (4) controls the rate at which the noise process "forgets" its past values. The function κ will be called a covariance envelope. Note that, for each envelope κ, the family $\mathcal{N}(\kappa)$ contains every zero mean i.i.d. process with finite variance.

Theorem 1 *Given a reference distribution μ_0, constant α, and covariance envelope κ there exist estimates $\{\hat{f}_n\}$ such that for each sampling process $\{X_i\} \in \mathcal{S}(\mu_0, \alpha)$, each noise process $\{Z_i\} \in \mathcal{N}(\kappa)$ independent of $\{X_i\}$, and every measurable function $f : \mathbb{R}^d \to \mathbb{R}$ with $Ef^2(X) < \infty$, the estimates $\hat{f}_n$ derived from the noisy samples $(X_i, f(X_i) + Z_i)$, $1 \leq i \leq n$*

are such that

$$E(\hat{f}_n(X) - f(X))^2 \to 0$$

with probability one.

The estimates $\hat{f}_n$ of Theorem 1 are histograms constructed from a fixed sequence of nested partitions. As choosing a partition appropriate for a given sample is rather complicated, the estimates are not readily implementable. An explicit description of the estimates is given in Section II below.

The principal assumptions of the theorem concern the one dimensional distribution of the sampling process. No conditions are imposed on the regularity of the sampled function f, or on the mixing rates of the sampling process, and minimal conditions are placed on the moments of $f(X)$ and Z_i. When L_1-consistent estimates of $d\mu/d\mu_0$ are available and f is bounded, the comparability assumption (2) can be dropped: see Section IV for more details. The estimates of Theorem 1 are evidently consistent for any subfamilies $\mathcal{S} \subseteq \mathcal{S}(\mu_0, \alpha)$ and $\mathcal{N} \subseteq \mathcal{N}(\kappa)$.

Concerning the proof of Theorem 1, the key feature of the additive noise model (1) is that one can estimate the expected squared error $\sigma^2 = E(Y - E(Y|X))^2$ of the regression function $f(x) = E(Y|X = x)$ from the available observations (see Section A below). In fact, the same proof shows that one can consistently estimate $E(Y|X = x)$ from any stationary ergodic process $\{(X_i, Y_i)\}$ with $\{X_i\} \in \mathcal{S}(\mu_0, \alpha)$ and $EY^2 < \infty$ if the error σ^2 is known or can be consistently estimated from the available observations.

The assumption that $\{Z_i\}$ is weakly mixing (in the ergodic theory sense) and independent of $\{X_i\}$ ensures that the joint process $\{(X_i, Z_i)\}$ is ergodic. (For a proof of an analogous result for transformations, see [30].) Ergodicity of the joint process $\{(X_i, Y_i)\}$ follows immediately from that of $\{(X_i, Z_i)\}$.

B Connection to Previous Work

As noted above, the problem considered here is a special case of regression estimation from ergodic processes. The existence of regression estimates that are weakly L_2-consistent for any i.i.d. process was first established by Stone [35] using nearest neighbor methods. Beginning with the papers [33, 34, 31] 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 majority of this work is devoted to central limit theorems and rates of convergence for kernel and histogram type estimates. See the monographs [15, 32, 4] for further references and discussion. Local polynomial regression from weakly dependent processes is considered in [23]. There is also a substantial body of work on regression estimation from stationary processes exhibiting long range (also called strong) dependence. For an overview of these results and additional references, see [3].

Concerning the general setting of interest here, it has recently been shown [2, 27, 40] that no sequence of regression estimates is consistent for every stationary ergodic process $\{(X_i, Y_i)\}$, even if X and Y are assumed to take values in the unit interval. Thus restrictions of some sort must be placed on the family of possible observations in order to establish the consistency of a regression scheme under study. One possible set of restrictions is considered here. Minimal conditions for the existence of consistent regression estimates, if such exist, are at present unknown.

Modha and Masry [25] study lag-adaptive autoregression from stationary processes with exponentially decreasing weak mixing coefficients. Strengthening earlier work of Delecroix [8], Györfi, Härdle, Sarda and Vieu [15] show that one can consistently estimate the k-step autoregression function of an ergodic process if the conditional density of $X_1, \ldots, X_r$ given $X_0, X_{-1}, \ldots$ exists and is continuous for every $r \geq 1$. Similar results are also obtained in [9]. Yakowitz [37, 38] establishes the consistency of kernel and nearest neighbor autoregression estimates for Markov chains under a variety of regularity conditions.

Yakowitz, Györfi, Kieffer and Morvai [39] propose truncated histogram estimates for Lipschitz continuous regression functions. For each constant $L > 0$ they exhibit a sequence of estimates that is almost surely pointwise consistent for every ergodic process $\{(X_i, Y_i)\}$ with regression function f satisfying $|f(x) - f(y)| \leq L||x - y||$. Morvai, Kulkarni, and Nobel [26] proposed adaptive histogram regression estimates for processes with one dimensional covariates X. Given constants $\alpha_1, \alpha_2, \ldots$, they define estimates $\hat{f}_n$ that are strongly L_2 consistent for every ergodic process $\{(X_i, Y_i)\}$ such that X is non-atomic and the variation of the regression function f on $[-k, k]$ is at most α_k for each $k \geq 1$. Kulkarni and Posner [20] consider nearest neighbor regression for general sampling schemes.

Existing work on consistent regression estimation from dependent processes places assumptions on the dependence of the observations, or the regularity of the regression function, or both. By contrast, Theorem 1 shows that, in exchange for prior knowledge about the one dimensional distribution of the sampling process, one may, at a level sufficient for consistency, adapt simultaneously to both the regularity of the sampled function and the mixing rate of the sampling process.

C Outline

The estimates $\{\hat{f}_n\}$ of Theorem 1 are defined in the next section. Section III is devoted to deterministic sampling schemes. It is shown there how the results of Theorem 1 can be used to estimate a measure preserving, ergodic transformation from one of its trajectories. Alternative estimates, similar to those defined in the next section, are briefly discussed in Section IV, where it is shown that the comparability condition (2) can be replaced by the assumption that f is bounded, and that L_1-consistent estimates of $d\mu/d\mu_0$ are available. Proofs of the principle results are given in Section V.

II Description of the 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 $x \in \mathbb{R}^d$,

$$\lim_{k\to\infty} \operatorname{diam}(\pi_k[x]) = 0\,. \tag{5}$$

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. Condition (5) allows the partitions π_k to have unbounded cells, provided that the sequence of cells containing each fixed vector x eventually shrinks down to x. The partition π_l may be obtained, for example, by dividing $[-l, l)^d$ into cubes of side-length 2^{-l}, and letting the complement of $[-l, l)^d$ comprise a single cell.

The estimates $\{\hat{f}_n\}$ are histograms, obtained by partitioning the samples X_i according to one of the partitions π_k and then averaging the corresponding noisy values within each cell. In order to obtain consistent estimates under weak assumptions on the sampling process and the sampled function, some care must be taken when choosing the partition π_k appropriate for a particular set of observations. Choice of an index k is based in part on estimates of the noise variance EZ^2. Variance estimates are defined in the next subsection, and the regression estimates $\{\hat{f}_n\}$ are defined in Section B.

A Estimates of Noise Variance

Let $\{Z_i\}$ be a stationary, zero mean noise process with covariance envelope κ. The condition (4) yields bounds on the probability that weighted averages of the noise sequence deviate from zero. In particular, for every $c > 0$, every $s \geq 1$, every sequence of integers $1 \leq k_1 < k_2 < \cdots < k_s$, and every sequence $\alpha_1, \ldots, \alpha_s \in [-1, 1]$, Chebyshev's inequality and (4) imply that

$$\begin{aligned}
\mathbb{P}\left\{ \frac{1}{s} \sum_{i=1}^{s} \alpha_i Z_{k_i} > c\sqrt{EZ^2} \right\} &\leq \left(s^2 c^2\, EZ^2\right)^{-1} \sum_{1 \leq i,j \leq s} |\operatorname{Cov}(Z_{k_i}, Z_{k_j})| \\
&\leq \left(s^2 c^2\right)^{-1} \sum_{1 \leq i,j \leq s} \kappa(|k_i - k_j|) \\
&\leq \left(s^2 c^2\right)^{-1} \sum_{1 \leq i,j \leq s} \kappa(|i - j|) \\
&\leq \frac{1}{c^2}\left[\frac{\kappa(0)}{s} + \frac{2}{s}\sum_{l=1}^{s} \kappa(l)\right] \triangleq R(c, s)\,.
\end{aligned}$$

The third inequality follows from monotonicity of κ the fact that $|k_i - k_j| \geq |i - j|$. Note that for each $c > 0$, the quantity $R(c, s) \to 0$ as $s \to \infty$.

Suppose now that $\{X_i\}$ is stationary and ergodic, and that $\{(X_i, Y_i) = (X_i, f(X_i) + Z_i)\}$ are noisy samples of a measurable function $f : \mathbb{R}^d \to \mathbb{R}$. The bound $R(c, s)$ can be used to obtain estimates of EZ^2 in the following way. First, let $c_0 = 1$ and select $c_1, c_2, \ldots \in (0, 1)$ such that $\sum_u c_u < \infty$. Choose integers $s_u \to \infty$ such that $R(c_u, s_u) \leq c_u/|\pi_u|$ for each $u \geq 1$, where $|\pi_u|$ denotes the number of cells in the partition π_u. Using $\{s_u\}$, define a subsequence of $1, 2, \ldots$ based on the samples $X_1, X_2, \ldots$ as follows. Set $l(0) = 1$, and for each $u \geq 1$ define

$$l(u) = \min\left\{ l > l(u-1) : \sum_{i=1}^{l} I\{X_i \in A\} \geq s_u \text{ or } = 0 \text{ for every } A \in \pi_u \right\}.$$

Thus $l(u)$ is the first time $l > l(u - 1)$ that each cell of π_u is either empty or contains at least s_u of the samples X_i. For each $n \geq 1$ define the random partition and time indices

$$u_n = \max\{u : l(u) \leq n\} \quad \text{and} \quad m_n = \max\{l(u) : l(u) \leq n\}, \tag{6}$$

both of which depend only on $X_1, \ldots, X_n$ Finally, define the histogram

$$\theta_n(x) \;=\; \frac{\sum_{i=1}^{m_n} Y_i\, I\{X_i \in \pi_{u_n}[x]\}}{\sum_{i=1}^{m_n} I\{X_i \in \pi_{u_n}[x]\}} \tag{7}$$

and the corresponding error estimate

$$\hat{\Gamma}_n \;=\; \frac{1}{m_n}\sum_{i=1}^{m_n}\left(\theta_n(X_i) - Y_i\right)^2 \;. \tag{8}$$

Note that $\hat{\Gamma}_n$ depends only on $(X_1, Y_1), \ldots, (X_n, Y_n)$. The following result is proved in Section V.

Lemma 1 *If $\{Z_i\} \in \mathcal{N}(\kappa)$ is independent of $\{X_i\}$ and $Ef(X)^2 < \infty$, then $\hat{\Gamma}_n \to EZ^2$ with probability one.*

B Definition of the Estimates

Suppose that a reference distribution μ_0, constant α, and envelope κ have been fixed, and that the partitions $\pi_0, \pi_1, \ldots$ are defined as above. Given observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ of the form (1), define candidate histograms

$$\phi_{k,n}(x) = \frac{\sum_{i=1}^n Y_i\, I\{X_i \in \pi_k[x]\}}{\sum_{i=1}^n I\{X_i \in \pi_k[x]\}} \qquad k = 1, 2, \ldots\,. \tag{9}$$

If no vector X_i lies in some cell $\pi_k[x]$, then set $\phi_{k,n}(x) = 0$. Let

$$\Delta_{k,n} = \left(\frac{1}{n}\sum_{i=1}^n |\phi_{k,n}(X_i) - Y_i|^2\right)^{1/2} \tag{10}$$

be the empirical loss of $\phi_{k,n}$. Both $\phi_{k,n}$ and $\Delta_{k,n}$ depend only on the available observations. The estimate $\hat{f}_n$ is chosen from among the candidates $\{\phi_{k,n} : k \geq 1\}$ by selecting a suitable partition index, based on the available data, and on μ_0, α and κ.

Fix a sequence $\epsilon_1, \epsilon_2, \cdots$ of positive numbers tending monotonically to zero. Let s_n be the largest $s \geq 1$ such that

$$||\phi_{j,n} - \phi_{l,n}||_{\mu_0} \;\leq\; 2\alpha^{-1/2}(\Delta_{j,n}^2 - \hat{\Gamma}_n)_+^{1/2} \;+\; 2(1 + \alpha^{-1/2})\epsilon_j \;\;\text{for}\;\; 1 \leq j \leq l \leq s \tag{11}$$

and define $\hat{f}_n(x) = \phi_{s_n,n}(x)$. Here $||f||_{\mu_0} = (\int f^2\, d\mu_0)^{1/2}$ is the usual $L_2(\mu_0)$-norm of f, and $\hat{\Gamma}_n$ is the variance estimate defined in (8) above. Each of the quantities appearing in (11) is either known in advance of the observations, or may be evaluated once the observations are obtained. Thus the estimate is well defined. In general, the partition index s_n will not increase monotonically with the sample size n, nor will it grow at any prespecified rate.

Related estimates for the noiseless setting $Z_i \equiv 0$ are defined in [1], and used there to estimate a measure preserving transformation from a suitable reconstruction sequence. See

Section B for more details. The definition of $\hat{f}_n$ via (11) is similar to the construction of minimax adaptive regression estimates given by Lepskii [21] in a different context. The two estimates differ, however, as we seek the most complex candidate compatible with the available data, while Lepskii seeks the simplest.

The proof of Theorem 1 is carried out in three steps. In the first step it is shown that the partition index s_n tends to infinity as the sample size n increases. Then it is shown that the estimates $\hat{f}_n$ form a Cauchy sequence in $L_2(\mu_0)$, and that their limit is the sampled function f.

III Deterministic Sampling

Deterministic samples are generated by repeatedly applying a fixed, non-linear map $F : \mathbb{R}^d \to \mathbb{R}^d$ to an initial vector $x \in \mathbb{R}^d$. In this case the successive sampling points are given by the trajectory

$$x, Fx, F^2x, \ldots \in \mathbb{R}^d \tag{12}$$

of F starting at x, where F^i denotes the i-fold composition of F with itself. The samples in (12) exhibit dependence across large time scales: from knowledge of F and any individual sample one can, in principle, reconstruct every subsequent sample. The estimates of Theorem 1 can be used to reconstruct measurable functions from the deterministic samples (12) when F is measure preserving and ergodic. Let $\mathcal{B}$ denote the Borel subsets of $\mathbb{R}^d$. We assume in what follows that F is measurable, *i.e.* $F^{-1}A \in \mathcal{B}$ for every $A \in \mathcal{B}$.

Definition: A 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 $\mu(F^{-1}A) = \mu(A)$ for every $A \in \mathcal{B}$, and is said to be ergodic with respect to μ if in addition $F^{-1}A = A$ implies $\mu(A) = 0$ or 1.

A Estimation from Deterministic Samples

To place deterministic sampling in the context of ergodic processes, note that if $F : \mathbb{R}^d \to \mathbb{R}^d$ is μ-preserving and ergodic, and if $X \in \mathbb{R}^d$ is a random vector with distribution μ, then $\mathbf{X} = \{X, FX, F^2X, \ldots\}$ is a stationary ergodic process with one-dimensional distribution μ. The process $\mathbf{X}$ evolves in a deterministic fashion, with uncertainty entering only through choice of the initial vector X. In particular, $\mathbf{X}$ exhibits very long range dependence and fails to satisfy most standard mixing assumptions.

Consider again the estimation problem discussed in the introduction. Let regression estimates $\{\hat{f}_n\}$ be defined as in (11), in terms of a reference distribution μ_0, constant α, and covariance envelope κ. Suppose that $F : \mathbb{R}^d \to \mathbb{R}^d$ is a μ-preserving ergodic map, where μ satisfies $\alpha \leq d\mu/d\mu_0 < \beta$ for some $\beta < \infty$. The following result is an immediate corollary of Theorem 1.

Proposition 1 *For μ_0-almost every $x \in \mathbb{R}^d$, each noise process $\{Z_i\} \in \mathcal{N}(\kappa)$ independent of the choice of x, and every measurable function $f : \mathbb{R}^d \to \mathbb{R}$ with $\int f^2\, d\mu < \infty$, the estimates $\hat{f}_n$ produced from the deterministic samples*

$$(x,\, f(x) + Z_1), \ldots, (F^{n-1}x,\, f(F^{n-1}x) + Z_n)$$

are such that $\int (\hat{f}_n - f)^2\, d\mu \to 0$ with probability one. In particular, the estimates are consistent for any i.i.d. noise sequence with finite second moment.

B Estimating an Ergodic Map

The results of Theorem 1 can be applied to the problem of estimating an ergodic μ-preserving map F from a single trajectory $x, Fx, F^2x, \ldots$ in the absence of noise. To motivate the problem, note that if successive points in the trajectory are grouped together, each of the resulting pairs (x, Fx), (Fx, F^2x), $\ldots$ 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. When F is irregular, or noise is present, selective local averaging is necessary to obtain consistent estimates.

Fix a reference distribution μ_0 and a constant $\alpha \in (0, 1)$. Suppose that F is a measure preserving ergodic map. Given a trajectory $x, Fx, \ldots$ of F starting at a vector $x \in \mathbb{R}^d$, define $(x_i, y_i) = (F^{i-1}x, F^ix)$ to be the i'th pair of successive terms in the sequence, and let y_i^j be the j'th component of the vector y_i. For each $j = 1, \ldots, d$ let s_n^j be the partition index selected according to (11) on the basis of observations $\{(x_i, y_i^j)\}_{i=1}^n$ with $\hat{\Gamma}_n = 0$. Set $w_n = \min\{s_n^1, \ldots, s_n^d\}$ and define the multivariate histogram

$$\hat{F}_n(u) \;=\; \frac{\sum_{i=1}^n F^i x \cdot I\{F^{i-1}x \in \pi_{w_n}[u]\}}{\sum_{i=1}^n I\{F^{i-1}x \in \pi_{w_n}[u]\}}, \tag{13}$$

which is based solely on $x, Fx, \ldots, F^n x$. Let $||u||$ denote the ordinary Euclidean norm of a vector $u \in \mathbb{R}^d$.

Proposition 2 *For every distribution μ such that $\alpha \leq d\mu/d\mu_0 \leq \beta$ for some $\beta < \infty$, and every μ-preserving ergodic map $F : \mathbb{R}^d \to \mathbb{R}^d$ such that $\int ||F||^2 d\mu < \infty$,*

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

for μ-almost every initial vector $x \in \mathbb{R}^d$.

Proof: If the initial value of the trajectory is selected according to a random vector X with distribution μ then the paired process $(X, FX), (FX, F^2X), \ldots$ is stationary and ergodic. These are noiseless measurements of F generated by deterministic sampling according to F. Note that one can express the estimate $\hat{F}_n$ as a vector of estimates $(\psi^1_{n,w_n}, \ldots, \psi^d_{n,w_n})$, where the j'th coordinate is the univariate histogram ψ^j_{n,w_n} based on (x_i, y^j_i). As shown in the proof of Theorem 1, each of the partition indices $s^j_n \to \infty$ as $n \to \infty$ with μ-probability one, and therefore $w_n \to \infty$ with μ-probability one as well. Arguments analogous to those in the proof of Theorem 1 show that ψ^j_{n,w_n} converges in $L_2(\mu)$ to the j'th component function of F with μ-probability one, and the result follows.

The problem of estimating an iterated map has previously been studied primarily in the context of smooth dynamical systems, 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 [12, 6, 7, 19, 29, 22, 36], and the surveys [11, 18, 17]. In most of this work it is assumed that the map under study is differentiable, and that successive iterates of the map are perturbed by observational or dynamical noise. Central limit theorems for U-statistics and smooth functionals of noiseless dynamical systems generated by piecewise-monotone maps are studied in [16, 24, 10].

In the noiseless setting considered here, Bosq and Guégan [5] studied the estimation of uniform mixing continuous maps F using kernel density estimators. Aside from the integrability condition $\int ||F||^2 d\mu < \infty$, the assumptions of Proposition 2 concern only the measure preserved by F; no conditions are placed on the regularity of F or on its mixing properties. An extension of Proposition 2 can be found in [1] where estimates like those in Section II are defined for noiseless samples, and used to estimate a measure preserving (not necessarily ergodic) transformation of a Polish space $\mathcal{X}$ from a suitable reconstruction sequence. The estimates constructed in [1] are consistent in the stronger sense that $\mu(\hat{F}_n^{-1} A \Delta F^{-1} A) \to 0$ for every Borel subset A of $\mathcal{X}$.

The family of transformations F for which the estimates $\hat{F}_n$ are consistent may be quite large. To illustrate this, let $\alpha = d = 1$ and let μ_0 be Lebesgue measure on $[0,1]$. The following corollary of Proposition 2 is immediate.

Corollary 1 *For every ergodic Lebesgue measure preserving map $F : [0,1] \to [0,1]$, for almost every initial value $x \in [0,1]$, the estimates $\hat{F}_n$ obtained from $x, Fx, F^2x, \ldots$ are such that $\int_0^1 (\hat{F}_n - F)^2 \, du \to 0$.*

Among the ergodic Lebesgue measure preserving transformations of $[0,1]$ there is an uncountable subfamily $\mathcal{S}$ with the property that no two transformations in $\mathcal{S}$ are isomorphic. The histograms $\hat{F}_n$ give consistent estimates of each transformation in $\mathcal{S}$ from almost every one of its trajectories.

IV Related Estimates

Modifications of the estimates $\hat{f}_n$ proposed above can recover functions under general conditions somewhat different than those considered in Theorem 1. In particular, when consistent estimates of the density $d\mu/d\mu_0$ of the sampling process are available, the comparability assumption (2) can be dropped.

Suppose now that μ_0 is d-dimensional Lebesgue measure, and that a covariance envelope κ, and sequence $\{\hat{h}_n : n \geq 1\}$ of d-variate density estimates have been fixed. Consider as before observations

$$(X_i, f(X_i) + Z_i) \quad i \geq 1$$

generated by sampling an unknown measurable function f according to an ergodic process $\{X_i\}$ whose marginal distribution μ has density $h = d\mu/dx$ with respect to Lebesgue measure. Let $\hat{h}_n = \hat{h}_n(X_1, \ldots, X_n)$ be an estimate of $d\mu/d\mu_0$ based on the first n sampling points. With $\phi_{j,n}$, $\Delta_{j,n}$, and $\hat{\Gamma}_n$ defined as in Section B, let t_n be the largest $t \geq 1$ such that

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

and define the estimate $\tilde{f}_n(x) = \phi_{t_n,n}(x)$. By arguments analogous to those in the proof of Theorem 1 one may establish the following result.

Theorem 2 *If $\int |\hat{h}_n - h| \, dx \to 0$ with probability one, then for every bounded measurable function f and every bounded noise process $\{Z_i\} \in \mathcal{N}(\kappa)$ independent of $\{X_i\}$, $\int (\tilde{f}_n - f)^2 d\mu \to 0$ with probability one.*

Remark: If n is replaced by m_n and t_n is replaced by $\min\{t_n, u_n\}$, with m_n and u_n defined as in (6), the resulting estimates are consistent under the weaker assumption that $EZ^2 < \infty$.

One choice of candidate density estimates $\{\hat{h}_n\}$ are the recursive kernel estimates proposed by Györfi and Masry [14], which they show to be strongly L_1 consistent for every ergodic process such that the conditional distribution of X_0 given the infinite past $X_{-1}, X_{-2}, \ldots$ is absolutely continuous with probability one. The regression estimates $\tilde{f}_n$ defined via (14) using the recursive kernel estimates will be consistent for every ergodic sampling process satisfying the absolute continuity condition. Another choice of candidate estimates are the univariate histograms studied in [28], which are L_1 consistent for every univariate ergodic process whose one-dimensional marginal density has variation less than a known constant. This includes processes with monotone and multi-modal densities, where both the number of modes and the value of the density are bounded by known constants.

V Derivations

Fix a probability distribution μ on $\mathbb{R}^d$ and let $||f||_\mu = \left(\int |f|^2 d\mu\right)^{1/2}$ be the usual $L_2(\mu)$ norm of a measurable function $f : \mathbb{R}^d \to \mathbb{R}$. For each finite partition π of $\mathbb{R}^d$ and each function f such that $||f||_\mu < \infty$ define

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

provided that $\mu(\pi[x]) > 0$, and set $(f \circ \pi)(x) = 0$ otherwise. Thus $f \circ \pi$ is the conditional expectation of f given π, and is constant on the cells of π. Note that $||f \circ \pi||_\mu \leq ||f||_\mu$.

Suppose that $(X_1, Y_1), (X_2, Y_2), \ldots \in \mathbb{R}^d \times \mathbb{R}$ is a stationary ergodic sequence such that X has distribution μ and $EY^2 < \infty$. Let $f(x) = E(Y|X = x)$ be the regression function of Y on X, and define $f_k = f \circ \pi_k$, where $\{\pi_k\}$ is the sequence of partitions described in Section II. Part (a) of the next lemma follows from standard results on conditional expectations and martingales. A direct proof is given for completeness.

Lemma 2 *If $\phi_{k,n}$ and $\Delta_{k,n}$ are defined as in equations (9) and (10) then the following relations hold.*

a. $||f_k - f||_\mu$ *decreases to zero as* $k \to \infty$;

b. With probability one, $\max_{x \in \mathbb{R}^d} |\phi_{k,n}(x) - f_k(x)| \to 0$ *for each* $k \geq 1$.

c. With probability one, $\Delta_{k,n}^2 \to E(f(X) - Y)^2 + ||f_k - f||_\mu^2$ *for each* $k \geq 1$.

Proof: Fix $\epsilon > 0$ and let f' be a bounded, continuous function such that $||f - f'||_\mu \leq \epsilon$. By an obvious upper bound,

$$\begin{aligned} ||f - f_k||_\mu &\leq ||f - f'||_\mu + ||f' - f' \circ \pi_k||_\mu + ||(f' - f) \circ \pi_k||_\mu \\ &\leq 2\epsilon + ||f' - f' \circ \pi_k||_\mu \,. \end{aligned}$$

The continuity of f' and the shrinking cell condition (5) ensure that $f' \circ \pi_k(x) \to f'(x)$ for each $x \in \mathbb{R}^d$. It then follows from the dominated convergence theorem and the last inequality above that

$$\limsup_{k \to \infty} ||f - f_k||_\mu \leq 2\epsilon \,.$$

As $\epsilon > 0$ was arbitrary, $||f - f_k||_\mu \to 0$. For any set $A \subseteq \mathbb{R}^d$ with $\mu(A) > 0$ the integral $\int_A |f(x) - c|^2 d\mu$ is minimized over constants by $c = \mu(A)^{-1} \int_A f d\mu$. Therefore, among all those functions g that are constant on the cells of π_{k+1}, the integral $\int |f - g|^2 d\mu$ is minimized by $g = f_{k+1}$. This establishes assertion a.

Fix $k \geq 1$ for the moment and let A be a cell of π_k. The ergodic theorem implies that with probability one

$$\frac{1}{n} \sum_{i=1}^n Y_i I\{X_i \in A\} \to E(Y I\{X \in A\}) = \int_A f d\mu$$

and

$$\frac{1}{n} \sum_{i=1}^n I\{X_i \in A\} \to \mu(A) \,.$$

As π_k has finitely many cells, and k ranges over a countable index set, (b) is immediate. (Recall that $\phi_{k,n}(x) = f_k(x) = 0$ if $\mu(\pi_k[x]) = 0$.) To establish (c) define $\tilde{\Delta}_{k,n} = (\, n^{-1} \sum_{i=1}^n |f_k(X_i) - Y_i|^2 \,)^{1/2}$. By part (b) of the lemma,

$$|\Delta_{k,n} - \tilde{\Delta}_{k,n}| \leq \left(\frac{1}{n} \sum_{i=1}^n |\phi_{k,n}(X_i) - f_k(X_i)|^2 \right)^{1/2} \leq \max_{x \in \mathbb{R}^d} |\phi_{k,n}(x) - f_k(x)| \,,$$

which tends to zero with probability one as $n \to \infty$. On the other hand, the ergodic theorem ensures that $\tilde{\Delta}_{k,n}^2 \to E(f_k(X) - Y)^2 = E(f(X) - Y)^2 + ||f_k - f||_\mu^2$ with probability one as $n \to \infty$, and the proof is complete. ♣

A Proof of Lemma 1

Let θ_n and $\hat{\Gamma}_n$ be defined as in (7) and (8), respectively. Note that $\theta_n(x) = U_n(x) + V_n(x)$, where

$$U_n(x) = \frac{\sum_{i=1}^{m_n} Z_i I\{X_i \in \pi_{u_n}[x]\}}{\sum_{i=1}^{m_n} I\{X_i \in \pi_{u_n}[x]\}} \quad \text{and} \quad V_n(x) = \frac{\sum_{i=1}^{m_n} f(X_i) I\{X_i \in \pi_{u_n}[x]\}}{\sum_{i=1}^{m_n} I\{X_i \in \pi_{u_n}[x]\}}.$$

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

$$\left| \hat{\Gamma}_n - \frac{1}{m_n} \sum_{i=1}^{m_n} Z_i^2 \right| \le |\Theta_{1,n}| + \cdots + |\Theta_{5,n}|,$$

where the quantities $\Theta_{i,n}$ are defined as follows:

$$\Theta_{1,n} = \frac{1}{m_n} \sum_{i=1}^{m_n} (V_n(X_i) - f(X_i))^2 \qquad \Theta_{2,n} = \frac{2}{m_n} \sum_{i=1}^{m_n} Z_i (V_n(X_i) - f(X_i))$$

$$\Theta_{3,n} = \frac{1}{m_n} \sum_{i=1}^{m_n} U_n^2(X_i)$$

$$\Theta_{4,n} = \frac{2}{m_n} \sum_{i=1}^{m_n} U_n(X_i)(V_n(X_i) - f(X_i)) \qquad \Theta_{5,n} = \frac{1}{m_n} \sum_{i=1}^{m_n} Z_i U_n(X_i).$$

The ergodic theorem ensures that m_n and r_n tend to infinite along almost every sample sequence of $\{X_i\}$. In particular, $m_n^{-1} \sum_{i=1}^{m_n} Z_i^2 \to EZ_1^2$ with probability one, as $\{Z_i\}$ is ergodic and independent of $\{X_i\}$. It is therefore enough to show that $\Theta_{j,n} \to 0$ with probability one for $j = 1, \ldots, 5$.

Consider first $\Theta_{1,n}$, which depends only on the sampling process $\{X_i\}$. For fixed u and each $n \ge 1$ define the histograms

$$\psi_{u,n}(x) = \frac{\sum_{i=1}^{m_n} f(X_i) I\{X_i \in \pi_u[x]\}}{\sum_{i=1}^{m_n} I\{X_i \in \pi_u[x]\}}.$$

By arguments like those in the proof of Lemma 2 one may show that

$$\Theta_{1,n} \le \frac{1}{m_n} \sum_{i=1}^{m_n} (\psi_{u,n}(X_i) - f(X_i))^2,$$

when $u_n \ge u$, and that $\max_{x \in \mathbb{R}^d} |\psi_{u,n}(x) - (f \circ \pi_u)(x)| \to 0$ as $n \to \infty$. It follows that

$$0 \le \limsup_{n \to \infty} \Theta_{1,n} \le E((f \circ \pi_u)(X) - f(X))^2$$

with probability one for each $u \ge 1$. Letting u tend to infinity shows that $\Theta_{1,n} \to 0$. By the Cauchy-Schwartz inequality,

$$\Theta_{2,n} \le 2 \left(\frac{1}{m_n} \sum_{i=1}^{m_n} Z_i^2 \right)^{1/2} \Theta_{1,n}^{1/2}$$

As n tends to infinity, the first term on the right hand side tends to a finite limit, while the second term tends to zero. Thus $\Theta_{2,n} \to 0$.

In order to evaluate the remaining terms, we investigate the conditional behavior of $U_n(\cdot)$ given the sampling process. Suppose that the sampling points $X_1^\infty = X_1, X_2, \ldots$ are fixed. Then u_n and m_n are fixed, and as $\{Z_i\}$ is independent of $\{X_i\}$, the collection $\{Z_i : X_i \in A,\, i \leq m_n\}$ is a fixed (non-random) subsequence of $Z_1, \ldots, Z_{m_n}$ for each cell $A \in \pi_{u_n}$. Now observe that

$$\begin{aligned} & \mathbb{P}\left\{ \sup_{x \in \mathbb{R}^d} |U_n(x)| \geq (EZ^2)^{1/2} \cdot c_{u_n} \mid X_1^\infty \right\} \\ & \leq \sum_{A \in \pi_{u_n}} \mathbb{P}\left\{ \left| \frac{\sum_{i=1}^{m_n} Z_i I\{X_i \in A\}}{\sum_{i=1}^{m_n} I\{X_i \in A\}} \right| \geq (EZ^2)^{1/2} \cdot c_{u_n} \mid X_1^\infty \right\} \end{aligned} \tag{15}$$

and consider each term in the sum. If no covariate X_i lies in A then the corresponding probability is zero. Alternatively, if A contains at least one X_i, then by the definition of m_n and the assumption that $\{Z_i\} \in \mathcal{N}(\kappa)$, the corresponding probability is at most $c_{u_n}/|\pi_{u_n}|$. Thus the probability on the left side of (15) is at most c_{u_n}. (This and each of the conditional statements below holds for almost every sample sequence of X_1^∞.) This implies that

$$\mathbb{P}\left\{ \Theta_{3,n} \geq EZ^2 \cdot c_{u_n}^2 \mid X_1^\infty \right\} \;\leq\; c_{u_n},$$

and therefore for each $\epsilon > 0$,

$$\sum_{k=1}^{\infty} \mathbb{P}\left\{ \Theta_{3,l(k)} \geq \epsilon \mid X_1^\infty \right\} \;\leq\; \min\{k : EZ^2 \cdot c_k^2 \leq \epsilon\} \;+\; \sum_{k=1}^{\infty} c_k \;<\; \infty.$$

It follows from this last inequality and the Borel-Cantelli Lemma that

$$\mathbb{P}\left\{ \lim_{k\to\infty} \Theta_{3,l(k)} = 0 \mid X_1^\infty \right\} = 1\,.$$

By definition, $\Theta_{3,n} = \Theta_{3,l(k)}$ for $n = l(k), \ldots, l(k+1) - 1$ and consequently

$$\mathbb{P}\left\{ \Theta_{3,n} \to 0 \mid X_1^\infty \right\} = 1\,.$$

Integrating over X_1^∞ shows that $\Theta_{3,n} \to 0$ with probability one, as desired. An argument like that used to show $\Theta_{2,n} \to 0$ now shows that $\Theta_{4,n}, \Theta_{5,n} \to 0$ as well. ♣

B Proof of Theorem 1

Fix a sampling process $\{X_i\}$, noise process $\{Z_i\}$, and measurable function f satisfying the conditions of the theorem. Suppose that $\{X_i\}$ and $\{Z_i\}$ are defined on an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Given $s, N \geq 1$ let $A(s, N) \in \mathcal{F}$ be the event that for each $r = 1, \ldots, s$ and each $n \geq N$,

$$\max_{x \in \mathbb{R}^d} |\phi_{r,n}(x) - f_r(x)| \leq \epsilon_r \quad \text{and} \quad \Big| \, ||f_r - f||_\mu - (\Delta_{r,n}^2 - \hat{\Gamma}_n)_+^{1/2} \, \Big| \leq \epsilon_r \tag{16}$$

where $f_r = f \circ \pi_r$. Suppose that $A(s, N)$ occurs. Fix $n \geq N$ and $1 \leq j \leq l \leq s$. The inequalities (16) and the assumption that $d\mu/d\mu_0 \geq \alpha$ imply that

$$\begin{aligned}
||\phi_{j,n} - \phi_{l,n}||_{\mu_0} &\leq ||f_j - f_l||_{\mu_0} + 2\epsilon_j \\
&\leq \alpha^{-1/2} \, ||f_j - f_l||_\mu + 2\epsilon_j \\
&\leq \alpha^{-1/2} \, (||f_j - f||_\mu + ||f_l - f||_\mu) + 2\epsilon_j \\
&\leq 2\alpha^{-1/2} \, ||f_j - f||_\mu + 2\epsilon_j \\
&\leq 2\alpha^{-1/2} \, (\Delta_{j,n}^2 - \hat{\Gamma}_n)_+^{1/2} + 2(1 + \alpha^{-1/2})\epsilon_j \, .
\end{aligned}$$

It follows from the definition (11) of s_n that $A(s, N)$ implies $s_n \geq s$ for each $n \geq N$, and therefore,

$$\left\{ \lim_{n \to \infty} s_n = \infty \right\} \supseteq \bigcap_{s=1}^{\infty} \bigcup_{N=1}^{\infty} A(s, N) \, .$$

Lemmas 1 and 2 imply that $\mathbb{P}(A(s, N)) \to 1$ as $N \to \infty$ for each $s \geq 1$, and consequently $s_n \to \infty$ with probability one.

If $A(s, N)$ occurs then $s_m, s_n \geq s$ for each $m, n \geq N$ and therefore the inequalities (16) and (11) imply that

$$\begin{aligned}
||\hat{f}_n - \hat{f}_m||_{\mu_0} &\leq ||\phi_{s_n,n} - \phi_{s,n}||_{\mu_0} + ||\phi_{s_m,m} - \phi_{s,m}||_{\mu_0} + 2\epsilon_s \\
&\leq 2\alpha^{-1/2} \, [\, (\Delta_{s,n}^2 - \hat{\Gamma}_n)_+^{1/2} + (\Delta_{s,m}^2 - \hat{\Gamma}_m)_+^{1/2} \,] + (6 + 4\alpha^{-1/2})\epsilon_s \\
&\leq 4\alpha^{-1/2} \, ||f_s - f||_\mu + (6 + 8\alpha^{-1/2})\epsilon_s \, .
\end{aligned}$$

Given $\delta > 0$ let $s(\delta)$ be any integer for which this last expression is less then δ. The last inequality ensures that

$$\left\{ \inf_{N \geq 1} \sup_{n,m \geq N} \int |\hat{f}_n - \hat{f}_m| \, d\mu_0 \leq \delta \right\} \supseteq \bigcup_{N=1}^{\infty} A(s(\delta), N) \, .$$

Lemma 2 implies $\mathbb{P}(A(s(\delta), N)) \to 1$ as $N \to \infty$, and therefore the estimates $\{\hat{f}_n\}$ form a Cauchy sequence in $L_2(\mu_0)$ with probability one.

Consider a sample sequence of $\{(X_i, Y_i)\}$ for which the estimates $\{\hat{f}_n\}$ are Cauchy. As $L_2(\mu_0)$ is complete, there is a function $f^* \in L_2(\mu_0)$ such that $||\hat{f}_n - f^*||_{\mu_0} \to 0$. For each n

such that $s_n \geq r$,

$$\begin{aligned}
&||f - f^*||_{\mu_0} \\
&\quad\leq\; ||f - f_r||_{\mu_0} + ||f_r - \phi_{r,n}||_{\mu_0} + ||\phi_{r,n} - \phi_{s_n,n}||_{\mu_0} + ||\hat{f}_n - f^*||_{\mu_0} \\
&\quad\leq\; \alpha^{-1/2}\, ||f - f_r||_\mu \;+\; \max_{x \in \mathbb{R}^d} |f_r(x) - \phi_{r,n}(x)| \;+\; 2\alpha^{-1/2}\, (\Delta^2_{r,n} - \hat{\Gamma}_n)_+^{1/2} \\
&\qquad\; + 2(1 + \alpha^{-1/2})\epsilon_r \;+\; ||\hat{f}_n - f^*||_{\mu_0}\,.
\end{aligned}$$

Letting n, and then r, tend to infinity shows that $||f - f^*||_{\mu_0} = 0$. By assumption, $d\mu/d\mu_0 \leq \beta$ for some $\beta < \infty$, and consequently $||\hat{f}_n - f||_\mu \to 0$. ♣

References

[1] T.M. Adams and A.B. Nobel, "Finitary reconstruction of a measure preserving transformation", to appear in the *Israel Journal of Mathematics*, 2001.

[2] T.M. Adams, "Families of ergodic processes without consistent density or regression estimates", preprint, 1999.

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

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

[5] D. Bosq and D. Guégan, "Nonparametric estimation of the chaotic function and the invariant measure of a dynamical system", *Stat. and Prob. Let.*, vol.25, pp.201–212, 1995.

[6] M. Casdagli, "Nonlinear prediction of chaotic time series", *Physica D*, vol.35, pp.335–356, 1989.

[7] M. Casdagli, "Chaos and deterministic *versus* stochastic non-linear modeling", *J. R. Stat. Soc. B*, vol.54, pp.303–328, 1992.

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

[9] M. Delecroix and A.C. Rosa, "Nonparametric estimation of a regression function and its derivatives under an ergodic hypothesis", *Nonparam. Statistics*, vol.6, pp.367–382, 1996.

[10] M. Denker and G. Keller, "Rigorous statistical procedures for data from dynamical systems", *J. Stat. Phys.*, vol.44, pp.67–93, 1986.

[11] J.-P. Eckmann and D. Ruelle, "Ergodic theory of chaos and strange attractors", *Rev. Mod. Phys.*, vol.57, pp.617–656, 1985.

[12] J.D. Farmer and J.J. Sidorowich, "Predicting chaotic time series", *Phys. Rev. Let.*, vol.59, pp.845–848, 1987.

[13] L. Györfi, "Strongly consistent density estimate from ergodic sample", *J. Multivariate Analysis*, vol.11, pp.81–84, 1981.

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

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

[16] F. Hofbauer and G. Keller, "Ergodic properties of invariant measures for piecewise monotonic functions", *Mathematische Zeitschrift*, vol.180, pp.119–140, 1982.

[17] V. Isham, "Statistical aspects of chaos: a review", in *Networks and Chaos – Statistical and Probabilistic Aspects*, O.E. Barndorff-Nielsen, J.L. Jensen, and W.S. Kendall editors, Chapman and Hall, London, 1993.

[18] J.L. Jensen, "Chaotic dynamical systems with a view towards statistics: a review" in *Networks and Chaos – Statistical and Probabilistic Aspects*, O.E. Barndorff-Nielsen, J.L. Jensen, and W.S. Kendall editors, Chapman and Hall, London, 1993.

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

[20] S.R. Kulkarni and S.E. Posner, "Rates of convergence of nearest neighbor estimation under arbitrary sampling", *IEEE Trans. Inform. Theory*, vol.41, pp.1028-1039, 1995.

[21] O.V. Lepskii, "Asymptotically minimax adaptive estimation I: upper bounds, optimally adaptive estimates", *Theory Probab. Appl.*, vol.36, pp.682–697, 1991.

[22] Z.-Q. Lu and R.L. Smith, "Estimating local Lyapunov exponents", *Fields Institute Communications*, vol.11, pp.135–151, 1997.

[23] E. Masry, "Multivariate local polynomial regression for time series: uniform strong consistency and rates", *J. Time Sers. Anal.*, vol.17, pp.571-599, 1996.

[24] D.H. Mayer, "Approach to equilibrium for locally expanding maps in R^k", *Communications in Mathematical Physics*, vol.95, pp.1–15, 1984.

[25] D. Modha and E. Masry, "Memory-universal prediction of stationary random processes", *IEEE Trans. Inform. Theory*, vol.44, pp.117-133, 1998.

[26] G. Morvai, S. Kulkarni, and A.B. Nobel, "Regression estimation from an individual stable sequence", *Statistics*, vol.33, pp.99-118, 1999.

[27] A.B. Nobel, "Limits to classification and regression estimation from ergodic processes", *Annals of Statistics*, vol.27, pp.262-273, 1999.

[28] A.B. Nobel, G. Morvai, and S. Kulkarni, "Density estimation from an individual numerical sequence" *IEEE Trans. Inform. Theory*, vol.44, pp.537–541, 1998.

[29] D. Nychka, S. Ellner, A.R. Gallant, and D. McCaffrey, "Finding chaos in noisy systems", *J. R. Stat. Soc. B*, vol.54, pp.399-426, 1992.

[30] K. Petersen, *Ergodic Theory*, Cambridge Univ. Press, 1989.

[31] M. Rosenblatt, "Density estimates and Markov sequences", In *Nonparametric Techniques in Statistical Inference*, M. Puri editor, Cambridge Univ. Press, London, 199-210, 1970.

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

[33] G. Roussas, "Nonparametric estimation in Markov processes", *Ann. Inst. Statist. Math.*, vol.21, pp.73-87, 1967.

[34] G. Roussas, "Nonparametric estimation of the transition distribution function of a Markov process", *Ann. Math. Stat.*, vol.40, pp.1386-1400, 1969.

[35] C. Stone, "Consistent nonparametric regression", *Ann. Stat.*, vol.5, pp.595-620, 1977.

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

[37] S. Yakowitz, "Nonparametric density and regression estimation for Markov sequences without mixing assumptions", *J. Multivar. Anal.*, vol.30, pp124–136, 1989.

[38] S. Yakowitz, "Nearest neighbor regression estimation for null-recurrent Markov time series", *Stoc. Proc. Appl.*, vol.48, pp.311–318, 1993.

[39] S. Yakowitz, L. Györfi, J. Kieffer, and G. Morvai, "Strongly-consistent nonparametric estimation of smooth regression functions for stationary ergodic sequences" *J. Multivar. Anal.*, vol.71, pp.24-41, 1999.

[40] S. Yakowitz and C. Heyde, "Long range dependency effects with implications for forecasting and queuing inference", preprint, 1998.