Estimating beta-mixing coefficients

Daniel J. McDonald
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213
danielmc@stat.cmu.edu

Cosma Rohilla Shalizi
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213
cshalizi@stat.cmu.edu

Mark Schervish
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213
mark@cmu.edu

Abstract

The literature on statistical learning for time series assumes the asymptotic independence or “mixing” of the data-generating process. These mixing assumptions are never tested, and there are no methods for estimating mixing rates from data. We give an estimator for the beta-mixing rate based on a single stationary sample path and show it is L1-risk consistent.

1 Introduction

Relaxing the assumption of independence is an active area of research in the statistics and machine learning literature. For time series, independence is replaced by the asymptotic independence of events far apart in time, or “mixing”. Mixing conditions make the dependence of the future on the past explicit by quantifying the decay in dependence as the future moves farther from the past. There are many definitions of mixing of varying strength with matching dependence coefficients (see [9, 7, 4] for reviews), but most of the results in the learning literature focus on β-mixing or absolute regularity. Roughly speaking (see Definition 2.1 below for a precise statement), the β-mixing coefficient at lag a is the total variation distance between the actual joint distribution of events separated by a time steps and the product of their marginal distributions, i.e., the L^1 distance from independence.

Numerous results in the statistical machine learning literature rely on knowledge of the β-mixing coefficients. As Vidyasagar [25, p. 41] notes, β-mixing is “just right” for the extension of IID results to dependent data, and so recent work has consistently focused on it. Meir [15] derives generalization error bounds for nonparametric methods based on model selection via structural risk minimization. Baraud et al. [1] study the finite sample risk performance of penalized least squares regression estimators under β-mixing. Lozano et al. [13] examine regularized boosting algorithms under absolute regularity and prove consistency. Karandikar and Vidyasagar [12] consider “probably approximately correct” learning algorithms, proving that PAC algorithms for IID inputs remain PAC with β-mixing inputs under some mild conditions. Ralaivola et al. [20] derive PAC bounds for ranking statistics and classifiers using a decomposition of the dependency graph. Finally, Mohri and Rostamizadeh [16] derive stability bounds for β-mixing inputs, generalizing existing stability results for IID data.

All these results assume not just β-mixing, but known mixing coefficients. In particular, the risk bounds in [15, 16] and [20] are incalculable without knowledge of the rates. This knowledge is *never* available. Unless researchers are willing to assume specific values for a sequence of β-mixing coefficients, the results mentioned in the previous paragraph are generally useless when confronted with data.To illustrate this deficiency, consider Theorem 18 of [16]:

Theorem 1.1 (Briefly). *Assume a learning algorithm is λ-stable.*[1] *Then, for any sample of size n drawn from a stationary β-mixing distribution, and $\epsilon > 0$*

$$\mathbb{P}(|R - \widehat{R}| > \epsilon) \leq \Gamma(n, \lambda, \epsilon, a, b) + \beta(a)(\mu_n - 1)$$

where $n = (a + b)\mu_n$, Γ has a particular functional form, and $R - \widehat{R}$ is the difference between the true risk and the empirical risk.

Ideally, one could use this result for model selection or to control the size of the generalization error of

Appearing in Proceedings of the 14^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15 of JMLR: W&CP 15. Copyright 2011 by the authors.

[1] The literature on algorithmic stability refers to this as β-stability (e.g. Bousquet and Elisseeff [3]).

competing prediction algorithms (support vector machines, support vector regression, and kernel ridge regression are a few of the many algorithms known to satisfy λ-stability). However the bound depends explicitly on the mixing coefficient $\beta(a)$. To make matters worse, there are *no* methods for estimating the β-mixing coefficients. According to Meir [15, p. 7], "there is no efficient practical approach known at this stage for estimation of mixing parameters." We begin to rectify this problem by deriving the first method for estimating these coefficients. We prove that our estimator is consistent for arbitrary β-mixing processes. In addition, we derive rates of convergence for Markov approximations to these processes.

Application of statistical learning results to β-mixing data is highly desirable in applied work. Many common time series models are known to be β-mixing, and the rates of decay are known given the true parameters of the process. Among the processes for which such knowledge is available are ARMA models [17], GARCH models [5], and certain Markov processes — see [9] for an overview of such results. To our knowledge, only Nobel [18] approaches a solution to the problem of estimating mixing rates by giving a method to distinguish between different polynomial mixing rate regimes through hypothesis testing.

We present the first method for estimating the β-mixing coefficients for stationary time series data. Section 2 defines the β-mixing coefficient and states our main results on convergence rates and consistency for our estimator. Section 3 gives an intermediate result on the L^1 convergence of the histogram estimator with β-mixing inputs. Section 4 proves the main results from §2. Section 5 concludes and lays out some avenues for future research.

2 Estimation of β-mixing

In this section, we present one of many equivalent definitions of absolute regularity and state our main results, deferring proof to §4.

To fix notation, let $\mathbf{X} = \{X_t\}_{t=-\infty}^{\infty}$ be a sequence of random variables where each X_t is a measurable function from a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ into a measurable space $\mathcal{X}$. A block of this random sequence will be given by $\mathbf{X}_i^j \equiv \{X_t\}_{t=i}^j$ where i and j are integers, and may be infinite. We use similar notation for the sigma fields generated by these blocks and their joint distributions. In particular, σ_i^j will denote the sigma field generated by $\mathbf{X}_i^j$, and the joint distribution of $\mathbf{X}_i^j$ will be denoted $\mathbb{P}_i^j$.

2.1 Definitions

There are many equivalent definitions of β-mixing (see for instance [9], or [4] as well as Meir [15] or Yu [28]), however the most intuitive is that given in Doukhan [9].

Definition 2.1 (β-mixing)**.** *For each positive integer a, the* coefficient of absolute regularity*, or* β-mixing coefficient*, $\beta(a)$, is*

$$\beta(a) \equiv \sup_t \left|\left|\mathbb{P}_{-\infty}^t \otimes \mathbb{P}_{t+a}^{\infty} - \mathbb{P}_{t,a}\right|\right|_{TV} \tag{1}$$

where $||\cdot||_{TV}$ is the total variation norm, and $\mathbb{P}_{t,a}$ is the joint distribution of $(\mathbf{X}_{-\infty}^t, \mathbf{X}_{t+a}^{\infty})$. A stochastic process is said to be absolutely regular*, or* β-mixing*, if $\beta(a) \to 0$ as $a \to \infty$.*

Loosely speaking, Definition 2.1 says that the coefficient $\beta(a)$ measures the total variation distance between the joint distribution of random variables separated by a time units and a distribution under which random variables separated by a time units are independent. The supremum over t is unnecessary for stationary random processes $\mathbf{X}$ which is the only case we consider here.

Definition 2.2 (Stationarity)**.** *A sequence of random variables $\mathbf{X}$ is* stationary *when all its finite-dimensional distributions are invariant over time: for all t and all non-negative integers i and j, the random vectors $\mathbf{X}_t^{t+i}$ and $\mathbf{X}_{t+j}^{t+i+j}$ have the same distribution.*

Our main result requires the method of blocking used by Yu [27, 28]. The purpose is to transform a sequence of dependent variables into subsequence of nearly IID ones. Consider a sample $\mathbf{X}_1^n$ from a stationary β-mixing sequence with density f. Let m_n and μ_n be non-negative integers such that $2m_n\mu_n = n$. Now divide $\mathbf{X}_1^n$ into $2\mu_n$ blocks, each of length m_n. Identify the blocks as follows:

$$U_j = \{X_i : 2(j-1)m_n + 1 \leq i \leq (2j-1)m_n\},$$
$$V_j = \{X_i : (2j-1)m_n + 1 \leq i \leq 2jm_n\}.$$

Let $\mathbf{U}$ be the entire sequence of odd blocks U_j, and let $\mathbf{V}$ be the sequence of even blocks V_j. Finally, let $\mathbf{U}'$ be a sequence of blocks which are independent of $\mathbf{X}_1^n$ but such that each block has the same distribution as a block from the original sequence:

$$U_j' \overset{D}{=} U_j \overset{D}{=} U_1. \tag{2}$$

The blocks $\mathbf{U}'$ are now an IID block sequence, so standard results apply. (See [28] for a more rigorous analysis of blocking.) With this structure, we can state our main result.

2.2 Results

Our main result emerges in two stages. First, we recognize that the distribution of a finite sample depends only on finite-dimensional distributions. This leads to an estimator of a finite-dimensional version of $\beta(a)$. Next, we let the finite-dimension increase to infinity with the size of the observed sample.

For positive integers t, d, and a, define

$$\beta^d(a) \equiv \left|\left|\mathbb{P}_{t-d+1}^{t} \otimes \mathbb{P}_{t+a}^{t+a+d-1} - \mathbb{P}_{t,a,d}\right|\right|_{TV}, \tag{3}$$

where $\mathbb{P}_{t,a,d}$ is the joint distribution of $(\mathbf{X}_{t-d+1}^{t}, \mathbf{X}_{t+a}^{t+a+d-1})$. Also, let $\widehat{f}^d$ be the d-dimensional histogram estimator of the joint density of d consecutive observations, and let $\widehat{f}_a^{2d}$ be the $2d$-dimensional histogram estimator of the joint density of two sets of d consecutive observations separated by a time points.

We construct an estimator of $\beta^d(a)$ based on these two histograms.[2] Define

$$\widehat{\beta}^d(a) \equiv \frac{1}{2}\int \left|\widehat{f}_a^{2d} - \widehat{f}^d \otimes \widehat{f}^d\right| \tag{4}$$

We show that, by allowing $d = d_n$ to grow with n, this estimator will converge on $\beta(a)$. This can be seen most clearly by bounding the ℓ^1-risk of the estimator with its estimation and approximation errors:

$$|\widehat{\beta}^d(a) - \beta(a)| \leq |\widehat{\beta}^d(a) - \beta^d(a)| + |\beta^d(a) - \beta(a)|.$$

The first term is the error of estimating $\beta^d(a)$ with a random sample of data. The second term is the non-stochastic error induced by approximating the infinite dimensional coefficient, $\beta(a)$, with its d-dimensional counterpart, $\beta^d(a)$.

Our first theorem in this section establishes consistency of $\widehat{\beta}^{d_n}(a)$ as an estimator of $\beta(a)$ for all β-mixing processes provided d_n increases at an appropriate rate. Theorem 2.4 gives finite sample bounds on the estimation error while some measure theoretic arguments contained in §4 show that the approximation error must go to zero as $d_n \to \infty$.

Theorem 2.3. *Let $\mathbf{X}_1^n$ be a sample from an arbitrary β-mixing process. Let $d_n = O(\exp\{W(\log n)\})$ where W is the Lambert W function.*[3] *Then $\widehat{\beta}^{d_n}(a) \xrightarrow{P} \beta(a)$ as $n \to \infty$.*

A finite sample bound for the approximation error is the first step to establishing consistency for $\widehat{\beta}^d(a)$. This result gives convergence rates for estimation of the finite dimensional mixing coefficient $\beta^d(a)$ and also for Markov processes of known order d, since in this case, $\beta^d(a) = \beta(a)$.

Theorem 2.4. *Consider a sample $\mathbf{X}_1^n$ from a stationary β-mixing process. Let μ_n and m_n be positive integers such that $2\mu_n m_n = n$ and $\mu_n \geq d > 0$. Then*

$$\begin{aligned}
&\mathbb{P}(|\widehat{\beta}^d(a) - \beta^d(a)| > \epsilon) \\
&\leq 2\exp\left\{-\frac{\mu_n \epsilon_1^2}{2}\right\} + 2\exp\left\{-\frac{\mu_n \epsilon_2^2}{2}\right\} \\
&+ 4(\mu_n - 1)\beta(m_n),
\end{aligned}$$

where $\epsilon_1 = \epsilon/2 - \mathbb{E}\left[\int |\widehat{f}^d - f^d|\right]$ and $\epsilon_2 = \epsilon - \mathbb{E}\left[\int |\widehat{f}_a^{2d} - f_a^{2d}|\right]$.

Consistency of the estimator $\widehat{\beta}^d(a)$ is guaranteed only for certain choices of m_n and μ_n. Clearly $\mu_n \to \infty$ and $\mu_n\beta(m_n) \to 0$ as $n \to \infty$ are necessary conditions. Consistency also requires convergence of the histogram estimators to the target densities. We leave the proof of this theorem for section 4. As an example to show that this bound can go to zero with proper choices of m_n and μ_n, the following corollary proves consistency for first order Markov processes. Consistency of the estimator for higher order Markov processes can be proven similarly. These processes are geometrically β-mixing as shown in e.g. Nummelin and Tuominen [19].

Corollary 2.5. *Let $\mathbf{X}_1^n$ be a sample from a first order Markov process with $\beta(a) = \beta^1(a) = O(r^a)$ for some $0 \leq r < 1$. Then under the conditions of Theorem 2.4, $\widehat{\beta}^1(a) \xrightarrow{P} \beta(a)$ at a rate of $o(\sqrt{n})$ up to a logarithmic factor.*

Proof. Recall that $n = 2\mu_n m_n$. Then,

$$\begin{aligned}
4(\mu_n - 1)\beta(m_n) &= 4\mu_n\beta(m_n) + 4\beta(m_n) \\
&= K_1\frac{n}{m_n}r^{m_n} + K_2 r^{m_n} \\
&\to 0
\end{aligned}$$

if $m_n = \Omega(\log n)$ for constants K_1 and K_2. But the exponential terms are

$$\exp\left\{-K_3\frac{n\epsilon_j^2}{m_n}\right\}$$

for $j = 1, 2$ and a constant K_3. Therefore, both exponential terms go to 0 as $n \to \infty$ for $m_n = o(n)$. Balancing the rates gives the optimal choice of $m_n = o(\sqrt{n})$ with corresponding rate of convergence (up to a logarithmic factor) of $o(\sqrt{n})$. □

[2] While it is clearly possible to replace histograms with other choices of density estimators (most notably kernel density estimators), histograms in this case are more convenient theoretically and computationally. See §5 for more details.

[3] The Lambert W function is defined as the (multivalued) inverse of $f(w) = w\exp\{w\}$. Thus, $O(\exp\{W(\log n)\})$ is bigger than $O(\log\log n)$ but smaller than $O(\log n)$. See for example Corless et al. [6].

Proving Theorem 2.4 requires showing the L^1 convergence of the histogram density estimator with β-mixing data. We do this in the next section.

3 L^1 convergence of histograms

Convergence of density estimators is thoroughly studied in the statistics and machine learning literature. Early papers on the L^∞ convergence of kernel density estimators (KDEs) include [26, 2, 22]; Freedman and Diaconis [10] look specifically at histogram estimators, and Yu [27] considered the L^∞ convergence of KDEs for β-mixing data and shows that the optimal IID rates can be attained. Devroye and Györfi [8] argue that L^1 is a more appropriate metric for studying density estimation, and Tran [23] proves L^1 consistency of KDEs under α- and β-mixing. As far as we are aware, ours is the first proof of L^1 convergence for histograms under β-mixing.

Additionally, the dimensionality of the target density is analogous to the order of the Markov approximation. Therefore, the convergence rates we give are asymptotic in the bandwidth h_n which shrinks as n increases, but also in the dimension d which increases with n. Even under these asymptotics, histogram estimation in this sense is not a high dimensional problem. The dimension of the target density considered here is on the order of $\exp\{W(\log n)\}$, a rate somewhere between $\log n$ and $\log\log n$.

Theorem 3.1. *If $\widehat{f}$ is the histogram estimator based on a (possibly vector valued) sample $\mathbf{X}_1^n$ from a β-mixing sequence with stationary density f, then for all $\epsilon > \mathbb{E}\left[\int |\widehat{f} - f|\right]$,*

$$\mathbb{P}\left(\int |\widehat{f} - f| > \epsilon\right) \leq 2\exp\left\{-\frac{\mu_n \epsilon_1^2}{2}\right\} + 2(\mu_n - 1)\beta(m_n) \qquad (5)$$

where $\epsilon_1 = \epsilon - \mathbb{E}\left[\int |\widehat{f} - f|\right]$.

To prove this result, we use the blocking method of Yu [28] to transform the dependent β-mixing into a sequence of nearly independent blocks. We then apply McDiarmid's inequality to the blocks to derive asymptotics in the bandwidth of the histogram as well as the dimension of the target density. For completeness, we state Yu's blocking result and McDiarmid's inequality before proving the doubly asymptotic histogram convergence for IID data. Combining these lemmas allows us to derive rates of convergence for histograms based on β-mixing inputs.

Lemma 3.2 (Lemma 4.1 in Yu [28])**.** *Let ϕ be a measurable function with respect to the block sequence $\mathbf{U}$ uniformly bounded by M. Then,*

$$|\mathbb{E}[\phi] - \tilde{\mathbb{E}}[\phi]| \leq M\beta(m_n)(\mu_n - 1), \qquad (6)$$

where the first expectation is with respect to the dependent block sequence, $\mathbf{U}$, and $\tilde{\mathbb{E}}$ is with respect to the independent sequence, $\mathbf{U}'$.

This lemma essentially gives a method of applying IID results to β-mixing data. Because the dependence decays as we increase the separation between blocks, widely spaced blocks are nearly independent of each other. In particular, the difference between expectations over these nearly independent blocks and expectations over blocks which are actually independent can be controlled by the β-mixing coefficient.

Lemma 3.3 (McDiarmid Inequality [14])**.** *Let $X_1, \ldots, X_n$ be independent random variables, with X_i taking values in a set A_i for each i. Suppose that the measurable function $f : \prod A_i \to \mathbb{R}$ satisfies*

$$|f(\mathbf{x}) - f(\mathbf{x}')| \leq c_i$$

whenever the vectors $\mathbf{x}$ and $\mathbf{x}'$ differ only in the i^{th} coordinate. Then for any $\epsilon > 0$,

$$\mathbb{P}(f - \mathbb{E}f > \epsilon) \leq \exp\left\{-\frac{2\epsilon^2}{\sum c_i^2}\right\}.$$

Lemma 3.4. *For an IID sample $X_1, \ldots, X_n$ from some density f on $\mathbb{R}^d$,*

$$\mathbb{E}\int |\widehat{f} - \mathbb{E}\widehat{f}| dx = O\left(1/\sqrt{nh_n^d}\right) \qquad (7)$$

$$\int |\mathbb{E}\widehat{f} - f| dx = O(dh_n) + O(d^2 h_n^2), \qquad (8)$$

where $\widehat{f}$ is the histogram estimate using a grid with sides of length h_n.

Proof of Lemma 3.4. Let p_j be the probability of falling into the j^{th} bin B_j. Then,

$$\begin{aligned}
\mathbb{E}\int |\widehat{f} - \mathbb{E}\widehat{f}| &= h_n^d \sum_{j=1}^J \mathbb{E}\left|\frac{1}{nh_n^d}\sum_{i=1}^n I(X_i \in B_j) - \frac{p_j}{h^d}\right| \\
&\leq h_n^d \sum_{j=1}^J \frac{1}{nh_n^d}\sqrt{\mathbb{V}\left[\sum_{i=1}^n I(X_i \in B_j)\right]} \\
&= h_n^d \sum_{j=1}^J \frac{1}{nh_n^d}\sqrt{np_j(1-p_j)} \\
&= \frac{1}{\sqrt{n}}\sum_{j=1}^J \sqrt{p_j(1-p_j)} \\
&= O(n^{-1/2})O(h_n^{-d/2}) = O\left(1/\sqrt{nh_n^d}\right).
\end{aligned}$$

For the second claim, consider the bin B_j centered at $\mathbf{c}$. Let I be the union of all bins B_j. Assume the following:

1. $f \in L_2$ and f is absolutely continuous on I, with a.e. partial derivatives $f_i = \frac{\partial}{\partial y_i} f(\mathbf{y})$
2. $f_i \in L_2$ and f_i is absolutely continuous on I, with a.e. partial derivatives $f_{ik} = \frac{\partial}{\partial y_k} f_i(\mathbf{y})$
3. $f_{ik} \in L_2$ for all i, k.

Using a Taylor expansion

$$f(\mathbf{x}) = f(\mathbf{c}) + \sum_{i=1}^{d} (x_i - c_i) f_i(\mathbf{c}) + O(d^2 h_n^2),$$

where $f_i(\mathbf{y}) = \frac{\partial}{\partial y_i} f(\mathbf{y})$. Therefore, p_j is given by

$$p_j = \int_{B_j} f(x)dx = h_n^d f(c) + O(d^2 h_n^{d+2})$$

since the integral of the second term over the bin is zero. This means that for the j^{th} bin,

$$\begin{aligned}\mathbb{E}\widehat{f}_n(x) - f(x) &= \frac{p_j}{h_n^d} - f(x)\\ &= -\sum_{i=1}^{d} (x_i - c_i) f_i(\mathbf{c}) + O(d^2 h_n^2).\end{aligned}$$

Therefore,

$$\begin{aligned}&\int_{B_j} \left|\mathbb{E}\widehat{f}_n(x) - f(x)\right|\\ &= \int_{B_j} \left| -\sum_{i=1}^{d} (x_i - c_i) f_i(\mathbf{c}) + O(d^2 h_n^2)\right|\\ &\le \int_{B_j} \left| -\sum_{i=1}^{d} (x_i - c_i) f_i(\mathbf{c})\right| + \int_{B_j} O(d^2 h^2)\\ &= \int_{B_j} \left| \sum_{i=1}^{d} (x_i - c_i) f_i(\mathbf{c})\right| + O(d^2 h_n^{2+d})\\ &= O(d h_n^{d+1}) + O(d^2 h_n^{2+d})\end{aligned}$$

Since each bin is bounded, we can sum over all J bins. The number of bins is $J = h_n^{-d}$ by definition, so

$$\begin{aligned}&\int |\mathbb{E}\widehat{f}_n(x) - f(x)|dx\\ &= O(h_n^{-d}) \left(O(d h_n^{d+1}) + O(d^2 h_n^{2+d})\right)\\ &= O(d h_n) + O(d^2 h_n^2).\end{aligned}$$

□

We can now prove the main result of this section.

Proof of Theorem 3.1. Let g be the L^1 loss of the histogram estimator, $g = \int |f - \widehat{f}_n|$. Here $\widehat{f}_n(x) = \frac{1}{n h_n^d} \sum_{i=1}^n I(X_i \in B_j(x))$ where $B_j(x)$ is the bin containing x. Let $\widehat{f}_{\mathbf{U}}$, $\widehat{f}_{\mathbf{V}}$, and $\widehat{f}_{\mathbf{U}'}$ be histograms based on the block sequences $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{U}'$ respectively. Clearly $\widehat{f}_n = \frac{1}{2}(\widehat{f}_{\mathbf{U}} + \widehat{f}_{\mathbf{V}})$. Now,

$$\begin{aligned}\mathbb{P}(g > \epsilon) &= \mathbb{P}\left(\int |f - \widehat{f}_n| > \epsilon\right)\\ &= \mathbb{P}\left(\int \left|\frac{f - \widehat{f}_{\mathbf{U}}}{2} + \frac{f - \widehat{f}_{\mathbf{V}}}{2}\right| > \epsilon\right)\\ &\le \mathbb{P}\left(\frac{1}{2}\int |f - \widehat{f}_{\mathbf{U}}| + \frac{1}{2}\int |f - \widehat{f}_{\mathbf{V}}| > \epsilon\right)\\ &= \mathbb{P}(g_{\mathbf{U}} + g_{\mathbf{V}} > 2\epsilon)\\ &\le \mathbb{P}(g_{\mathbf{U}} > \epsilon) + \mathbb{P}(g_{\mathbf{V}} > \epsilon)\\ &= 2\mathbb{P}(g_{\mathbf{U}} - \mathbb{E}[g_{\mathbf{U}}] > \epsilon - \mathbb{E}[g_{\mathbf{U}}])\\ &= 2\mathbb{P}(g_{\mathbf{U}} - \mathbb{E}[g_{\mathbf{U}'}] > \epsilon - \mathbb{E}[g_{\mathbf{U}'}])\\ &= 2\mathbb{P}(g_{\mathbf{U}} - \mathbb{E}[g_{\mathbf{U}'}] > \epsilon_1),\end{aligned}$$

where $\epsilon_1 = \epsilon - \mathbb{E}[g_{\mathbf{U}'}]$. Here,

$$\mathbb{E}[g_{\mathbf{U}'}] \le \tilde{\mathbb{E}} \int |\widehat{f}_{\mathbf{U}'} - \tilde{\mathbb{E}}\widehat{f}_{\mathbf{U}'}|dx + \int |\tilde{\mathbb{E}}\widehat{f}_{\mathbf{U}'} - f|dx,$$

so by Lemma 3.4, as long as for $\mu_n \to \infty$, $h_n \downarrow 0$ and $\mu_n h_n^d \to \infty$, then for all ϵ there exists $n_0(\epsilon)$ such that for all $n > n_0(\epsilon)$, $\epsilon > \mathbb{E}[g] = \mathbb{E}[g_{\mathbf{U}'}]$. Now applying Lemma 3.2 to the expectation of the indicator of the event $\{g_{\mathbf{U}} - \mathbb{E}[g_{\mathbf{U}'}] > \epsilon_1\}$ gives

$$\begin{aligned}2\mathbb{P}(g_{\mathbf{U}} - \mathbb{E}[g_{\mathbf{U}'}] > \epsilon_1) &\le 2\mathbb{P}(g_{\mathbf{U}'} - \mathbb{E}[g_{\mathbf{U}'}] > \epsilon_1)\\ &\quad + 2(\mu_n - 1)\beta(m_n)\end{aligned}$$

where the probability on the right is for the σ-field generated by the independent block sequence $\mathbf{U}'$. Since these blocks are independent, showing that $g_{\mathbf{U}'}$ satisfies the bounded differences requirement allows for the application of McDiarmid's inequality 3.3 to the blocks. For any two block sequences $u'_1, \ldots, u'_{\mu_n}$ and $\bar{u}'_1, \ldots, \bar{u}'_{\mu_n}$ with $u'_\ell = \bar{u}'_\ell$ for all $\ell \neq j$, then

$$\begin{aligned}&|g_{\mathbf{U}'}(u'_1, \ldots, u'_{\mu_n}) - g_{\mathbf{U}'}(\bar{u}'_1, \ldots, \bar{u}'_{\mu_n})|\\ &= \left|\int |\widehat{f}(y; u'_1, \ldots, u'_{\mu_n}) - f(y)|dy\right.\\ &\quad \left. - \int |\widehat{f}(y; \bar{u}'_1, \ldots, \bar{u}'_{\mu_n}) - f(y)|dy\right|\\ &\le \int |\widehat{f}(y; u'_1, \ldots, u'_{\mu_n}) - \widehat{f}(y; \bar{u}'_1, \ldots, \bar{u}'_{\mu_n})|dy\\ &= \frac{2}{\mu_n h_n^d} h_n^d = \frac{2}{\mu_n}.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{P}(g > \epsilon) &\le 2\mathbb{P}(g_{\mathbf{U}'} - \mathbb{E}[g_{\mathbf{U}'}] > \epsilon_1) + 2(\mu_n - 1)\beta(m_n)\\ &\le 2\exp\left\{-\frac{\mu_n \epsilon_1^2}{2}\right\} + 2(\mu_n - 1)\beta(m_n).\end{aligned}$$

□

4 Proofs

The proof of Theorem 2.4 relies on the triangle inequality and the relationship between total variation distance and the L^1 distance between densities.

Proof of Theorem 2.4. For any probability measures ν and λ defined on the same probability space with associated densities f_ν and f_λ with respect to some dominating measure π,

$$||\nu - \lambda||_{TV} = \frac{1}{2}\int |f_\nu - f_\lambda| d(\pi).$$

Let P be the d-dimensional stationary distribution of the d^{th} order Markov process, i.e. $P = \mathbb{P}^{t}_{t-d+1} = \mathbb{P}^{t+a+d-1}_{t+a}$ in the notation of equation 3. Let $\mathbb{P}_{a,d}$ be the joint distribution of the bivariate random process created by the initial process and itself separated by a time steps. By the triangle inequality, we can upper bound $\beta^d(a)$ for any $d = d_n$. Let $\widehat{P}$ and $\widehat{\mathbb{P}}_{a,d}$ be the distributions associated with histogram estimators $\widehat{f}^d$ and $\widehat{f}^{2d}_a$ respectively. Then,

$$\begin{aligned}
\beta^d(a) &= ||P\otimes P - \mathbb{P}_{a,d}||_{TV} \\
&= \Big|\Big|P\otimes P - \widehat{P}\otimes\widehat{P} + \widehat{P}\otimes\widehat{P} \\
&\quad - \widehat{\mathbb{P}}_{a,d} + \widehat{\mathbb{P}}_{a,d} - \mathbb{P}_{a,d}\Big|\Big|_{TV} \\
&\leq \Big|\Big|P\otimes P - \widehat{P}\otimes\widehat{P}\Big|\Big|_{TV} + \Big|\Big|\widehat{P}\otimes\widehat{P} - \widehat{\mathbb{P}}_{a,d}\Big|\Big|_{TV} \\
&\quad + \Big|\Big|\widehat{\mathbb{P}}_{a,d} - \mathbb{P}_{a,d}\Big|\Big|_{TV} \\
&\leq 2\Big|\Big|P - \widehat{P}\Big|\Big|_{TV} + \Big|\Big|\widehat{P}\otimes\widehat{P} - \widehat{\mathbb{P}}_{a,d}\Big|\Big|_{TV} \\
&\quad + \Big|\Big|\widehat{\mathbb{P}}_{a,d} - \mathbb{P}_{a,d}\Big|\Big|_{TV} \\
&= \int |f^d - \widehat{f}^d| + \frac{1}{2}\int |\widehat{f}^d\otimes\widehat{f}^d - \widehat{f}^{2d}_a| \\
&\quad + \frac{1}{2}\int |f^{2d}_a - \widehat{f}^{2d}_a|
\end{aligned}$$

where $\frac{1}{2}\int|\widehat{f}^d\otimes\widehat{f}^d - \widehat{f}^{2d}_a|$ is our estimator $\widehat{\beta}^d(a)$ and the remaining terms are the L^1 distance between a density estimator and the target density. Thus,

$$\beta^d(a) - \widehat{\beta}^d(a) \leq \int |f^d - \widehat{f}^d| + \frac{1}{2}\int |f^{2d}_a - \widehat{f}^{2d}_a|.$$

A similar argument starting from $\beta^d(a) = ||P\otimes P - \mathbb{P}_{a,d}||_{TV}$ shows that

$$\beta^d(a) - \widehat{\beta}^d(a) \geq -\int |f^d - \widehat{f}^d| - \frac{1}{2}\int |f^{2d}_a - \widehat{f}^{2d}_a|,$$

so we have that

$$\left|\beta^d(a) - \widehat{\beta}^d(a)\right| \leq \int |f^d - \widehat{f}^d| + \frac{1}{2}\int |f^{2d}_a - \widehat{f}^{2d}_a|.$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\left(\left|\beta^d(a) - \widehat{\beta}^d(a)\right| > \epsilon\right) \\
&\leq \mathbb{P}\left(\int |f^d - \widehat{f}^d| + \frac{1}{2}\int |f^{2d}_a - \widehat{f}^{2d}_a| > \epsilon\right) \\
&\leq \mathbb{P}\left(\int |f^d - \widehat{f}^d| > \frac{\epsilon}{2}\right) + \mathbb{P}\left(\frac{1}{2}\int |f^{2d}_a - \widehat{f}^{2d}_a| > \frac{\epsilon}{2}\right) \\
&\leq 2\exp\left\{-\frac{\mu_n\epsilon_1^2}{2}\right\} + 2\exp\left\{-\frac{\mu_n\epsilon_2^2}{2}\right\} \\
&+ 4(\mu_n - 1)\beta(m_n),
\end{aligned}$$

where $\epsilon_1 = \epsilon/2 - \mathbb{E}\left[\int|\widehat{f}^d - f^d|\right]$ and $\epsilon_2 = \epsilon - \mathbb{E}\left[\int|\widehat{f}^{2d}_a - f^{2d}_a|\right]$. □

The proof of Theorem 2.3 requires two steps which are given in the following Lemmas. The first specifies the histogram bandwidth h_n and the rate at which d_n (the dimensionality of the target density) goes to infinity. If the dimensionality of the target density were fixed, we could achieve rates of convergence similar to those for histograms based on IID inputs. However, we wish to allow the dimensionality to grow with n, so the rates are much slower as shown in the following lemma.

Lemma 4.1. *For the histogram estimator in Lemma 3.4, let*

$$\begin{aligned}
d_n &\sim \exp\{W(\log n)\}, \\
h_n &\sim n^{-k_n},
\end{aligned}$$

with

$$k_n = \frac{W(\log n) + \frac{1}{2}\log n}{\log n\left(\frac{1}{2}\exp\{W(\log n)\} + 1\right)}.$$

These choices lead to the optimal rate of convergence.

Proof. Let $h_n = n^{-k_n}$ for some k_n to be determined. Then we want $n^{-1/2}h_n^{-d_n/2} = n^{(k_n d_n - 1)/2} \to 0$, $d_n h_n = d_n n^{-k} \to 0$, and $d_n^2 h_n^2 = d_n^2 n^{-2k} \to 0$ all as $n \to \infty$. Call these A, B, and C. Taking A and B first gives

$$\begin{aligned}
n^{(k_n d_n - 1)/2} &\sim d_n n^{-k_n} \\
\Rightarrow \frac{1}{2}(k_n d_n - 1)\log n &\sim \log d_n - k_n \log n \\
\Rightarrow k_n \log n\left(\frac{1}{2}d_n + 1\right) &\sim \log d_n + \frac{1}{2}\log n \\
\Rightarrow k_n &\sim \frac{\log d_n + \frac{1}{2}\log n}{\log n\left(\frac{1}{2}d_n + 1\right)}. \qquad (9)
\end{aligned}$$

Similarly, combining A and C gives

$$k_n \sim \frac{2\log d_n + \frac{1}{2}\log n}{\log n\left(\frac{1}{2}d_n + 2\right)}. \qquad (10)$$

Equating (9) and (10) and solving for d_n gives

$$\Rightarrow d_n \sim \exp\{W(\log n)\}$$

where $W(\cdot)$ is the Lambert W function. Plugging back into (9) gives that

$$h_n = n^{-k_n}$$

where

$$k_n = \frac{W(\log n) + \frac{1}{2}\log n}{\log n\left(\frac{1}{2}\exp\{W(\log n)\} + 1\right)}.$$

□

It is also necessary to show that as d grows, $\beta^d(a) \to \beta(a)$. We now prove this result.

Lemma 4.2. *$\beta^d(a)$ converges to $\beta(a)$ as $d \to \infty$.*

Proof. By stationarity, the supremum over t is unnecessary in Definition 2.1, so without loss of generality, let $t = 0$. Let $\mathbb{P}^0_{-\infty}$ be the distribution on $\sigma^0_{-\infty} = \sigma(\ldots, X_{-1}, X_0)$, and let $\mathbb{P}^\infty_a$ be the distribution on $\sigma^\infty_{a+1} = \sigma(X_{a+1}, X_{a+2}, \ldots)$. Let $\mathbb{P}_a$ be the distribution on $\sigma = \sigma^0_{-\infty} \otimes \sigma^\infty_{a+1}$ (the product sigma-field). Then we can rewrite Definition 2.1 using this notation as

$$\beta(a) = \sup_{C\in\sigma} |\mathbb{P}_a(C) - [\mathbb{P}^0_{-\infty} \otimes \mathbb{P}^\infty_{a+1}](C)|.$$

Let σ^0_{-d+1} and σ^{a+d}_{a+1} be the sub-σ-fields of $\sigma^0_{-\infty}$ and σ^∞_{a+1} consisting of the d-dimensional cylinder sets for the d dimensions closest together. Let σ^d be the product σ-field of these two. Then we can rewrite $\beta^d(a)$ as

$$\beta^d(a) = \sup_{C\in\sigma^d} ||\mathbb{P}_a(C) - [\mathbb{P}^0_{-\infty} \otimes \mathbb{P}^\infty_{a+1}](C)|. \qquad (11)$$

As such $\beta^d(a) \leq \beta(a)$ for all a and d. We can rewrite (11) in terms of finite-dimensional marginals:

$$\beta^d(a) = \sup_{C\in\sigma^d} |\mathbb{P}_{a,d}(C) - [\mathbb{P}^0_{-d+1} \otimes \mathbb{P}^{a+d}_{a+1}](C)|,$$

where $\mathbb{P}_{a,d}$ is the restriction of $\mathbb{P}$ to $\sigma(X_{-d+1}, \ldots, X_0, X_{a+1}, \ldots, X_{a+d})$. Because of the nested nature of these sigma-fields, we have

$$\beta^{d_1}(a) \leq \beta^{d_2}(a) \leq \beta(a)$$

for all finite $d_1 \leq d_2$. Therefore, for fixed a, $\{\beta^d(a)\}^\infty_{d=1}$ is a monotone increasing sequence which is bounded above, and it converges to some limit $L \leq \beta(a)$. To show that $L = \beta(a)$ requires some additional steps.

Let $R = \mathbb{P}_a - [\mathbb{P}^0_{-\infty} \otimes \mathbb{P}^\infty_a]$, which is a signed measure on σ. Let $R^d = \mathbb{P}_{a,d} - [\mathbb{P}^0_{-d} \otimes \mathbb{P}^{a+d}_a]$, which is a signed measure on σ^d. Decompose R into positive and negative parts as $R = Q^+ - Q^-$ and similarly for $R^d = Q^{+d} - Q^{-d}$. Notice that since R^d is constructed using the marginals of $\mathbb{P}$, then $R(E) = R^d(E)$ for all $E \in \sigma^d$. Now since R is the difference of probability measures, we must have that

$$\begin{aligned} 0 = R(\Omega) &= Q^+(\Omega) - Q^-(\Omega) \\ &= Q^+(D) + Q^+(D^c) - Q^-(D) - Q^-(D^c) \qquad (12) \end{aligned}$$

for all $D \in \sigma$.

Define $Q = Q^+ + Q^-$. Let $\epsilon > 0$. Let $C \in \sigma$ be such that

$$Q(C) = \beta(a) = Q^+(C) = Q^-(C^c). \qquad (13)$$

Such a set C is guaranteed by the Hahn decomposition theorem (letting C^* be a set which attains the supremum in (11), we can throw away any subsets with negative R measure) and (12) assuming without loss of generality that $\mathbb{P}_a(C) > [\mathbb{P}^0_{-\infty} \otimes \mathbb{P}^\infty_a](C)$. We can use the field $\sigma_f = \bigcup_d \sigma^d$ to approximate σ in the sense that, for all ϵ, we can find $A \in \sigma_f$ such that $Q(A\Delta C) < \epsilon/2$ (see Theorem D in Halmos [11, §13] or Lemma A.24 in Schervish [21]). Now,

$$\begin{aligned} Q(A\Delta C) &= Q(A \cap C^c) + Q(C \cap A^c) \\ &= Q^-(A \cap C^c) + Q^+(C \cap A^c) \end{aligned}$$

by (13) since $A \cap C^c \subseteq C^c$ and $C \cap A^c \subseteq C$. Therefore, since $Q(A\Delta C) < \epsilon/2$, we have

$$\begin{aligned} &Q^-(A \cap C^c) \leq \epsilon/2 \qquad (14) \\ &Q^+(A^c \cap C) \leq \epsilon/2. \end{aligned}$$

Also,

$$\begin{aligned} Q(C) &= Q(A \cap C) + Q(A^c \cap C) \\ &= Q^+(A \cap C) + Q^+(A^c \cap C) \\ &\leq Q^+(A) + \epsilon/2 \end{aligned}$$

since $A \cap C$ and $A^c \cap C$ are contained in C and $A \cap C \subseteq A$. Therefore

$$Q^+(A) \geq Q(C) - \epsilon/2.$$

Similarly,

$$Q^-(A) = Q^-(A \cap C) + Q^-(A \cap C^c) \leq 0 + \epsilon/2 = \epsilon/2$$

since $A \cap C \subseteq C$ and $Q^-(C) = 0$ by (14). Finally,

$$\begin{aligned} Q^{+d}(A) \geq Q^{+d}(A) - Q^{-d}(A) &= R^d(A) \\ &= R(A) = Q^+(A) - Q^-(A) \\ &\geq Q(C) - \epsilon/2 - \epsilon/2 = Q(C) - \epsilon \\ &= \beta(a) - \epsilon. \end{aligned}$$

And since $\beta^d(a) \geq Q^{+d}(A)$, we have that for all $\epsilon > 0$ there exists d such that for all $d_1 > d$,

$$\begin{aligned}\beta^{d_1}(a) \geq \beta^d(a) &\geq Q^{+d}(A)\\ &\geq \beta(a) - \epsilon.\end{aligned}$$

Thus, we must have that $L = \beta(a)$, so that $\beta^d(a) \to \beta(a)$ as desired. □

Proof of Theorem 2.3. By the triangle inequality,

$$|\widehat{\beta}^{d_n}(a) - \beta(a)| \leq |\widehat{\beta}^{d_n}(a) - \beta^{d_n}(a)| + |\beta^{d_n}(a) - \beta(a)|.$$

The first term on the right is bounded by the result in Theorem 2.4, where we have shown that $d_n = O(\exp\{W(\log n)\})$ is slow enough for the histogram estimator to remain consistent. That $\beta^{d_n}(a) \xrightarrow{d_n\to\infty} \beta(a)$ follows from Lemma 4.2. □

5 Discussion

We have shown that our estimator of the β-mixing coefficients is consistent for the true coefficients $\beta(a)$ under some conditions on the data generating process. There are numerous results in the statistics and machine learning literatures which assume knowledge of the β-mixing coefficients, yet as far as we know, this is the first estimator for them. An ability to estimate these coefficients will allow researchers to apply existing results to dependent data without the need to arbitrarily assume their values. Despite the obvious utility of this estimator, as a consequence of its novelty, it comes with a number of potential extensions which warrant careful exploration as well as some drawbacks.

The reader will note that Theorem 2.3 does not provide a convergence rate. The rate in Theorem 2.4 applies only to the difference between $\hat{\beta}^d(a)$ and $\beta^d(a)$. In order to provide a rate in Theorem 2.3, we would need a better understanding of the non-stochastic convergence of $\beta^d(a)$ to $\beta(a)$. It is not immediately clear that this quantity can converge at any well-defined rate. In particular, it seems likely that the rate of convergence depends on the tail of the sequence $\{\beta(a)\}_{a=1}^{\infty}$.

Several other mixing and weak-dependence coefficients also have a total-variation flavor, perhaps most notably α-mixing [9, 7, 4]. None of them have estimators, and the same trick might well work for them, too.

The use of histograms rather than kernel density estimators for the joint and marginal densities is surprising and perhaps not ultimately necessary. As mentioned above, Tran [23] proved that KDEs are consistent for estimating the stationary density of a time series with β-mixing inputs, so perhaps one could replace the histograms in our estimator with KDEs. However, this would need an analogue of the double asymptotic results proven for histograms in Lemma 3.4. In particular, we need to estimate increasingly higher dimensional densities as $n \to \infty$. This does not cause a problem of small-n-large-d since d is chosen as a function of n, however it will lead to increasingly higher dimensional integration. For histograms, the integral is always trivial, but in the case of KDEs, the numerical accuracy of the integration algorithm becomes increasingly important. This issue could swamp any efficiency gains obtained through the use of kernels. However, this question certainly warrants further investigation.

The main drawback of an estimator based on a density estimate is its complexity. The mixing coefficients are functionals of the joint and marginal distributions derived from the stochastic process $\mathbf{X}$, however, it is unsatisfying to estimate densities and solve integrals in order to estimate a single number. Vapnik's main principle for solving problems using a restricted amount of information is

> When solving a given problem, try to avoid solving a more general problem as an intermediate step [24, p. 30].

This principle is clearly violated here, but perhaps our seed will precipitate a more aesthetically pleasing solution.

Acknowledgements

The authors wish to thank Darren Homrighausen and two anonymous reviewers for helpful comments and the Institute for New Economic Thinking for supporting this research.

References

[1] Baraud, Y., Comte, F., and Viennet, G. (2001), “Adaptive estimation in autoregression or β-mixing regression via model selection,” *Annals of statistics*, 29, 839–875.

[2] Bickel, P. and Rosenblatt, M. (1973), “On Some Global Measures of the Deviations of Density Function Estimates,” *The Annals of Statistics*, 1, 1071–1095.

[3] Bousquet, O. and Elisseeff, A. (2002), “Stability and Generalization,” *The Journal of Machine Learning Research*, 2, 499–526.

[4] Bradley, R. C. (2005), “Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions,” *Probability Surveys*, 2, 107–144.

[5] Carrasco, M. and Chen, X. (2002), "Mixing and Moment Properties of Various GARCH and Stochastic Volatility Models," *Econometric Theory*, 18, 17–39.

[6] Corless, R., Gonnet, G., Hare, D., Jeffrey, D., and Knuth, D. (1996), "On the Lambert W Function," *Advances in Computational Mathematics*, 5, 329–359.

[7] Dedecker, J., Doukhan, P., Lang, G., Leon R., J. R., Louhichi, S., and Prieur, C. (2007), *Weak Dependence: With Examples and Applications*, vol. 190 of *Lecture Notes in Statistics*, Springer Verlag, New York.

[8] Devroye, L. and Györfi, L. (1985), *Nonparametric Density Estimation: The L_1 View*, Wiley, New York.

[9] Doukhan, P. (1994), *Mixing: Properties and Examples*, vol. 85 of *Lecture Notes in Statistics*, Springer Verlag, New York.

[10] Freedman, D. and Diaconis, P. (1981), "On the Maximum Deviation Between the Histogram and the Underlying Density," *Probability Theory and Related Fields*, 58, 139–167.

[11] Halmos, P. (1974), *Measure Theory*, Graduate Texts in Mathematics, Springer-Verlag, New York.

[12] Karandikar, R. L. and Vidyasagar, M. (2009), "Probably Approximately Correct Learning with Beta-Mixing Input Sequences," submitted for publication.

[13] Lozano, A., Kulkarni, S., and Schapire, R. (2006), "Convergence and Consistency of Regularized Boosting Algorithms with Stationary Beta-Mixing Observations," *Advances in Neural Information Processing Systems*, 18, 819.

[14] McDiarmid, C. (1989), "On the Method of Bounded Differences," in *Surveys in Combinatorics*, ed. J. Siemons, vol. 141 of *London Mathematical Society Lecture Note Series*, pp. 148–188, Cambridge University Press.

[15] Meir, R. (2000), "Nonparametric Time Series Prediction Through Adaptive Model Selection," *Machine Learning*, 39, 5–34.

[16] Mohri, M. and Rostamizadeh, A. (2010), "Stability Bounds for Stationary φ-mixing and β-mixing Processes," *Journal of Machine Learning Research*, 11, 789–814.

[17] Mokkadem, A. (1988), "Mixing properties of ARMA processes," *Stochastic processes and their applications*, 29, 309–315.

[18] Nobel, A. (2006), "Hypothesis Testing for Families of Ergodic Processes," *Bernoulli*, 12, 251–269.

[19] Nummelin, E. and Tuominen, P. (1982), "Geometric Ergodicity of Harris Recurrent Markov Chains with Applications to Renewal Theory," *Stochastic Processes and Their Applications*, 12, 187–202.

[20] Ralaivola, L., Szafranski, M., and Stempfel, G. (2010), "Chromatic PAC-Bayes Bounds for Non-IID Data: Applications to Ranking and Stationary β-Mixing Processes," *Journal of Machine Learning Research*, 11, 1927–1956.

[21] Schervish, M. (1995), *Theory of Statistics*, Springer Series in Statistics, Springer Verlag, New York.

[22] Silverman, B. (1978), "Weak and Strong Uniform Consistency of the Kernel Estimate of a Density and its Derivatives," *The Annals of Statistics*, 6, 177–184.

[23] Tran, L. (1989), "The L_1 Convergence of Kernel Density Estimates under Dependence," *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, 17, 197–208.

[24] Vapnik, V. (2000), *The Nature of Statistical Learning Theory*, Statistics for Engineering and Information Science, Springer Verlag, New York, 2nd edn.

[25] Vidyasagar, M. (1997), *A Theory of Learning and Generalization: With Applications to Neural Networks and Control Systems*, Springer Verlag, Berlin.

[26] Woodroofe, M. (1967), "On the Maximum Deviation of the Sample Density," *The Annals of Mathematical Statistics*, 38, 475–481.

[27] Yu, B. (1993), "Density Estimation in the L_∞ Norm for Dependent Data with Applications to the Gibbs Sampler," *Annals of Statistics*, 21, 711–735.

[28] Yu, B. (1994), "Rates of Convergence for Empirical Processes of Stationary Mixing Sequences," *The Annals of Probability*, 22, 94–116.