Estimating beta-mixing coefficients

Daniel J. McDonald, Cosma Rohilla Shalizi and Mark Schervish
Department of Statistics, Carnegie Mellon University

Introduction

- Statistical learning for time series assumes asymptotic independence or "mixing"
- Mixing behavior assumed known
- Processes known to be β-mixing:
  - Independent RVs
  - Markov processes
  - ???
- No way to estimate

Literature using β-mixing

- Vidyasagar (1997) – β-mixing "just right" for extension of IID results
- Meir (ML 2000) – GEBs for nonparametric methods
- Lozano et al. (NIPS 2006) – boosting
- Karandikar and Vidyasagar (2009) – PAC algorithms
- Mohri and Rostamizadeh (NIPS 2008, JMLR 2010) – Rademacher complexity and stability bounds

Definitions

- **(β-mixing)** For each positive integer a, the the *coefficient of absolute regularity*, or β*-mixing coefficient*, $\beta(a)$, is
$$\beta(a) \equiv \sup_t \left|\left|\mathbb{P}^t_{-\infty} \otimes \mathbb{P}^{\infty}_{t+a} - \mathbb{P}_{t,a}\right|\right|_{TV}$$
where $||\cdot||_{TV}$ is the total variation norm, and $\mathbb{P}_{t,a}$ is the joint distribution of $(\mathbf{X}^t_{-\infty}, \mathbf{X}^{\infty}_{t+a})$. A stochastic process is said to be *absolutely regular*, or β*-mixing*, if $\beta(a) \to 0$ as $a \to \infty$.
- **(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 $\mathbf{X}^{t+i}_t$ and $\mathbf{X}^{t+i+j}_{t+j}$ have the same distribution.
- **(Finite dimensional coefficients)** For positive integers t, d, and a, define
$$\beta^d(a) = \left|\left|\mathbb{P}^t_{t-d+1} \otimes \mathbb{P}^{t+a+d-1}_{t+a} - \mathbb{P}_{t,a,d}\right|\right|_{TV},$$
where $\mathbb{P}_{t,a,d}$ is the joint distribution of $(\mathbf{X}^t_{t-d+1}, \mathbf{X}^{t+a+d-1}_{t+a})$.

Estimator

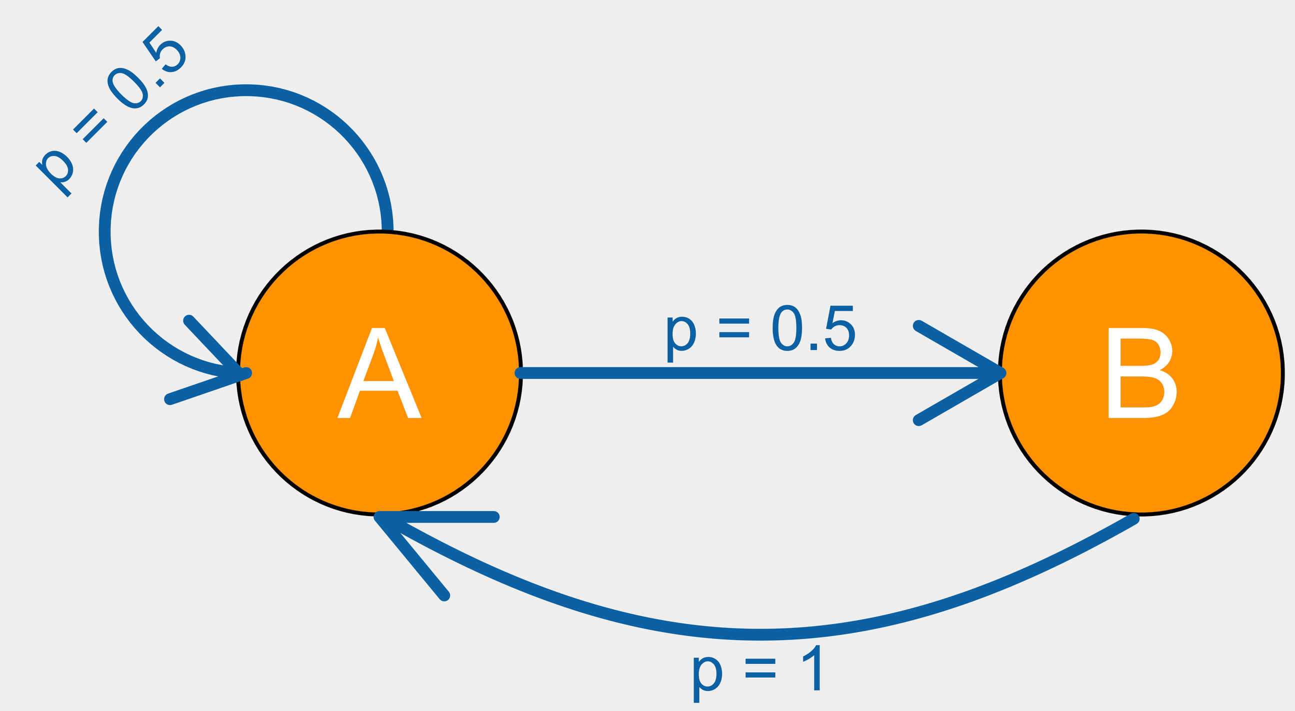
$$\widehat{\beta}^d(a) = \frac{1}{2}\int \left|\widehat{f}^{2d}_a - \widehat{f}^d \otimes \widehat{f}^d\right|,$$
where $\widehat{f}^{2d}_a$ is a histogram based on 2 length d sequences separated by $a-1$ points and $\widehat{f}^d$ is a histogram based on a length d sequence.

Performance evaluation (theory)

- Decompose the ℓ_1 risk of the estimator
$$|\widehat{\beta}^d(a) - \beta(a)| \leq |\widehat{\beta}^d(a) - \beta^d(a)| + |\beta^d(a) - \beta(a)|.$$
- Estimation error goes to zero
  - Can use McDiarmid type results
  - Speed depends on $\beta(a)$
  - If $\mathbf{X}$ is a Markov chain, then $\beta(a) = o(a^{-r})$
  - In this case $|\widehat{\beta}^d(a) - \beta^d(a)| = o(n^{r/(1+r)})$
- Approximation error goes to zero
  - Requires measure theoretic proof
  - Speed depends on $\beta(a)$

Stochastic processes

Define the Markov chain S_t

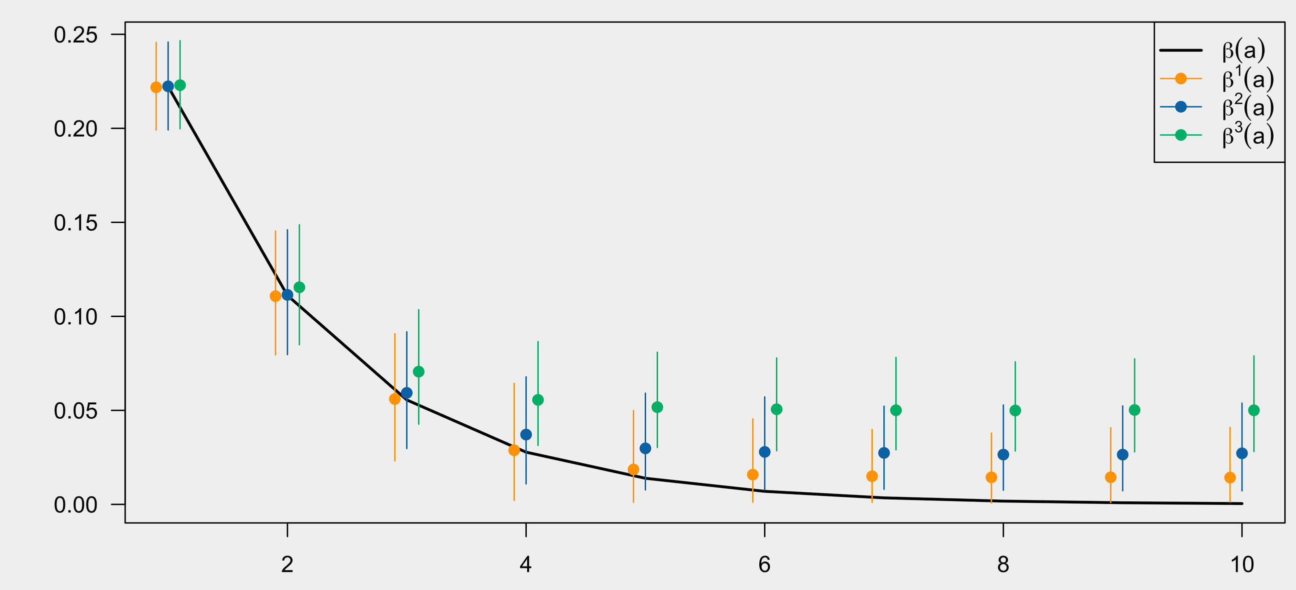


1. Observe S_t directly
   - The true β-mixing coefficient
$$\beta(a) = \beta^1(a) = \frac{4}{9}2^{-a}$$
   - Can be calculated exactly
   - Finite dimensional rate is exact, since the process is Markovian
2. Observe HMM ("Even" process)
$$O_t = \begin{cases} 1 & \text{if } (S_t, S_{t-1}) = (A,B) \text{ or } (B,A) \\ 0 & \text{else} \end{cases}$$
   - β-mixing coefficient is upper bounded
$$\beta(a) \leq \frac{8}{9}2^{-a}$$
   - Rate for the observed process is unknown — process is non-Markovian
   - Observed process is a function of the joint process which is Markovian — gives upper bound

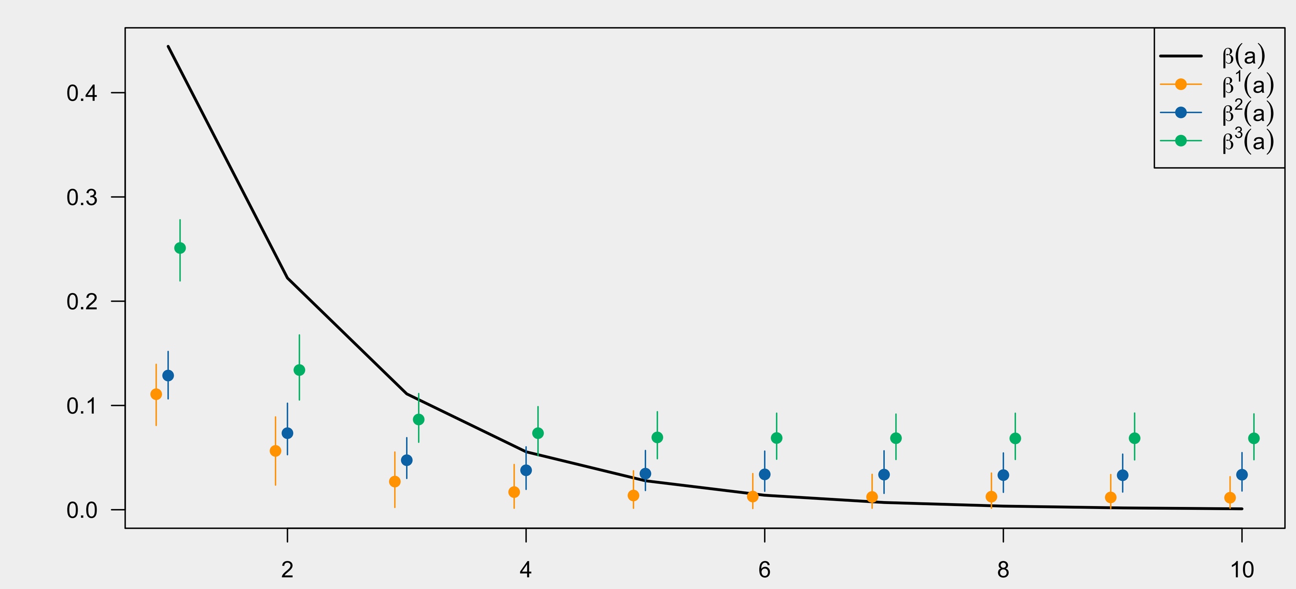
Simulations

- $n = 1000$
- $a \in \{1, 2, \ldots, 10\}$
- $d = o(\log_2 n)$ is the optimal rate
- 1000 replications

S_t



O_t ("Even" process)



Black line is an upper bound

Conclusions

- First procedure to estimate β-mixing coefficients
- Works reasonably well
- There is an upward bias to the estimates as d increases (estimator is nonnegative)
- Future work
  1. Eliminate the bias (decreases as $n \to \infty$)
  2. Rates of convergence for approximation error (likely impossible)
  3. Do we even need $\beta(a)$? Maybe just $\beta^d(a)$ for fixed d.

http://www.stat.cmu.edu/~danielmc
danielmc@stat.cmu.edu