Supplementary Materials: Information Theoretic Estimators for Dependence in Time Series

Anonymous Author(s)
Affiliation
Address
email

1 Convergence Rates

1.1 Density estimation for non-IID data

Most of the estimators discussed above rely on first estimating certain stationary probability densities functions of the variables involved. A natural first question to ask when studying these quantities is how well we can estimate the stationary distributions of time series data. In fact, under sufficient mixing conditions and boundedness assumptions on the domain of the time series, we can show uniform convergence rates as follows:

Theorem 5.3 of Fan and Yao (2003): Suppose the α-mixing coefficient of $\mathcal{X} = \{X_i\}_{i=1}^{\infty}$ satisfies $\alpha(\ell) \leq c\ell^{-\beta}$ for some $c > 0, \beta > 15/4$ and suppose the stationary density p of $\mathcal{X}$ has support in the interval $[a, b]$, then, for $h \asymp \left(\frac{\log T}{T}\right)^{\frac{6}{2\beta+5}}$ and a Lipschitz continuous kernel K,

$$\|\widehat{p}_h - p\|_\infty \in O_P\left(\left(\frac{\log T}{T}\right)^{\frac{2\beta-1}{4\beta+10}}\right),$$

where $\widehat{p}_h : \mathscr{X} \to [0, \infty)$ denotes the kernel density estimate trained on T samples, with kernel K and bandwidth h. Matching lower bounds are also known.

For continuous data, the α-mixing condition above is fairly weak and holds for many standard time series models (including, e.g., ARMA models). The boundedness of the domain is typically a necessary assumption for estimating information theoretic functionals of continuous data, and so it is not problematic either.

1.2 Consequences for estimating information theoretic functionals

An easy corollary relevant for our problem is the following:

Corollary: Suppose $\mathcal{X}, K$, and h satisfy the conditions of Theorem 5.3 above, and suppose, in addition, that the density f is lower bounded (i.e., $0 < \kappa := \inf_{x\in[a,b]} f(x)$). Then,

$$|\widehat{H}(\mathcal{X}) - H(\mathcal{X})| \in O_P\left(\left(\frac{\log T}{T}\right)^{\frac{2\beta-1}{4\beta+10}}\right),$$

where $\widehat{H}(\mathcal{X})$ is the plug-in estimator $\widehat{H}(\mathcal{X}) = -\int_a^b \widetilde{p}_h(x) \log \widetilde{p}_h(x)\, dx$ using the clipped density estimate $\widetilde{p}_h(x) = \max\{\kappa, \widehat{p}_h(x)\}$. [1]

[1] Note that we can avoid having to compute the integral here by splitting the data, using part to estimate the density and the remainder to compute the sample mean of the logarithm of the density estimate, without negatively affecting the convergence rates.

Proof: By Theorem 5.3 and the definition of O_P, it suffices to bound $|\widehat{H}(\mathcal{X}) - H(\mathcal{X})|$ by some constant multiple of $\|\widehat{p}_h - p\|_\infty$. Note that, for $x > \kappa$,

$$\left|\frac{d}{dx} x \log x\right| = |1 + \log x| \leq 1 + |\log \kappa|,$$

and so, by the Mean Value Theorem,

$$\begin{aligned}
|H(\mathcal{X}) - \widehat{H}(\mathcal{X})| &= \left| \int_a^b p(x) \log p(x) - \widetilde{p}_h(x) \log \widetilde{p}_h(x)\, dx \right| \\
&\leq \int_a^b |\|p - \widetilde{p}_h(x)\|_\infty (1 + |\log \kappa|)|\, dx \\
&= \|p - \widetilde{p}_h(x)\|_\infty (1 + |\log \kappa|)(b - a) \\
&= \|p - \widehat{p}_h(x)\|_\infty (1 + |\log \kappa|)(b - a),
\end{aligned}$$

since clipping improves the estimator point-wise. ■

A slight generalization of Theorem 5.3 above to multivariate time series (with, presumably, a slower rate depending exponentially on the number of variables) likely also holds. This would be sufficient to derive convergence rates for plug-in or von Mises estimators of information theoretic functionals involving multiple variables, such as conditional entropies. Since estimating both transfer entropy and the mutual information rate can be reduced to estimating conditional entropies, following this line of thought, it should be possible to derive convergence rates for plug-in (and von Mises) estimators of these quantities.

2 Derivation of the first-order von Mises estimators

We use estimators based on the von Mises expansion as described in Kandasamy et al. (2014), although other approaches are also possible.

As noted in equation (7), we can write the MIR in terms of conditional entropies, and, similarly, we can write transfer entropy as

$$\begin{aligned}
T^n_{\mathcal{X}\to\mathcal{Y}} &= I(Y_n; \{X_i\}_{i=i-\beta}^{n-1} | \{Y_i\}_{i=i-\beta}^{n-1}) \\
&= H(Y_n | \{Y_i\}_{i=i-\beta}^{n-1}) + H(\{X_i\}_{i=i-\beta}^{n-1} | \{Y_i\}_{i=i-\beta}^{n-1}) - H(Y_n, \{X_i\}_{i=i-\beta}^{n-1} | \{Y_i\}_{i=i-\beta}^{n-1})
\end{aligned}$$

Thus, estimation of both transfer entropy and MIR reduces to estimating conditional entropies, for which we derive an estimator below.

2.1 Derivation of the first-order von Mises Estimator for Conditional Entropy

Consider a random pair (X, Y) distributed on $\mathcal{X} \times \mathcal{Y}$ with joint density p, and let $q(y) = \int_{\mathcal{X}} p(x, y), dx$ denotes the marginal distribution of Y. Then,

$$H(X|Y) = -\int_{\mathcal{Y}} q(y) \int_{\mathcal{X}} \frac{p(x,y)}{q(y)} \log \frac{p(x,y)}{q(y)}\, dx\, dy = \int_{\mathcal{X}\times\mathcal{Y}} p(x,y) \log \frac{q(y)}{p(x,y)}\, d(x,y).$$

Let $\widehat{p}$ and $\widehat{q}$ be estimates of p and q, respectively. Note that

$$\begin{aligned}
&\int_{\mathcal{X}\times\mathcal{Y}} \left(\frac{d}{d\widehat{p}(x,y)} p(x,y) \log \frac{\widehat{q}(y)}{\widehat{p}(x,y)} \right) (p(x,y) - \widehat{p}(x,y))\, d(x,y) \\
&= \int_{\mathcal{X}\times\mathcal{Y}} \left(\log \frac{\widehat{q}(y)}{\widehat{p}(x,y)} - 1 \right) (p(x,y) - \widehat{p}(x,y))\ d(x,y) \\
&= \int_{\mathcal{X}\times\mathcal{Y}} (p(x,y) - \widehat{p}(x,y)) \log \frac{\widehat{q}(y)}{\widehat{p}(x,y)}\, d(x,y) = \mathbb{E}\left[\log \frac{\widehat{q}(y)}{\widehat{p}(x,y)} \right] - H(\widehat{p}, \widehat{q}),
\end{aligned}$$

and that

$$\int_{\mathcal{X}\times\mathcal{Y}} \left(\frac{d}{dq(y)}\widehat{p}(x,y)\log\frac{\widehat{q}(y)}{\widehat{p}(x,y)}\right)(q(y)-\widehat{q}(y))\,d(x,y) = \int_{\mathcal{X}\times\mathcal{Y}} \frac{\widehat{p}(x,y)}{\widehat{q}(y)}(q(y)-\widehat{q}(y))\,d(x,y)$$
$$= \int_{\mathcal{Y}} q(y)-\widehat{q}(y)\,dy = 0.$$

Thus, von Mises expansion gives

$$\begin{aligned} H(p,q) &= H(\widehat{p},\widehat{q}) + \int_{\mathcal{X}\times\mathcal{Y}} \left(\frac{d}{dp(x,y)}\widehat{p}(x,y)\log\frac{\widehat{q}(y)}{\widehat{p}(x,y)}\right)(p(x,y)-\widehat{p}(x,y)) \\ &\quad + \left(\frac{d}{dq(y)}\widehat{p}(x,y)\log\frac{\widehat{q}(y)}{\widehat{p}(x,y)}\right)(q(y)-\widehat{q}(y))\,d(x,y) + O(\|p-\widehat{p}\|_2^2 + \|q-\widehat{q}\|_2^2) \\ &= \mathbb{E}_{(X,Y)\sim p}\left[\log\frac{\widehat{q}(Y)}{\widehat{p}(X,Y)}\right] + O(\|p-\widehat{p}\|_2^2 + \|q-\widehat{q}\|_2^2). \end{aligned}$$

Thus, using data points $\{(X_i,Y_i)\}_{i=1}^{2n}$, the first-order von Mises estimator for $H(X|Y)$ is thus

$$\widehat{H}(X|Y) = \frac{1}{n}\sum_{i=n+1}^{2n}\log\frac{\widehat{q}(Y_i)}{\widehat{p}(X_i,Y_i)}$$

where $\widehat{p}$ and $\widehat{q}$ are estimated using $\{(X_i,Y_i)\}_{i=1}^{2n}$.

References

Jianquing Fan and Qiwei Yao. *Nonlinear time series: nonparametric and parametric methods*. Springer, 2003.

Kirthevasan Kandasamy, Akshay Krishnamurthy, Barnabas Poczos, Larry Wasserman, and James M Robins. Influence functions for machine learning: Nonparametric estimators for entropies, divergences and mutual informations. *arXiv preprint arXiv:1411.4342*, 2014.