On the Estimation of Entropy for Markov Chains
Sur l'estimation de l'entropie des chaînes de Markov

Girardin, Valérie and Regnault, Philippe

Laboratoire de Mathématiques Nicolas Oresme, UMR6139, Campus II, Université de Caen, BP5186, 14032 Caen, France, girardin@math.unicaen.fr, regnault@math.unicaen.fr

RÉSUMÉ (ABSTRACT)

L'entropie marginale et le taux d'entropie de Shannon des chaînes de Markov à temps discret ou continu sont des fonctions régulires des coefficients de leur matrice de transition ou générateur. Des estimateurs par branchement peuvent ainsi être déduits des estimateurs de ces coefficients. Pour les entropies généralisées du type Rényi ou Tsallis, une expression exacte du taux d'entropie est obtenue avant d'être estimée, à l'aide d'outils de théorie des opérateurs. Les chaînes de Markov considérées incluent les espaces d'état binaires, finis et dénombrables, les suites indépendantes et identiquement distribuées (i.i.d.), et les chaînes paramétriques. Tous les estimateurs construits ont de bonnes propriétés asymptotiques, de lois asymptotiques totalement explicites. Pour les suites i.i.d., un principe de grande déviations impliquant l'entropie relative d'une loi escorte de la suite est énoncé.

The Shannon marginal entropy and entropy rate of ergodic Markov chains either with discrete or continuous time, are smooth functions of the coefficients of their transition matrices or generators. Plug-in estimators of entropy can thus be constructed from estimators of these coefficients. For generalized entropies such as Rényi or Tsallis, a closed form expression of the entropy rate has to be obtained prior to estimation by means of operator theory methods. Considered Markov chains include binary, finite or denumerable state spaces, independent identically distributed (i.i.d.) sequences and parametric chains. All the constructed estimators behave asymptotically well, with limit distributions generally depending on explicit parameters. Moreover, for an i.i.d. sequence, a large deviation principle holds, involving the relative entropy of an escort distribution of the sequence.

1 Introduction

Shannon (1948) adapted to the field of probability the concept of entropy introduced by Boltzmann and Gibbs in the XIX-th century by defining the entropy of a distribution P taking values in a countable set E as

$$(1) \quad \mathbb{S}(P) = -\sum_{i\in E} P(i)\log P(i),$$

with the convention $0\log 0 = 0$. Entropy measures the randomness or uncertainty of a random phenomenon. It now naturally applies to information theory and statistical mechanics, and to many other fields such as finance, statistics, cryptography, physics, artificial intelligence, etc.

Rényi (1960) proposed a one parameter family of entropy functionals extending Shannon entropy to new applications. Since then, many different generalized entropies have been defined to adapt to many different fields. Among them, Tsallis or Sharma-Mittal entropies are instances of what Menéndez *et al* (1997) call (h,ϕ)-entropy. Precisely, we set

$$(2) \quad \mathbb{S}_{h(y),\phi(x)}(P) = h\left(\sum_{i\in E}\phi\big(P(i)\big)\right)$$

for any measure P on a countable space E such that the quantity is finite. Note that Rényi entropy is obtained for $h_s(y) = (1-s)^{-1}\log y$ and $\phi_s(x) = x^s$ with $s > 0$, while Tsallis entropy involves the functions $h_r(y) = (r-1)^{-1}(1-y)$ and $\phi_r(x) = x^r$ for some positive $r \neq 1$. Shannon entropy is obtained for s tending to 1 and for $r = 1$. More generally, Sharma-Mittal entropies are obtained for $h_{s,r}(y) = (r-1)^{-1}[1 - y^{(1-r)/(1-s)}]$ and $\phi_s(x) = x^s$.

Entropy can be extended to a discrete or continuous-time stochastic process $\mathbf{X} = (X_t)$ by considering entropy for its marginal distributions, that is of the distributions of X_t, for $t \in \mathbb{N}$ or $\mathbb{R}_+$. If the process is stationary, these distributions are all equal to some P, and the marginal entropy is $\mathbb{S}(P)$. When the process is ergodic, its asymptotic behavior is described by some unique probability distribution, say P again; then, $\mathbb{S}(P)$ measures the information of the process at equilibrium. In case $\mathbf{X}$ is both stationary and ergodic, its marginal and limit distributions are equal.

The entropy rate of a discrete time process $\mathbf{X} = (X_n)_{n\in\mathbb{N}}$ is usually defined as its entropy per unit time. For an i.i.d. sequence, Shannon and Rényi entropy rates are well-known to be the entropy of the marginal distribution. This is due to additivity properties which are lost for non-extensive systems to which Tsallis entropy fits better; see Tsallis (2010). The Shannon entropy rate of an ergodic homogeneous Markov chain $\mathbf{X} = (X_n)_{n\in\mathbb{N}}$ with a countable state space E, transition matrix $\mathbf{P} = (\mathbf{P}(i,j))_{i,j\in E}$ and stationary distribution P (such that $P\mathbf{P} = P$) is

$$(3) \quad \mathbb{H}(\mathbf{X}) = H(\mathbf{P}) = -\sum_{i\in E} P(i) \sum_{j\in E} \mathbf{P}(i,j) \log \mathbf{P}(i,j).$$

Shannon (1948) proved the convergence of the time average $\frac{1}{n}\log \mathbb{P}(X_0 = i_0, \dots, X_{n-1} = i_{n-1})$ to $\mathbb{H}(\mathbf{X})$, a limit which defines the Shannon entropy rate of any discrete time stationary ergodic process. For generalized entropies, the entropy rate of the sequence is similarly defined as $\mathbb{H}_{h(y),\phi(x)}(\mathbf{X}) = \lim_{n\to\infty} \frac{1}{n}\mathbb{S}_{h(y),\phi(x)}(X_0, \dots, X_{n-1})$, where $\mathbb{S}_{h(y),\phi(x)}$ is given in (2).

A natural equivalent of the entropy rate for a continuous-time stochastic process $\mathbf{X} = (X_t)_{t\in\mathbb{R}_+}$ comes from considering the limit of the time average $\frac{1}{T}H_T(\mathbf{X})$, with $H_T(\mathbf{X}) = -\int_{\mathbb{R}_+} f_{X_{(T)}} \log f_{X_{(T)}} d\mu$ for $T > 0$, where the distribution $P_{X_{(T)}}$ of $X_{(T)} = (X_t)_{0\le t\le T}$ is supposed to be dominated by μ, with density $f_{X_{(T)}}$. Bad Dumitrescu (1986) proved that the entropy rate exists for any ergodic continuous-time Markov chain with finite state-space E, infinitesimal generator $\mathbf{A} = (\mathbf{A}(i,j))_{(i,j)\in E^2}$ and stationary distribution P (such that $P\mathbf{A} = 0$), is given by

$$(4) \quad \mathbb{H}(\mathbf{X}) = \mathbf{H}(\mathbf{A}) = \lim_{T\to+\infty} \frac{1}{T} H_T(X) = -\sum_{i\in E} P(i) \sum_{j\neq i} \mathbf{A}(i,j) \log \mathbf{A}(i,j) + \sum_{i\in E} P(i) \sum_{j\neq i} \mathbf{A}(i,j).$$

In this review paper, we will consider first in Section 2 the estimation of Shannon entropy for an i.i.d. sequence with a stress on a large deviation principle. We will address the problem of estimating generalized entropy for discrete time parametric Markov chains, precisely marginal entropy in Section 3.1, and entropy rate in Section 3.2. Finally, we will estimate Shannon marginal entropy and entropy rate for continuous-time Markov chains in Section 4.

2 Estimation of Shannon entropy for i.i.d. sequences

For an i.i.d. sample of some distribution P, the estimator of entropy obtained by plugging the empirical estimator of P into (1) has been considered in the 50's; see Harris (1977) and the references therein. A complete proof of its asymptotic properties is given in Girardin and Regnault (2011), where the following large deviations principle is also proven to hold.

Theorem 1 *Let $\mathbf{X} = (X_n)_{n\in\mathbb{N}}$ be an i.i.d. sequence of observations of a distribution P on a finite state space E. Let $\widehat{P}_n(i) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{(X_k=i)}$, denote the empirical estimator of $P(i)$. Then $\mathbb{S}(\widehat{P}_n)$ is a strongly consistent estimator of Shannon entropy $\mathbb{S}(P)$. Moreover:*

- *if P is not uniform, then $\sqrt{n}[\mathbb{S}(\widehat{P}_n) - \mathbb{S}(P)]$ is asymptotically normal with variance*

$$\sum_{i \in E^*} P(i)[1 - P(i)] \left(\log \frac{P(i)}{[1 - \sum_{j \in E} P(j)]} \right)^2 ;$$

- *if P is uniform, say $P = U$, then $2n[\mathbb{S}(\widehat{P}_n) - \mathbb{S}(U)]$ converges to $\sum_{i \in E} \beta_i Y_i$, where all Y_i are independent $\chi^2(1)$-distributed random variables and all $\beta_i \in \mathbb{R}$.*

The sequence of estimators $\mathbb{S}(\widehat{P}_n)$ satisfies a large deviations principle with good rate function

$$I_{\mathbb{S}}(s) = \begin{cases} -s - \log(p) & \text{if } 0 \leq s \leq \log m, \\ \mathbb{K}(E_P^k | P) & \text{if } \log(m) < s \leq \log(|E|), \text{ with } k > 0 \text{ such that } \mathbb{S}(E_P^k) = s, \\ +\infty & \text{otherwise,} \end{cases}$$

where $m = |\{i \in E : P(i) = p\}|$ is the number of modes of P, taking value $p = \max_{i \in E} P(i)$, the distribution E_P^k with $E_P^k(i) = P(i)^k / \sum_{j \in E} P(j)^k$, is the k-escort distribution of P with Kullback-Leibler information relative to P given by $\mathbb{K}(E_P^k \mid P) = \sum_{i \in E} E_P^k(i) \log[E_P^k(i)/P(i)]$.

To establish a closed-form expression for k as a function of the entropy level s seems to be a difficult task, whatever be the cardinality of E. As an alternative, the original rate function $I_{\mathbb{S}}$ can be replaced by an explicite approximation in all applications involving large deviations principles, without significative loss of accuracy; see Girardin and Regnault (2011) for details.

3 Generalized entropies for discrete time Markov chains

Very few results exist in the literature concerning estimation of entropy for non i.i.d. sequences, especially Markov chains, even less concerning the estimation of generalized entropy functionals.

We suppose that the transition probabilities of the Markov chain depend on an unknown parameter $\theta \in \Theta^d$, where Θ is an open subset of some Euclidean space and $d \geq 1$. Billingsley (1961) shows that a strongly consistent maximum likelihood estimator $\hat{\theta}_n$ of θ exists, such that $\sqrt{n}(\hat{\theta}_n - \theta)$ is asymptotically centered and normal, with variance matrix $\sigma^{-1}(\theta)$ which is the inverse of Fisher's Information of $\mathbf{X}$, under a set of regularity assumptions too long to be given here. Since the transition probabilities of the chain depend on θ, its stationary distribution P also depends on θ, say through $P[\theta]$, and it is natural to consider the plug-in estimator $\mathbb{S}_{h(y),\phi(x)}(P[\hat{\theta}_n])$ of $\mathbb{S}_{h(y),\phi(x)}(P[\theta])$.

3.1 Marginal entropy

Ciuperca *et al* (2011) establishes the good asymptotic properties of the plug-in estimator of the marginal entropy of a parametric Markov chain, by means of operator theory tools. The proof is based on the so-called quasi-power property. In dynamical systems theory, the Dirichlet series $\Lambda_n(s) = \sum_{i_k \in E} P^n(i_0, \dots, i_n)^s$ is a central tool for studying general sources, in pattern matching or in the analysis of data structures. For an i.i.d. sequence with non-degenerated distribution P over a finite set E, it can be simply written as the n-th power of an analytic function, precisely $\Lambda_n(s) = \left[\sum_{i \in E} P(i)^s\right]^n$. Similarly, for more general random sequences, the quasi-power property says that Λ_n behaves like the n-th power of some analytic function, precisely $\Lambda_n(s) = c(s) \cdot \lambda(s)^{n-1} + R_n(s)$ with $|R_n(s)| = O\left(\rho(s)^{n-1}\lambda(s)^{n-1}\right)$, where c and λ are positive analytic functions for $s > \sigma_0$, and λ is strictly decreasing with $\lambda(1) = c(1) = 1$, and $\rho(s) < 1$.

Theorem 2 *Let $\mathbf{X} = (X_n)_{n \in \mathbb{N}}$ be an ergodic homogeneous finite discrete time Markov chain satisfying the quasi-power property. If Billingsley (1961)'s set of asssumptions is satisfied, then the estimator $\mathbb{S}_{h(y),\phi(x)}(P[\hat{\theta}_n])$ is strongly consistent.*

If, moreover, the differential $D_\theta \mathbb{S}_{h(y),\phi(x)}(P[\theta])$ *is not null, then* $\sqrt{n}\left[\mathbb{S}_{h(y),\phi(x)}(P[\hat{\theta}_n]) - \mathbb{S}_{h(y),\phi(x)}(P[\theta])\right]$ *is asymptotically normal with variance* $\left[D_\theta \mathbb{S}_{h(y),\phi(x)}(P[\theta])\right]^t \sigma^{-1}(\theta) \left[D_\theta \mathbb{S}_{h(y),\phi(x)}(P[\theta])\right]$.

All finite chains are parametric chains, since the transition probabilities are functions of a number d of parameters equal or less than $|E|(|E|-1)$. For a finite-state ergodic Markov chain, the asymptotic normality of the plug-in empirical estimator of Shannon marginal entropy $\mathbb{S}(\widehat{P}_n)$, is established in Ciuperca and Girardin (2007), thanks to the ergodic theorem for Markov chains and delta method. Different schemes of observations are considered, according to whether one long trajectory or several short trajectories are observed. Due to the presence of the sum over all the states in the expression of the entropy, and to the dependencies of the $\sum_{k=1}^{n} 1\!\!1_{(X_k=i)}$ for different i, the asymptotic variance $\mathbb{S}(\widehat{P}_n)$ cannot be obtained explicitely.

3.2 Entropy rate

Rényi entropy rate is proven in Rached *et al* (1999) to be $\mathbf{h}_s = (1-s)^{-1} \log \lambda(s)$ for any finite state-space E, where $\lambda(s)$ for $s > 0$ is the unique dominant eigenvalue of the perturbated transition matrix $\mathbf{P}_s = (\mathbf{P}(i,j)^s)_{i,j \in E}$. Note that Shannon entropy rate given in (3) is equal to the derivative $\mathbf{h} = -\lambda'(1)$.

In Ciuperca *et al* (2011), the (h,ϕ)-entropy rates of discrete time random sequences – especially Markov chains, taking values in countable spaces, are computed and estimated by applying operators theory tools, thanks to the quasi-power property. When the (h,ϕ)-entropy rate is neither null nor infinite, only two cases happen : either, the entropy rate is equal to Shannon entropy rate, or it is a simple function of Rényi entropy rate. Therefore, only the estimation of Shannon and Rényi entropy rates is to be detailed. Let us define for a parametric Markov chain the plug-in estimators $\mathbf{h}(\hat{\theta}_n) = -\lambda'(1;\hat{\theta}_n)$ of Shannon entropy rate $\mathbb{H}(\mathbf{X}) = -\lambda'(1;\theta)$, and $\mathbf{h}_s(\hat{\theta}_n) = (1-s)^{-1}\log\lambda(s;\hat{\theta}_n)$ of Rényi entropy rate $\mathbb{H}_s(\mathbf{X}) = (1-s)^{-1}\log\lambda(s;\theta)$.

Theorem 3 *Let* $\mathbf{X} = (X_n)_{n\in\mathbb{N}}$ *be an ergodic homogeneous countable Markov chain satisfying the quasi-power property. If Billingsley (1961)'s set of assumptions is satisfied, then the estimators* $\mathbf{h}(\hat{\theta}_n)$ *and* $\mathbf{h}_s(\hat{\theta}_n)$ *for* $s \neq 1$, *are strongly consistent and* $\sqrt{n}[\mathbf{h}(\hat{\theta}_n) - \mathbb{H}(\mathbf{X})]$ *and* $\sqrt{n}[\mathbf{h}_s(\hat{\theta}_n) - \mathbb{H}_s(\mathbf{X})]$ *are asymptotically normal with respective variances*

$$\Sigma_1 = \left\{\frac{\partial}{\partial\theta}[-\lambda'(1;\theta)]\right\}^t \sigma^{-1}(\theta)\frac{\partial}{\partial\theta}[-\lambda'(1;\theta)] \quad \textit{and} \quad \Sigma_s = \frac{1}{(1-s)^2}\left\{\frac{\partial}{\partial\theta}\lambda(s;\theta)\right\}^t \sigma^{-1}(\theta)\frac{\partial}{\partial\theta}\lambda(s;\theta).$$

All finite chains are parametric chains, hence their entropy rate can be estimated as above. Alternatively, the plug-in estimator constructed from the empirical estimator

$$(5) \quad \widehat{\mathbf{P}}(i,j) = \frac{\sum_{m=0}^{n-1} 1\!\!1_{X_m=i, X_{m+1}=j}}{\sum_{m=0}^{n} 1\!\!1_{X_m=i}}$$

obtained from one long trajectory can be considered. For a finite-state ergodic Markov chain with non uniform transition matrix, the asymptotic normality of $H(\widehat{\mathbf{P}})$ is established in Ciuperca and Girardin (2007), thanks to the ergodic theorem for Markov chains and delta method. Different schemes of observations are also considered. Again, due to the presence both of the sum over all the states in the expression of the entropy and of dependencies of the $\widehat{\mathbf{P}}(i,j)$, the asymptotic variance of $H(\widehat{\mathbf{P}})$ cannot be obtained in general. For a Markov chain $\mathbf{X}$ with a two-state space, say $E = \{0,1\}$, the asymptotic properties have been fully established in Girardin and Sesboüé (2009). Indeed, the entropy rate (3) of the chain can be written

$$(6) \quad \mathbb{H}(\mathbf{X}) = H(p,q) = \frac{q}{p+q}S_p + \frac{p}{p+q}S_q,$$

where $\mathbf{P}(0,1) = p$ and $\mathbf{P}(1,0) = q$, so that $P(0) = q/(p+q)$ and $P(1) = p/(p+q)$, with $S_p = -p\log p - (1-p)\log(1-p)$ and $S_q = -q\log q - (1-q)\log(1-q)$.

Replacing in (6) p and q by their estimators $\widehat{p}_n$ and $\widehat{q}_n$ obtained by (5), we get the plug-in estimator of the entropy rate of the chain $H(\widehat{p}_n, \widehat{q}_n)$.

Proposition 1 *Let $\mathbf{X} = (X_n)_{n\in\mathbb{N}}$ be an ergodic homogeneous two-state Markov chain. The plug-in empirical estimator $H(\widehat{p}_n, \widehat{q}_n)$ of the entropy rate $\mathbb{H}(\mathbf{X})$ is strongly consistent. Moreover:*

- *if $(p,q) \neq (0.5, 0.5)$, then $\sqrt{n}[H(\widehat{p}_n, \widehat{q}_n) - \mathbb{H}(\mathbf{X})]$ is asymptotically normal with variance $\gamma_0^2[\partial_1^1 H(p,q)]^2 + \gamma_1^2[\partial_2^1 H(p,q)]^2$, where*

$$\gamma_0^2 = \frac{p(1-p)}{P(0)} = \frac{p(1-p)(p+q)}{q} \quad \textit{and} \quad \partial_1^1 H(p,q) = \frac{q}{(p+q)^2}[S_q - S_p] - \frac{q}{p+q}\log\frac{p}{1-p},$$
$$\gamma_1^2 = \frac{q(1-q)}{P(1)} = \frac{q(1-q)(p+q)}{p} \quad \textit{and} \quad \partial_2^1 H(p,q) = \frac{p}{(p+q)^2}[S_p - S_q] - \frac{p}{p+q}\log\frac{q}{1-q};$$

- *if $(p,q) = (0.5, 0.5)$, then $2n[H(\widehat{p}_n, \widehat{q}_n) - \mathbb{H}(\mathbf{X})]$ converges in distribution to a $\chi^2(2)$-distribution.*

4 Shannon entropy for continuous time Markov chains

The empirical estimator of the generator $\mathbf{A} = (\mathbf{A}(i,j))_{(i,j)\in E^2}$ of a continuous time Markov chain $\mathbf{X}$ (also called pure jump Markov process) is

$$(7) \quad \widehat{A}_T(i,j) = \frac{\sum_{t=0}^{N_T-1} \mathbb{1}_{X_m=i, X_{m+1}=j}}{\int_0^T \mathbb{1}_{X_t=i}dt}, \qquad \text{for } j \neq i, \qquad \text{and } \widehat{A}_T(i,i) = -\sum_{j\neq i} \widehat{A}_T(i,j),$$

where N_T is the number of jumps of the process in the time interval $[0,T]$. Albert (1962) showed that $\widehat{A}_T$ is strongly consistent and that $\sqrt{T}(\widehat{A}_T - A)$ is asymptotically normal with diagonal variance matrix with entries $\Sigma^2_{\mathbf{A}}(i,j) = \mathbf{A}(i,j)/P(i)$ for $i \neq j$. He also computed the stationary distribution P of the process, such that $P\mathbf{A} = 0$, through

$$(8) \quad P(i) = P[\mathbf{A}](i) = \frac{\det \mathbf{A}^{(i,i)}}{\sum_{j\in E} \det \mathbf{A}^{(j,j)}}, \quad i \in E,$$

where $\mathbf{A}^{(i,i)}$ is the $(|E|-1)^2$-matrix obtained from $\mathbf{A}$ by canceling both i-th row and i-th column.

4.1 Marginal entropy

Up to our knowledge, the only available results on the estimation of the entropy of continuous time Markov chains are the following, proven in Regnault (2011). The asymptotic properties of the estimator of the stationary distribution are a necessary first step.

Theorem 4 *Let $\mathbf{X} = (X_t)_{t\in\mathbb{R}_+}$ be an ergodic continuous time Markov chain with finite state space E, generator $\mathbf{A}$ and stationary distribution P.*

The plug-in estimator $P[\widehat{A}_T]$ of P obtained using (7) and (8) is strongly consistent and $\sqrt{T}(P[\widehat{A}_T] - P)$ is asymptotically normal with variance $\Sigma_P^2 = D_P(\mathbf{A}).\Sigma_A^2.D_P(\mathbf{A})^t$. This normal distribution is never degenerated and the rate of convergence is optimal.

The plug-in estimator $\mathbb{S}(P[\widehat{A}_T])$ of $\mathbb{S}(P)$ is strongly consistent. Moreover:

- *if the differential $D_S(\mathbf{A})$ is not null, then $\sqrt{T}(\mathbb{S}(P[\widehat{A}_T]) - \mathbb{S}(P))$ is asymptotically normal with variance $\Sigma_P^2 = D_S(\mathbf{A}).\Sigma^2_{\mathbf{A}}.D_S(\mathbf{A})^t$.*
- *if $D_S(\mathbf{A}) = 0$, then $2T[\mathbb{S}(P) - \mathbb{S}(P[\widehat{A}_T])]$ converges in distribution to $\sum_{(i,j)\in E^{2*}} \alpha_{(i,j)} Y_{(i,j)}$, where all random variables $Y_{(i,j)}$ are $\chi^2(1)$-distributed and the $\alpha_{(i,j)}$ explicitely depend on $\Sigma^2_{\mathbf{A}}$.*

Conditions on $\mathbf{A}$ are given in Regnault (2011) for the nullity of $D_S(\mathbf{A})$. Three situations of observation are also discussed, according to whether one long trajectory is observed, or several short independent trajectories are observed, or the process is observed at discrete times.

The case of a two-state process is studied in full details in Regnault (2009b). In particular, the asymptotic variances take the form

$$\Sigma_P^2 = \frac{2ab}{(a+b)^4}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \quad \text{and} \quad \Sigma_S^2 = \frac{2ab}{(a+b)^3}\left(\log\frac{a}{b}\right)^2,$$

where $a = \mathbf{A}(0,1)$ and $b = \mathbf{A}(1,0)$.

4.2 Entropy rate

For estimating the entropy rate $\mathbb{H}(\mathbf{X})$ given by (4) , it is natural to consider the plug-in estimator $\mathbf{H}(\widehat{A}_T)$ using the empirical estimator $\widehat{A}_T$ given in (7). Its following good asymptotic properties are proven to hold in Regnault(2009a). They derive from the fact that $\widehat{H}_T$ is a continuous mapping of the strongly consistent estimator $\widehat{A}_T$. The rest follows from the delta method.

Theorem 5 *Let $\mathbf{X} = (\mathbf{X_t})_{\mathbf{t}\in\mathbb{R}_+}$ be an ergodic continuous time Markov chain with finite state space E, generator $\mathbf{A}$ and stationary distribution P. The plug-in estimator $\mathbf{H}(\widehat{A}_T)$ of the entropy rate $\mathbb{H}(\mathbf{X})$ is strongly consistent. Moreover:*

- *if $D_{\mathbf{H}}(\mathbf{A}) = \left(\frac{\partial \mathbf{H}}{\partial A(i,j)}(\mathbf{A})\right)$ is not null, then $\sqrt{T}[\mathbf{H}(\widehat{A}_T) - \mathbb{H}(\mathbf{X})]$ is asymptotically normal with variance $\Sigma_{\mathbf{H}}^2 = \sum_{i\neq j}\mathbf{A}(i,j)[\frac{\partial \mathbf{H}}{\partial A(i,j)}(\mathbf{A})]^2/P(i)$.*
- *if $D_{\mathbf{H}}(\mathbf{A}) = 0$, then $2T[\mathbf{H}(\widehat{A}_T) - \mathbb{H}(\mathbf{X})]$ converges in distribution to $\sum_{(i,j)\in E^{2*}}\alpha_{(i,j)}Y_{(i,j)}$, where all random variables $Y_{(i,j)}$ are $\chi^2(1)$-distributed and the $\alpha_{(i,j)}$ explicitely depend on $\Sigma_{\mathbf{A}}^2$.*

The case of a two-state process is studied in full details in Regnault (2009b). In particular, $\Sigma_{\mathbf{H}}^2 = \left([b - a - b\log(ab)]^2 + [a - b - a\log(ab)]^2\right) ab/(a+b)^3$.

REFERENCES

A. Albert (1962) *Ann. Math. Stat.*, V33, pp.727–753.

M. Bad Dumitrescu (1986) *Casopis Pro Pestovani Matematiky* V4, pp.429–434.

P. Billingsley (1961) *Ann. Math. Stat.* V32, pp.12–40.

G. Ciuperca and V. Girardin (2007) *Comm. Stat.: Theory and Methods*, V36, pp2543–2557.

G. Ciuperca, V. Girardin and L. Lhote (2011) *IEEE Trans on Inform Theory*, to appear in June 2011.

V. Girardin and Ph. Regnault (2011) *Technical Report,* Université de Caen, France.

V. Girardin and A. Sesboüé (2009) *Method Comp Appl Probab*, V11, pp.181–200.

B. Harris (1977) *Colloq. Math. Soc. Janos Bolyai*, V16, pp323–355, North-Holland, Amsterdam.

M.L. Menéndez, D. Morales, L. Pardo, and M. Salicrú (1997) *Appl. Math.*, V42, pp.81–98.

Z. Rached, F. Alajaji, and L. Campbell (1999) *Proc. CISS*, pp613-618.

Ph. Regnault (2009a) *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, V1193, pp.153-160, AIPCP.

Ph. Regnault (2009b) *Compte-rendus des Journées de la Société Française de Statistique*, Bordeaux.

Ph. Regnault (2011) *Journal of Statistical Planning and Inference* V141, pp2503–2986.

A. Rényi (1960) *Proc. 4th Berkeley Symposium on Mathematics, Statistics and Probability*, pp.547-561.

C. E. Shannon (1948) *Bell Syst. Tech. J.*, V27, I, pp. 379–423, II, pp.623–656.

C. Tsallis (2009) *Introduction to Nonextensive Statistical Mechanics* Springer, New York.